nature ecology & evolution

ARTICLES
https://doi.org/10.1038/s41559-018-0575-6

Early metazoan cell type diversity and the evolution of multicellular gene regulation

Arnau Sebé-Pedrós[1,2]*, Elad Chomsky[1,2], Kevin Pang[3], David Lara-Astiaso[4], Federico Gaiti[5,6], Zohar Mukamel[1,2], Ido Amit[4], Andreas Hejnol[3], Bernard M. Degnan[5] and Amos Tanay[1,2]*

A hallmark of metazoan evolution is the emergence of genomic mechanisms that implement cell-type-specific functions. However, the evolution of metazoan cell types and their underlying gene regulatory programmes remains largely uncharacterized. Here, we use whole-organism single-cell RNA sequencing to map cell-type-specific transcription in Porifera (sponges), Ctenophora (comb jellies) and Placozoa species. We describe the repertoires of cell types in these non-bilaterian animals, uncovering diverse instances of previously unknown molecular signatures, such as multiple types of peptidergic cells in Placozoa. Analysis of the regulatory programmes of these cell types reveals variable levels of complexity. In placozoans and poriferans, sequence motifs in the promoters are predictive of cell-type-specific programmes. By contrast, the generation of a higher diversity of cell types in ctenophores is associated with lower specificity of promoter sequences and the existence of distal regulatory elements. Our findings demonstrate that metazoan cell types can be defined by networks of transcription factors and proximal promoters, and indicate that further genome regulatory complexity may be required for more diverse cell type repertoires.

The origin of animal multicellularity has been linked to the spatial coexistence of cell types with distinct roles[1,2]. Cell type specialization is achieved through asymmetric access to genomic information, which is interpreted in a cell-specific fashion through mechanisms of transcriptional gene regulation. However, it remains unclear how elaborate genome regulation relates to cell type diversity. Poorly characterized, early-branching metazoans represent an opportunity to explore these questions by studying how cell-type-specific genome regulation is implemented in species with (presumed) intermediate-to-low organismal complexity. Phylogenetically, sponges, comb jellies and placozoans—together with the remaining metazoans (Planulozoa)—are the earliest-branching animal lineages[3–6] (Fig. 1). These organisms possess characteristic body plans and have been traditionally considered to contain low numbers of cell types[7], although our current understanding of this diversity of cell behaviours remains very limited. Moreover, these three lineages diverged over 650 million years ago (Ma)[8], which has resulted in extremely different and specialized morphologies, life strategies and body plan organization[9]. Ctenophores are mostly pelagic, marine predators. They have tissue-level organization and they develop a nervous system of uncertain homology with their bilaterian counterparts[10–12]. By contrast, sponges are sessile filter-feeders that live both in marine and freshwater environments and that seem to have no or very rudimentary specialized tissues[13]. Finally, placozoans are tiny benthic marine animals with a body plan organization that is composed of two cell layers. They possess ciliary-based locomotion and feed on algae using external digestion[14]. Sponges, ctenophores and placozoans also vary considerably in their overall genome size, median intergenic space and repertoire of potential transcriptional and post-transcriptional regulators (Fig. 1). The genome of the sponge *Amphimedon queenslandica* measures 166 megabases (Mb), and its annotation suggests a relatively compact gene arrangement with very short (0.6-kilobase (kb)) intergenic regions[15,16]. In comparison, similar genome size (156 Mb) but longer (2 kb) intergenic regions are found in the ctenophore *Mnemiopsis leidyi*[17]. In the case of the placozoan *Trichoplax adhaerens*, a smaller genome (98 Mb) but longer intergenic regions (2.7 kb) are reported[18]. Annotation and comparison of the predicted proteome in these non-bilaterian species uncovered an extensive suite of gene families shared across Metazoa[15,17–19], suggesting the existence of ancient regulatory mechanisms for orchestrating cell type specification and maintenance. For example, sponge, ctenophore and placozoan genomes encode large repertoires of transcription factors (209–232) and chromatin modifiers and remodellers (99–134), representing intermediate diversity compared with unicellular species and other metazoans (for example, cnidarians or bilaterians) (Fig. 1). However, comparative analysis of genomic regulatory programmes in non-model organisms is confounded by the scarcity of direct molecular data on cell states and genome regulation. Whole-organism single-cell RNA sequencing (RNA-seq)[20,21] opens an opportunity to start closing this gap, by performing extensive sampling of transcriptional programmes and characterizing cell type repertoires in diverse metazoan lineages. Here, we generate transcriptional maps at single-cell resolution for *A. queenslandica*, *M. leidyi* and *T. adhaerens*. These maps, in combination with chromatin data and sequence analysis, allow us to survey the cell type diversity and compare the genomic regulatory programmes in these non-bilaterian animal lineages.

Results

An atlas of *A. queenslandica* adult and larval cell types. To study sponge cell type diversity, we collected adult and larval specimens

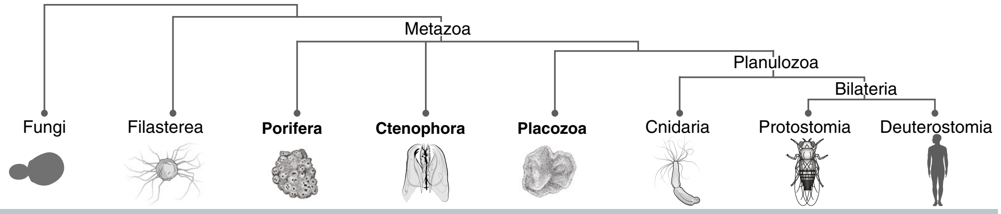
[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [2]Department of Biological Regulation, Weizmann Institute of Science, Rehovot, Israel. [3]Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway. [4]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. [5]School of Biological Sciences, University of Queensland, Brisbane, Queensland, Australia. Present address: [6]Department of Medicine, Weill Cornell Medicine and New York Genome Center, New York, NY, USA. *e-mail: arnau.sebe-pedros@weizmann.ac.il; amos.tanay@weizmann.ac.il

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

from *A. queenslandica*. We processed fresh cells using the massively parallel single-cell RNA-seq (MARS-seq) protocol with small adaptations[22] (see Methods), profiling a total of 4,992 adult and 3,840 larval *A. queenslandica* cells (Supplementary Fig. 1 and Supplementary Table 1). Whole-organism single-cell analysis involves processing cells with highly heterogeneous RNA content, given the expected differences in size and/or transcriptional activity between distinct cell types (Supplementary Fig. 1a,b). To maximize the sensitivity of our assay, we retained for subsequent analysis all sampled cells with at least 100 unique molecule identifiers (UMI). Applying the MetaCell framework (Supplementary Appendix 1), we found over 300 marker genes in each stage, which showed a high degree of intrapopulation transcriptional variance (Supplementary Fig. 1c). Using this approach, even cells with overall low UMI counts were characterized by a sufficient number of marker genes (Supplementary Fig. 1d). This allowed us to robustly group 81–94% of our single cells into transcriptionally coherent clusters, which we call metacells (Supplementary Fig. 1e,f; see also Methods and Supplementary Appendix 1), and to apply a bootstrap approach to support these metacells (Supplementary Fig. 1g; see also Methods). Moreover, we associated each of the derived metacells with a set of differentially expressed genes (Supplementary Tables 2 and 3) and used the functional annotation of these gene sets to annotate at least some of the metacells.

The power of whole-organism single-cell RNA-seq analysis to characterize cell types is demonstrated by visualizing *A. queenslandica* adult metacells (Fig. 2a), key marker genes in two-dimensional (2D) projection (Fig. 1b) and a heat map showing the distribution of marker genes at single-cell resolution (Fig. 2c). The sponge transcriptional landscape is dominated by large groups of choanocytes, pinacocytes and archaeocytes[13]. Even though these groups can be further subdivided into subclasses, their annotation into broad types is supported by common transcriptional signatures of key genes. Choanocytes are autonomous filter-feeding cells with a unique morphology, characterized by a flagellum surrounded by a microvilli collar[23]. Our data show that *A. queenslandica* choanocytes express RNA-binding proteins such as *mbnl*, *bruno2* and *nanos*, as well as multiple proteins of the flagellar apparatus, and annexins[24] (Fig. 2b and Supplementary Fig. 2b,h). They also specifically express multiple adhesion proteins, including cadherins and C-type lectins (Fig. 2b). Interestingly, not only choanocytes, but also other cell types we identified express unique combinations of adhesion proteins; for example, distinct integrin alpha/beta paralogue pairs (Supplementary Fig. 2a). These cell-type-specific adhesion molecules, especially those like cadherins and immunoglobulins that mediate homophilic interactions, are likely to be important in the spatial sorting of cell types and general sponge body plan organization. Finally, based on their expression, we can define two broad types of choanocytes (Fig. 2a) showing differences not only in their repertoire of effector genes but also in the expression of transcription factors.

Another abundant group of cells are pinacocytes (Fig. 2a,c). Pinacocytes are epidermal cells that cover the outer and inner surfaces of the sponge[13]. Our data show that *A. queenslandica* pinacocytes specifically express *pumilio* RNA-binding protein and multiple components of the actin contractility apparatus, including *tropomyosin*, *calponin* and *striated-type myosin II* (Fig. 2b and Supplementary Fig. 2a). This suggests that *A. queenslandica* pinacoderm has some contractile properties, as also indicated by experiments in the demosponge *Tethya wilhelma*[25]. Interestingly, we also identify a cluster of cells that show intermediate transcriptional profiles between choanocytes and pinacocytes, expressing both choanocyte markers, such as *fgf* and *bruno2* and pinacocyte



	Schizosaccharomcyes pombe	*Capsaspora owczarzaki*	***Amphimedon queenslandica***	***Mnemiopsis leidyi***	***Trichoplax adhaerens***	*Nematostella vectensis*	*Drosophila melanogaster*	*Homo sapiens*
Genome size (Mb)	12.6	28.0	166.7	155.9	98.0	356.6	143.7	3,095
Annotated genes	5,118	8,741	40,122	16,548	11,520	27,273	13,917	20,346
Intergenic distance Q10 (bp)	57	13	46	287	733	192	234	956
Intergenic distance Q50 (bp)	349	205	558	1,990	2,657	2,116	1,296	19,175
Intergenic distance Q90 (bp)	1,471	1,292	5,215	8,806	10,523	10,707	9,896	202,552
Intron median size	55	103	68	275	133	436	68	1,709
Exonic sequence (%)	56.69	61.00	24.45	14.00	14.73	6.94	15.39	1.09
Number of orphan genes[a]	191	2,259	7,394	4,582	1,595	4,284	292	211
Number of TFs	175	143	232	281	209	579	497	1,012
Number of chromatin modifiers	74	116	99	109	134	166	121	180
Number of RNA binding proteins	236	273	279	207	158	248	214	313
CTCF	–	–	–	–	–	–	+	+
DNA methylation[b]	–	–	+	+	–	+	–	+
Polycomb 2 complex	–	–	+	+	+	+	+	+

Fig. 1 | Comparison of genomic features of early metazoans and phylogenetically related species. Lineages and species sampled in this study are highlighted in bold. [a]Number of orphan genes based on Ensembl, except for *C. owczarzaki* (based on ref. [49]). [b]Presence/absence of DNA methylation in species without methylation data based on presence/absence of Dnmt1/3 orthologues. TF, transcription factor. Silhouettes were obtained from previous publications[42,74,75].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

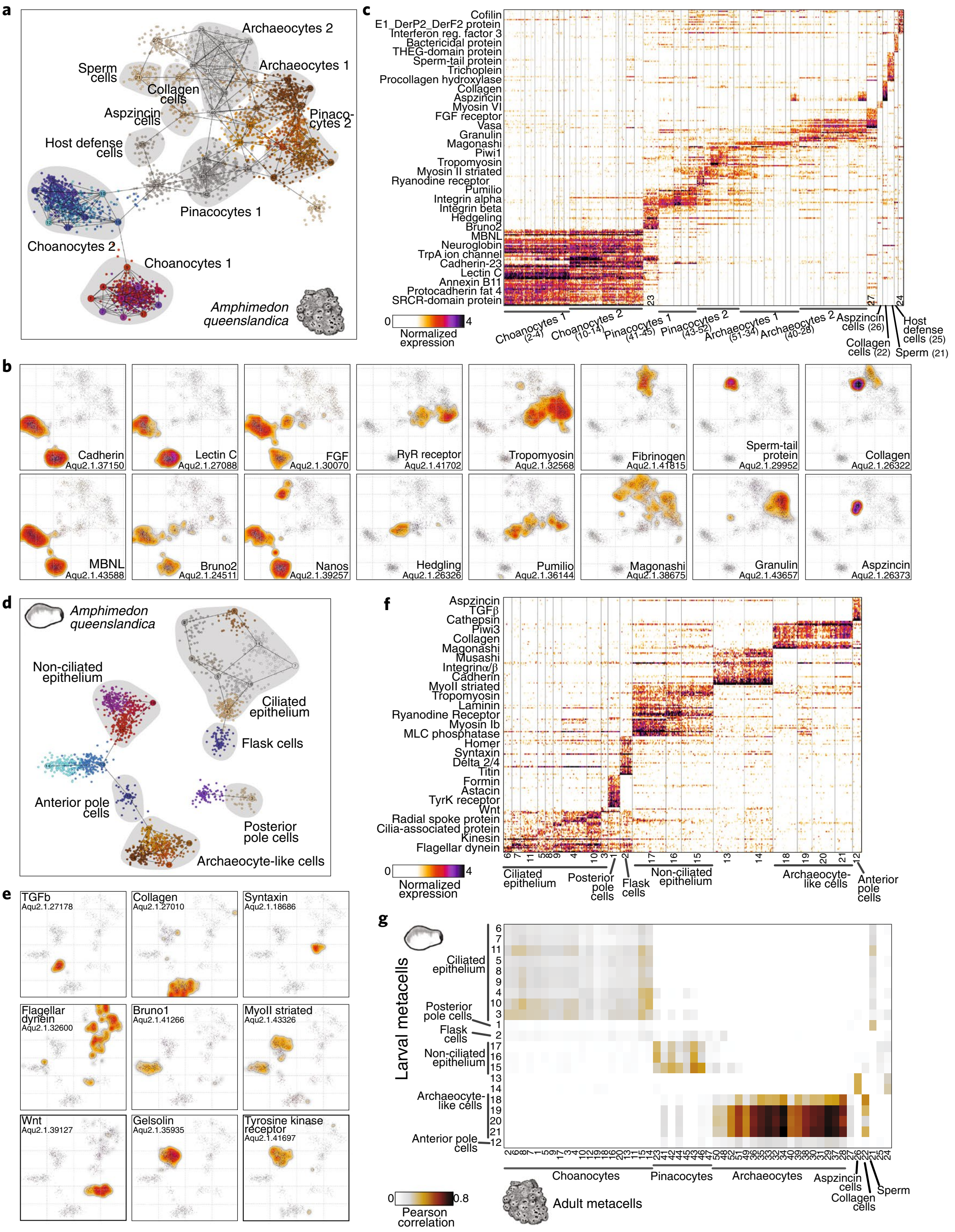


Fig. 2 | *A. queenslandica* adult and larval cell type atlases. a, 2D projection of *A. queenslandica* adult metacells and single cells. Cell clusters with known or hypothesized identity are annotated and highlighted in grey. **b**, Gene expression distribution on 2D projected *A. queenslandica* adult cells for selected gene markers. Gene identifiers from ref. [16]. The 2D cell projection is the same as in **a**. **c**, Normalized gene expression across 3,870 *A. queenslandica* adult single cells (columns), sorted by metacell. Metacell numbers are indicated in brackets. For each cluster, the top 25 genes sorted by fold change versus the other metacells were selected for visualization (with a fold-change threshold of ≥2). **d**, 2D projection of *A. queenslandica* larval metacells and single cells. **e**, Gene expression distribution on 2D projected *A. queenslandica* larval cells for selected gene markers. The 2D cell projection is the same as in **d. f**, Normalized gene expression across 1,932 *A. queenslandica* larval single cells (columns), sorted by metacell. Metacell numbers are indicated in brackets. Genes were selected as in **c**. **g**, Comparison of adult versus larval cell clusters. The heat map shows the correlation values between metacells based on highly variable genes (fold change > 2 in at least 1 adult and 1 larval metacell). Notice the strong association between adult archaeocytes and a group of larval cells, suggesting the re-usage of this specific cell type programme in two different post-embryonic stages. The colour-coding of cells and metacells in **a** and **d** is arbitrary.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

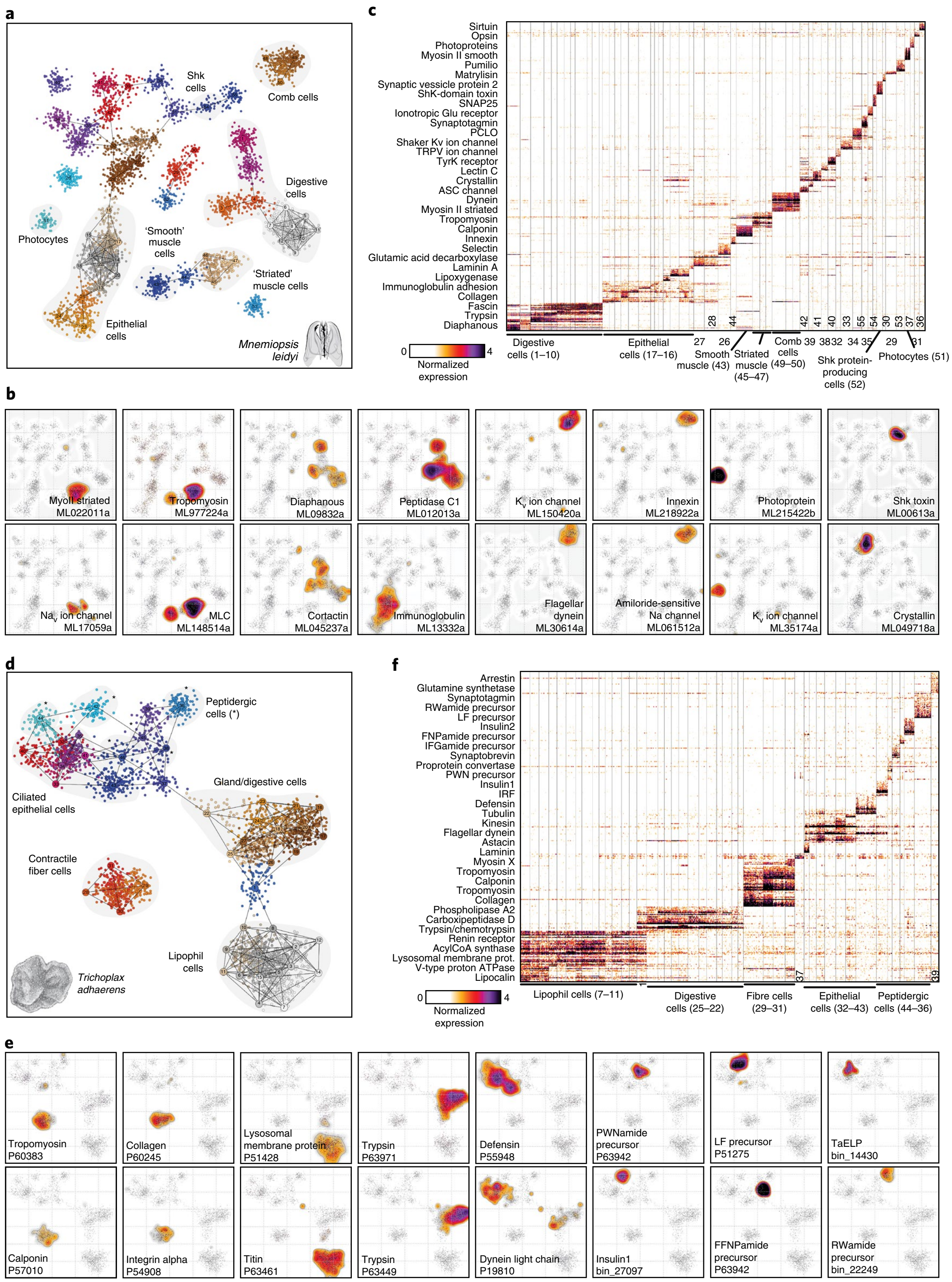


Fig. 3 | *M. leidyi* and *T. adhaerens* cell type atlases. a, 2D projection of *M. leidyi* metacells and single cells. Cell clusters with known or hypothesized identity are annotated and highlighted in grey. **b**, Gene expression distribution on 2D projected *M. leidyi* cells for selected gene markers. Gene identifiers from ref. [17]. The 2D cell projection is the same as in **a**. **c**, Normalized gene expression across 4,803 *M. leidyi* single cells (columns), sorted by metacell. Metacell numbers are indicated. For each cluster, the top 25 genes sorted by fold change versus the other metacells were selected for visualization (with a fold-change threshold of ≥2). **d**, 2D projection of *T. adhaerens* metacells and single cells. **e**, Gene expression distribution on 2D projected *T. adhaerens* cells for selected gene markers. Gene identifiers from ref. [18]. The 2D cell projection is the same as in **d**. **f**, Normalized gene expression across 3,209 *T. adhaerens* single cells (columns), sorted by metacell. Metacell numbers are indicated. Genes were selected as in **c**. The colour-coding of cells and metacells in **a** and **d** is arbitrary.

markers, such as *pumilio* (Fig. 2b). In addition, these cells specifically express *hedgling* (Fig. 2b), a cadherin with an amino-terminal hedgehog domain[26,27]. These data suggest the existence of transcriptional states representing trans-differentiation intermediates between cell types, a process known to occur in multiple sponge species, including *A. queenslandica*[13,28].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

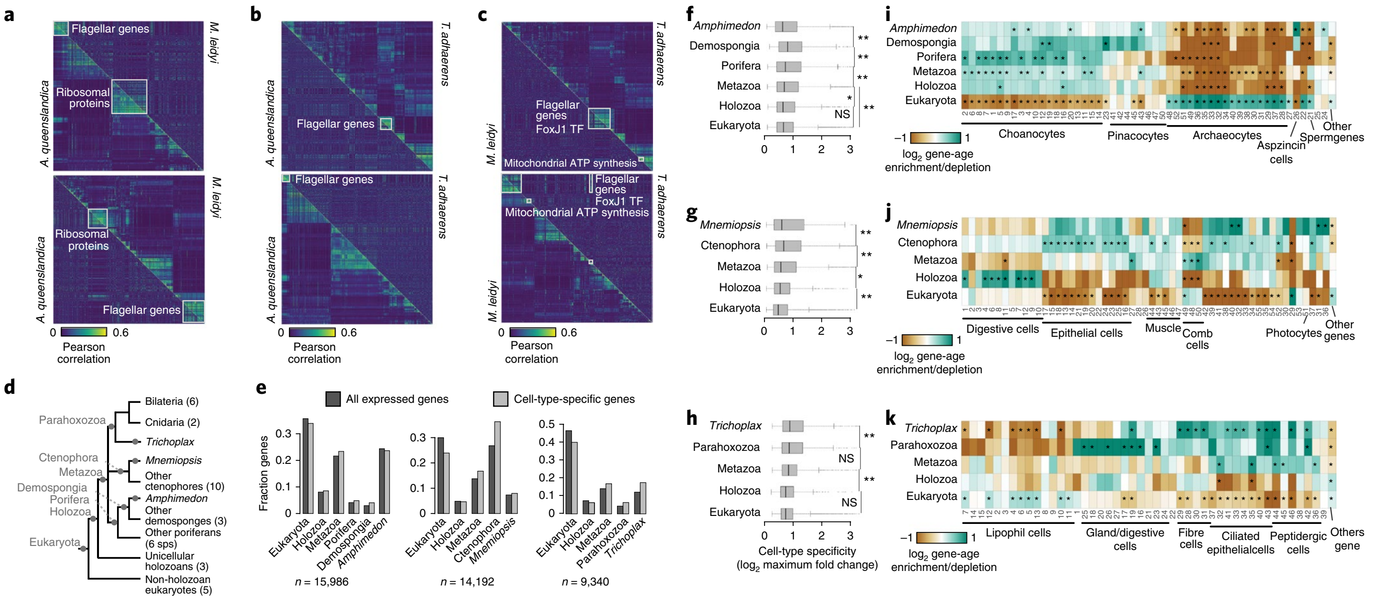


Fig. 4 | Phylogenetic patterns of cell-type-specific gene repertoires. a, Cross-species gene module analysis. Top: heat map showing the gene–gene correlation values for *A. queenslandica* gene modules (bottom triangle) compared with the gene–gene correlation values for gene orthologues in *M. leidyi* (sorted based on the clustering of *A. queenslandica* genes; top triangle). Bottom: reciprocal analysis, focusing on *M. leidyi* gene modules (top triangle) and showing the equivalent correlations for *A. queenslandica* orthologues (bottom triangle). Correlation values were computed based on expression profiles across metacells, and genes are hierarchically clustered based on these correlations in the species of focus. Conserved gene modules are highlighted with white squares. **b**, Same as **a** for *A. queenslandica* versus *T. adhaerens* comparisons. **c**, Same as **a** for *M. leidyi* versus *T. adhaerens* comparisons. Notice the highly conserved flagellar toolkit gene module, found in all pairwise comparisons. Interestingly, this module is associated with FoxJ1 transcription factor, both in *M. leidyi* and *T. adhaerens*. This is known to also be a ciliary regulator in bilaterians[76]. **d**, Schematic phylogenetic tree showing the lineages and number of species (in brackets) used in our orthology analysis. The gene-age categories derived from this analysis are shown in grey in the corresponding branches. **e**, Gene-age distributions in each species for all detected genes (dark grey) and cell-type-specific genes (light grey; genes with a maximum fold change of >2 in at least one metacell) for *A. queenslandica* (left), *M. leidyi* (middle) and *T. adhaerens* (right). **f**, *A. queenslandica* gene cell type specificity (calculated as the $\log_2$ maximum fold change across metacells) stratified by gene age. **g**,**h**, Same as **f**, but for *M. leidyi* (**g**) and *T. adhaerens* (**h**). Note that in all cases, cell type specificity is higher in evolutionarily younger genes, with a drop in orphan and species-specific genes in the case of *A. queenslandica* and *M. leidyi*. This is in line with previous observations suggesting that gene innovations tend to be associated with tissue and cell-type-specific functions[48,49]. $^{**}P<0.0001$, $^{*}P<0.05$, NS, not significant (Wilcoxon rank-sum test). **i**, Gene-age frequency enrichment (green) and depletion (brown) in gene sets specific to each *A. queenslandica* metacell. The enrichment and depletion is represented as the $\log_2$ of the frequency of gene ages (among the genes overexpressed in each metacell) versus the background frequency of gene ages (taking into account only detected genes, rather than all predicted genes). *Q value < 0.01 (χ^2 test with Benjamini–Hochberg correction). For example, choanocytes are strongly enriched in genes of poriferan origin and, to a lesser extent, of metazoan origin. By contrast, archaeocytes and sperm cells are strongly enriched in ancient, paneukaryotic genes. **j**,**k**, Same as **i**, but for *M. leidyi* (**j**) and *T. adhaerens* (**k**) metacells.

Another major sponge cell behaviour identified here corresponds to archaeocytes, which are pluripotent amoeboid cells found in the sponge mesohyl (the gelatinous matrix that fills the sponge body)[29]. We find that these cells express specific extracellular matrix proteins (such as, *fibrinogen*), granulins and large amounts of diverse RNA-binding proteins (such as, *magonashi*) (Fig. 2b and Supplementary Fig. 2b,c). The extensive usage of cell-type-specific RNA-binding proteins observed chiefly in archaeocytes, but also in other sponge cell types (Fig. 2b and Supplementary Fig. 2b), is in line with previous reports that suggest a pervasive role of this type of regulators in another sponge species, *Ephydatia fluviatilis*[30]. In addition to these abundant cell types, we detect in adult *A. queenslandica* remarkably distinct, yet much less abundant, cell types. These include sperm cells, defined by expression of *tprv* ion channel, *theg* and other genes associated with sperm function (Fig. 2b and Supplementary Fig. 2d), as well as collagen-producing cells (Fig. 1b), cells expressing multiple *aspzincin* protease paralogues (Fig. 2b and Supplementary Fig. 2e) and host defence cells producing antibacterial proteins (Supplementary Fig. 2f).

Unlike the other species included in this study, but similar to many marine invertebrates[31], *A. queenslandica* has a biphasic life cycle involving two dramatically different post-embryonic stages: adult and larva[32]. We therefore profiled single-cell transcriptomes in the lecitotrophic larva of *A. queenslandica* to identify larval cell types and compare them with those found in adult sponges. We sampled the transcriptomes of 3,840 larval single cells and identified metacells with specific expression signatures using the same strategy described for the adult (Fig. 2d,e and Supplementary Table 2). This analysis revealed at least seven different cell types in the larva (Fig. 2d,e). Based on published expression patterns for marker genes, we could identify some of these cell types. These include ciliated epithelial cells that express ciliary markers (Fig. 2e), flask cells[33], *wnt*-expressing posterior pole cells[34] and *tgfb*-expressing anterior pole cells[34]. When comparing transcriptional signatures, larval cell types show remarkable differences compared with adult cell types: 4.8% of the genes expressed in the larva (689/14,426) are not expressed in the adult and, reciprocally, 39.9% (9,010/22,567) of adult genes are not expressed in the larva. Direct metacell comparisons (Fig. 2g) show that, in fact, only one larval cell type shows strong similarity with an adult cell type: archaeocytes. Overall, this indicates that the *A. queenslandica* larval stage deploys a unique set of cell

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

behaviours with no counterparts in the cell types that emerge after the larva metamorphoses into an adult[28].

***M. leidyi* cell type diversity**. Ctenophores were traditionally considered to be a sister group to cnidarians[35]. However, recent phylogenomics studies clearly show they are one of the earliest-branching animal lineages, although it remains disputed whether they branched before or after sponges[3–6] (Fig. 1). Ctenophores have a complex body plan and cell types such as muscles and neurons. These features, together with the ctenophore phylogenetic position, open the question of whether neurons and other cell types have single or multiple origins within Metazoa[11,12,17,36]. We mapped the diversity of cell types in the ctenophore *M. leidyi* by profiling 6,144 single-cell transcriptomes. Compared with the sponge, mapping of the ctenophore *M. leidyi* transcriptional states uncovered a richer repertoire of cell types, some of which could be associated with putative functions and known cell types (Fig. 3a–c, Supplementary Fig. 3 and Supplementary Table 4). For example, we identified a group of photocyte cells (the cells responsible for ctenophore bioluminescence) expressing known photoproteins and opsins[37] (Fig. 3b). Unlike most other metazoans, ctenophore locomotion is based on the coordinated ciliary beating of rows of comb cells. We identified comb cells expressing multiple ciliary markers and specific potassium voltage-gated and amiloride-sensitive sodium ion channels (Fig. 3e and Supplementary Fig. 3). Comb cells also express a specific *innexin* gene (Fig. 3e), supporting the existence of gap junctions electrically coupling these groups of cells, as suggested by ultrastructural observations[38]. Another group of cells show expression of markers associated with muscle cell types in other species[39], such as *tropomyosin* and *myosin light chain* (Fig. 3b). Interestingly, although *M. leidyi* lacks striated muscles, we can distinguish a group of muscle cells expressing markers associated with striated muscles in other species[39], such as *striated-type myosin II*, while another group of muscle cells express markers of 'smooth' muscles, such as *calponin* (Fig. 3b and Supplementary Fig. 3a). We also detect cells showing expression of digestive enzymes and genes associated with the formation of microvilli and filopodia[40] (such as *diaphanous* and *cortactin*) (Fig. 3b), a group of cells expressing a secreted Shk-domain protein[41] (Fig. 3b), and epithelial cells expressing multiple transmembrane adhesion and extracellular matrix proteins (Supplementary Fig. 3b).

However, most of the cell clusters we identified cannot be assigned to known functions/types and many are strongly associated with unannotated proteins (Supplementary Table 3), often Ctenophora-specific (see Fig. 4e). This emphasizes our still very limited understanding of ctenophore biology[9]. Interestingly, we could not identify any metacell with distinct neuronal gene expression signatures such as those observed in cnidarians and bilaterians[36]. For example, different synaptic scaffold components are expressed across multiple cell types and no specific cell cluster shows co-expression of many voltage-gated ion channels. This lack of co-expression is similar to that observed for synaptic scaffold and other neuronal genes in *A. queenslandica* and *T. adherens* (see below)—two organisms without neuronal cells. Instead, we find in *M. leidyi* highly specific expression in multiple metacells of electrical synapse components (innexins), as well as specific expression of ASC, iGluR and K_v/Ca_v/Na_v ion channels[12,17] (Fig. 3b and Supplementary Fig. 3c–h). Overall, these findings indicate a dramatically different molecular composition of ctenophore synapses and neuronal-like cells from those of cnidarians and bilaterians, possibly suggesting convergence of these cell types[12,42].

***T. adhaerens* cell type diversity**. Placozoans are the simplest (non-parasitic) multicellular animals. They have no apparent body axis or tissue-level organization and they differentiate only six cell types according to ultrastructural studies[14,43]. These cells are organized in two ciliated epithelial layers and the flattened body is filled with extracellular matrix material and fibre cells. We dissociated and sampled the transcriptomes of 4,608 *T. adhaerens* cells (Fig. 3d–f, Supplementary Fig. 4 and Supplementary Table 5) and defined metacells and putative cell types using the same strategy as for *A. queenslandica* and *M. leidyi*. In line with the known biology and ultrastructure of *T. adhaerens*[43], we defined groups of fibre cells, lipophil cells, digestive and gland cells, and epithelial cells, comprising in total 79% of the sampled cells. Fibre cells express markers associated with cell contractility, such as *tropomyosin* and *calponin* (Fig. 3e and Supplementary Fig. 4g), as well as cell adhesion and extracellular matrix proteins such as integrins, collagens and fibronectins (Fig. 3e and Supplementary Fig. 4c,g). This suggests a dual role of these cells in generating the extracellular material that fills the body of *T. adhaerens*, as well as enabling the body contraction involved, for example, in placozoan feeding behaviour. Lipophil cells express multiple lysosome and lipid metabolism genes (Fig. 3e and Supplementary Fig. 4d), gland cells express different digestive enzymes such as trypsins (Fig. 3e and Supplementary Fig. 4h), and epithelial cells express multiple defensins—short peptides involved in host defence[44] (Fig. 3e and Supplementary Fig. 4e). Both gland and epithelial cells express ciliary markers (Fig. 3e and Supplementary Fig. 4f), as expected given that they are both ciliated cell types[43].

Besides these four abundant cell behaviours, our analysis reveals seven additional lower-frequency cell types, six of which are characterized by the production of unique regulatory peptides[45,46] and multiple specific transcription factors (Figs. 3e,5f). One of these regulatory peptides (*TaELP*; Fig. 3e) has recently been shown to regulate *T. adhaerens* locomotion through control of ciliary beating of the cells in the lower epithelial layer[45]. Therefore, we hypothesize that the five other peptidergic cell types uncovered in this study may be involved in the control of additional processes, such as the release of digestive enzymes from gland cells or the contraction of fibre cells. However, although the *T. adhaerens* genome encodes multiple genes involved in synaptic and neuronal functions[18], these genes do not show co-expression in these peptidergic cell types (Supplementary Fig. 4b), indicating the absence of a synaptic scaffold or any other neuronal gene module. Overall, the observed states indicate that elaborated peptidic regulation occurs in this simple animal within specialized cell types that lack the characteristics of synaptic neurons[45].

Phylogenetic patterns of cell-type-specific gene repertoires. To study the evolutionary dynamics of these cell-type-specific transcriptional programmes, we used phylogenetic mapping to define gene ages and orthology relationships in *A. queenslandica*, *M. leidyi* and *T. adhaerens* (Supplementary Table 6). First, we analysed the possible cross-species conservation of cell-type-specific expression correlation over orthologous gene pairs. This showed that, at the evolutionary distances separating these three species from their common ancestor (>635 Ma[8]), co-regulation of genes is almost completely divergent (Fig. 4a–c). In fact, we only observed conserved co-regulation of specific housekeeping functions, including ribosomal proteins and flagellar apparatus.

Next, we analysed how gene age correlates with cell type transcriptional specificity (Fig. 4d–k). We defined for each gene in each species an inferred evolutionary origin based on the presence of orthologues in species belonging to key taxonomic groups[4,12,47] (Fig. 4d and Supplementary Table 6). The global age distribution among expressed genes varied substantially across species (Fig. 4e). In *A. queenslandica*, most expressed genes are of eukaryotic origin (36%), followed by genes originated at the stem of Metazoa (23%) and *A. queenslandica*-specific genes (24%). In *T. adhaerens*, paneukaryotic genes are even more dominant, representing over 50% of all expressed genes, and a similar percentage of the genes that are expressed in a cell-specific manner; while there is only a modest

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

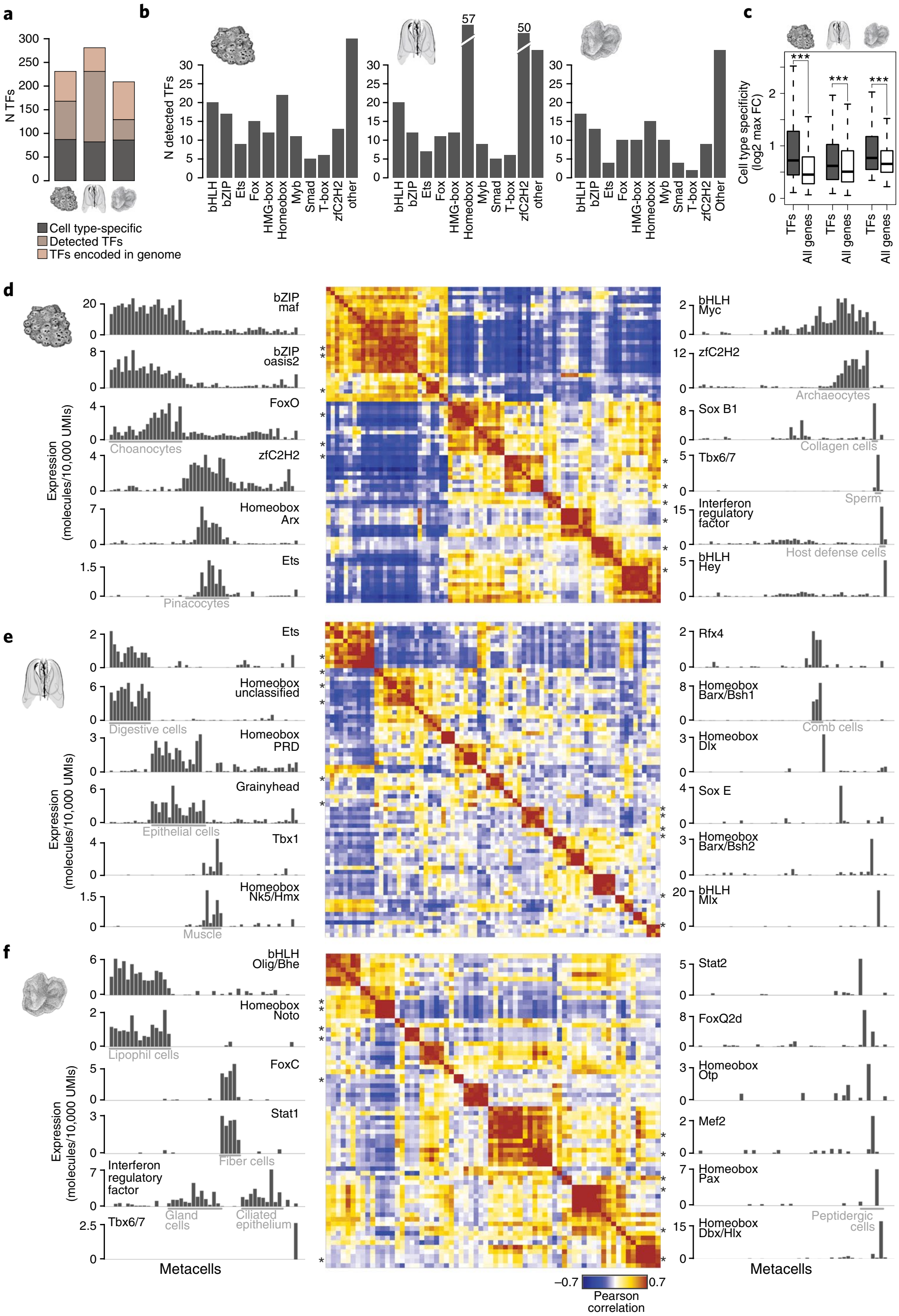


Fig. 5 | Transcription factor regulatory programmes in *A. queenslandica*, *M. leidyi* and *T. adhaerens*. a, Number of transcription factors encoded in the genome, detected in our single-cell RNA-seq analysis and showing cell-type-specific expression in each species. **b**, Number of transcription factors belonging to different structural classes detected in *A. queenslandica* (left), *M. leidyi* (middle) and *T. adhaerens* (right). **c**, Cell type specificity of transcription factors compared with all genes for each species. Cell-type specificity of each gene was measured as the maximum fold-change enrichment of its expression in any metacell. The midline indicates the median, the box ranges from the first to third quantile, the whiskers are the 1.5 times the interquartile range. **$P < 0.0001$ (Wilcoxon rank-sum test). **d**, Heat map (centre) showing *A. queenslandica* correlation between transcription factors based on expression profiles across metacells. Only transcription factors with >20 total molecules and a fold change of >1.8 in at least one metacell are included. On both sides, bar plots show the expression profile across metacells for representative transcription factors in each transcription factor module. Asterisks indicate the position of the transcription factors shown in bar plots in the heat map (in the same descending order). **e**,**f**, Same as **d**, but for *M. leidyi* (**e**) and *T. adhaerens* (**f**).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

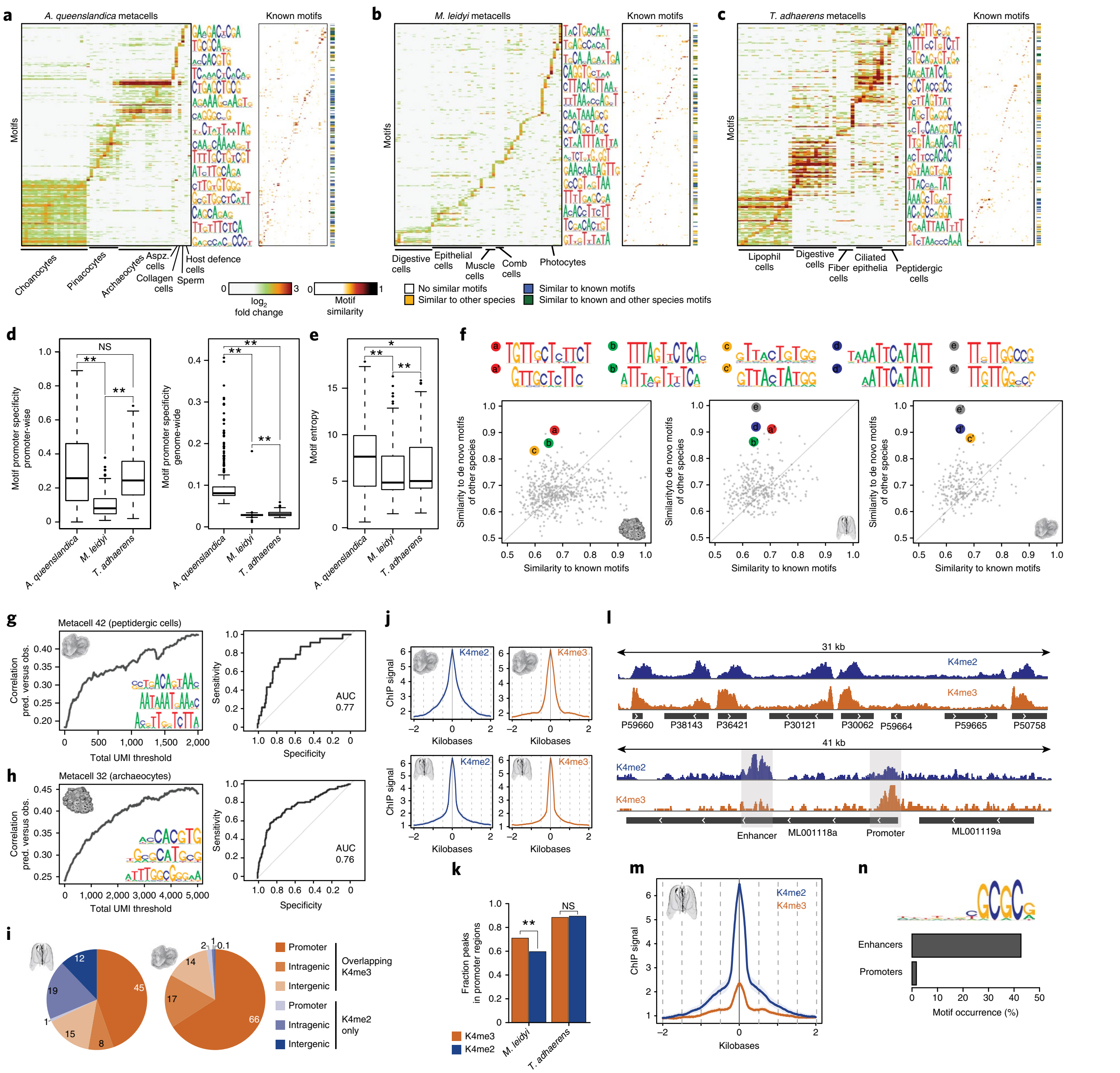


Fig. 6 | Regulatory sequence analysis in *A. queenslandica*, *M. leidyi* and *T. adhaerens*. a, De novo motif enrichments in *A. queenslandica* promoters. Left: heat map showing significant (FDR < 0.02) motif (rows) enrichments in the promoters of metacell-specific gene sets (columns). Right: heat map showing the similarity of each *A. queenslandica* promoter-enriched motif (rows) to known motifs in databases (columns). The coloured bar indicates whether the motif has high similarity (>0.7) with any known motifs and/or de novo motifs found in the other two species. **b**,**c**, Same as **a**, but for *M. leidyi* (**b**) and *T. adhaerens* (**c**). **d**, Boxplots showing, for each species, the frequency of occurrence of metacell-specific motifs in the promoters of metacell-specific genes compared with all other gene promoters (left) and compared with the whole genome (right). $^{**}P < 0.0001$, $^{*}P < 0.05$, NS, not significant (Wilcoxon rank-sum test). **e**, Boxplot showing, for each species, the distribution of de novo motif entropies. $^{**}P < 0.0001$, $^{*}P < 0.05$ (Wilcoxon rank-sum test). **f**, Scatterplots for *A. queenslandica* (left), *M. leidyi* (middle) and *T. adhaerens* (right), showing the maximum similarity of each de novo motif to known motifs (*x* axis) and motifs in the other two species (*y* axis). The highlighted cases (top) show examples of motifs that are highly similar between two species and not similar to any known motif in databases. **g**, Left: correlation between observed (obs.) and predicted (pred.) expression values derived from a linear model based on promoter motif content analysis for the *T. adhaerens* metacell 42 (peptidergic cells). Correlation is shown as a function of the total molecule count threshold applied to the genes considered in the analysis. The three motifs with the top coefficients according to the model are shown. Right: receiver operating characteristic curve of the linear regression model predicting gene expression in metacell 42 (peptidergic cells). **h**, Same as **g**, but for *A. queenslandica* metacell 32 (archaeocytes). **i**, Pie charts for *M. leidyi* (left) and *T. adhaerens* (right) showing the distribution of H3K4me2 peaks across different genomic features, grouped by overlap or lack of overlap with H3K4me3 peaks. H3K4me3 + K4me2 peaks in non-promoter regions are likely to represent unannotated promoter sites. Numbers indicate the percentage for each category. **j**, iChIP signal metaplots centred in promoter peak maximum coverage positions for H3K4me3 (left) and H3K4me2 (right) and in *T. adhaerens* (top) and *M. leidyi* (bottom). ChIP signal is indicated as $-\log_2(1 - \text{coverage quantile})$, see Methods. **k**, Fraction of H3K4me2/3 peaks observed in promoter regions in *M. leidyi* (left) and *T. adhaerens* (right). $^{**}P < 0.0001$, NS, not significant (χ^2 test). **l**, Example *T. adhaerens* (top) and *M. leidyi* (bottom) genomic regions showing normalized H3K4me2 and H3K4me3 iChIP coverage. **m**, *M. leidyi* H3K4me2/3 iChIP signal metaplots centred at enhancer element maximum H3K4me2 positions. **n**, De novo motif enriched in *M. leidyi* enhancers. The bar plot shows the frequency of occurrence of this motif in *M. leidyi* enhancers and promoters.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

contribution of genes specific to *T. adherens* (17%) in the cell-type-specific transcriptomes. By contrast with *A. queenslandica* and *T. adherens*, in the ctenophore *M. leidyi*, most cell-type-specific genes are of ctenophore origin (40%). This suggests an important contribution of ctenophore gene innovations to ctenophore cell type biology[9] and also explains the difficulty of determining the identity of many of the cell clusters we identified in this species (Fig. 3a–c).

In general, genes that are expressed broadly across tissues have been shown to have older phylogenetic origins, while genes expressed in a narrower subset of tissues tend to have more recent phylogenetic origins[48,49]. To test whether the same effect is observed in cell type transcriptomes, we defined for each gene a cell type specificity score (based on the maximum fold change in expression observed in any metacell) and stratified these values according to gene age (Fig. 4f,g). In all three species, we observed that evolutionarily more novel genes show a significantly higher degree of cell-type-specific regulation. At a higher resolution, specific cell clusters show distinct gene-age distributions (Fig. 4i–k). For example, sponge choanocytes are particularly enriched in genes specific to the sponge lineage, whereas, archaeocytes and sperm cells are enriched in paneukaryotic genes (Fig. 4i). In the ctenophore, digestive cells are enriched in genes of holozoan origin (that is, shared between animals and their closest unicellular relatives), while epithelial cells and multiple uncharacterized cell types are enriched in ctenophore genes (Fig. 4j). A similar pattern is observed in the placozoan *T. adherens*, with epithelial cells being enriched in lineage-specific genes, while lipophil cells are enriched in paneukaryotic genes and digestive cells are enriched in genes shared between placozoans, cnidarian and bilaterians (ParaHoxozoa) (Fig. 4k).

Cell-type-specific transcription factor modules. Transcription factors are key players in the gene regulatory networks that define cell type identity[50]. We examined transcription factor cell-type-specific expression to test whether the observed cell type transcriptional programmes are linked to a rich transcription factor repertoire. We detected expression for 168, 231 and 129 predicted transcription factors in *A. queenslandica*, *M. leidyi* and *T. adhaerens*, respectively (Fig. 5a and Supplementary Fig. 5a). The classification of predicted transcription factors into structural classes suggested expanded usage of homeobox and zf-C2H2 transcription factors in the ctenophore, but otherwise similar representation of transcription factor classes between these species (Fig. 5b). Consistent with their probable role as key drivers of cell type regulation, we found that transcription factors are much more likely to be expressed in a cell-type-specific fashion compared with all other genes (Fig. 5c). Accordingly, we found different transcription factors being specifically expressed in all cell types in each of the species. In *A. queenslandica*, we observed *maf*, *grainyhead* and 27 other transcription factors enriched in choanocytes; *ets* and *arx* homeobox are specific to pinacocytes; and Myc*myc* is expressed in archaeocytes (Fig. 5d and Supplementary Fig. 5b,c). Less frequent sponge cell types also show highly specific transcription factor expression. For example, sperm cells show co-expression of four *tbx6/7* paralogues, and host defence cells express *interferon regulatory factor* (Fig. 5d and Supplementary Fig. 5b,c). In *M. leidyi*, *grainyhead* transcription factor is enriched in epithelial cells and *rfx4* is enriched in the ciliated comb cells (Fig. 5e and Supplementary Fig. 6a). These transcription factors have been shown in other species to be expressed in epithelial cells and ciliated cells, respectively[51,52], suggesting conserved association of these transcription factors with epithelial and ciliary programmes. Examples of cell-type-specific transcription factor regulators in *T. adhaerens* include *noto* homeobox in lipohil cells and *foxC* in fibre cells (Fig. 5f and Supplementary Fig. 6b). Interestingly, while an overall similar number of transcription factors are expressed in a cell-type-specific fashion across the three species (Fig. 2a), in the ctenophore, the higher cell type complexity results in a smaller number of transcription factors linked to each transcriptional state, suggesting that additional epigenetic mechanisms might be involved in cell type specification for this species; for example, genomic compartmentalization and combinatorial gene regulation by distal regulatory elements. In summary, elaborated combinatorial expression of transcription factors is observed to correlate—and possibly drive—differentiated transcriptional programmes in sponges, ctenophores and placozoans.

Genomic embedding of cell type regulatory programmes in early metazoans. Transcription factors regulate their target genes by binding to sequence elements located at promoters and, most prominently in bilaterians, at distal enhancers. To reconstruct the degree to which information encoded into gene promoters can direct cell-type-specific transcriptional control in early metazoans, we defined sets of cell-type-specific gene modules for each species (Supplementary Tables 2–5). We then searched de novo for enriched sequence motifs in predicted gene promoters (−200 and +50 base pairs (bp) from the transcription start site (TSS)), controlling for false discovery rate and validating motif robustness by analysis of spatial motif distributions (Supplementary Fig. 7a) and shifted control sequences (Supplementary Fig. 7b). In *A. queenslandica*, we selected 325 motifs for downstream analysis (Supplementary Table 7), computed promoter affinity to each motif and visualized the distribution of motif enrichments for each cell-type-specific gene module (Fig. 6a). This resulted in remarkably rich landscapes of promoter motif content, covering all inferred cell types with 16–96 distinct motifs. For example, we observed 93 distinct motifs enriched in choanocyte gene promoters, consistent with the exceptionally rich combination of 29 transcription factors associated with choanocyte-specific expression (Fig. 5d and Supplementary Fig. 5b,c). Similar analysis in *M. leidyi* (Fig. 6b; 6–82 motifs per cell type) and *T. adhaerens* (Fig. 6c; 29–98 motifs per cell type) confirmed that promoter motifs are significantly enriched in these organisms as well. However, comparative analysis of the degree of motif genomic specificity (Fig. 6d) and entropy (Fig. 6e) suggested that, in the ctenophore *M. leidyi*, the strength of promoter motifs and their specificity to target genes given multiple potential genomic off-targets is significantly weaker compared with *A. queenslandica* and *T. adhaerens*.

Our de novo discovery approach is a priori not restricted to the identification of known transcription factor binding motifs characterized in model species. Nevertheless, we found that 33% of *A. queenslandica*, 25% of *M. leidyi* and 32% of *T. adhaerens* motifs matched (similarity >0.7) known models retrieved from databases covering transcription factor motifs for multiple eukaryotic species (Fig. 6a–c and Supplementary Fig. 7f). This indicates that at least some of the sequence elements defining the transcription-factor–genome interface are deeply evolutionarily conserved. Remarkably, out of the 570 novel motifs that could not be matched in databases, we detected 53 conserved between at least two species (Fig. 6f and Supplementary Fig. 7g). Discovering novel motifs independently in highly diverged species serves as further validation of the robustness of the promoter signals we characterize and indicates that comprehensive characterization of the repertoire of possible transcription-factor–DNA interfaces in metazoan genomes will require further analysis of phylogenetically diverse species.

Analysis of promoter information content by predictive expression models. In multicellular animals, stable differentiated transcriptional programmes are defined by multiple *cis*-regulatory modules, long-range control and powerful epigenetic mechanisms[53]. By contrast, in most unicellular eukaryotes, gene regulation involves exclusively regulatory elements that are proximal to the gene promoter[54]. Hence, we were surprised by the high degree of proximal promoter information content in *A. queenslandica* and

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

T. adhaerens. To further quantify this information content in cell-type-specific promoters, we implemented a simple model aiming to predict cell-type-specific expression from promoter sequences alone (see Methods). We tested the model by training on subsets of the genes and then predicting cell-type-specific gene expression from hidden promoter sequences. We found that this simple approach generated substantial predictive value in multiple *A. queenslandica* and *T. adhaerens* metacells (Fig. 6g,h and Supplementary Fig. 7c–e), despite the clear limitations of predicting combinatorial regulation using linear models. Accuracy improved as the total number of RNA molecules captured for a gene increased (Fig. 6g,h and Supplementary Fig. 7c–e), indicating that some of the inaccuracy of our predictions stems from experimental noise in the estimation of differential expression. For example, using promoter sequences alone, we could predict 50% of the *A. queenslandica* metacell 32 gene expression with 90% specificity (area under the curve (AUC) = 0.76) and 50% of the *T. adhaerens* metacell 42 gene expression with 84% specificity (AUC = 0.77). Interestingly, predictions based on promoter sequence were less powerful in the ctenophore (Supplementary Fig. 7d), suggesting an important contribution of additional, perhaps distal, regulatory elements in this group.

Characterizing distal epigenetically marked loci in *M. leidyi*. To test the potential contribution of long-range regulatory elements in *M. leidyi* and, as a control, *T. adhaerens*, we used indexing-first chromatin immunoprecipitation (iChIP)[55] in these two species. We profiled chromatin extracted from whole organisms with antibodies against histone modifications associated with promoter (H3K4me2/3) and enhancer (H3K4me2-only) activities. We found that whole-organism iChIP was sufficiently sensitive to detect H3K4me2/3 enrichment in 45% of *M. leidyi* and 66% of *T. adhaerens* promoters (Fig. 6i), showing quantitatively stronger enrichment for promoters that were expressed in a larger fraction of the cells (Supplementary Fig. 8a,b). Spatial analysis showed that H3K4me3 and H3K4me2 are localized around annotated promoters at a distance scale of less than 500 bp in both species (Fig. 6j). Interestingly, we found that while in *T. adhaerens* the fraction of H3K4me2 and H3K4me3 peaks mapping in promoter regions is the same (Fig. 6k), a significant fraction of H3K4me2 in *M. leidyi* does not co-localize with H3K4me3 in promoters, suggesting the existence of non-promoter distal regulatory elements[55]. Examples of epigenomic profiles (Fig. 6l and Supplementary Fig. 8c,d) and spatial mapping around distal H3K4me2 in the ctenophore (Fig. 6m) both support the existence of a distinct class of distal epigenetically marked loci in this species. Furthermore, sequence analysis revealed that these loci are 20-fold enriched for a specific GCGC-rich motif compared with promoters (fivefold compared with the genomic background) (Fig. 6n and Supplementary Fig. 8e,f). The strong chromatin signature we observe in whole-organism iChIP for this class of distal elements and the strong sequence specificity observed within it suggest that this class represents some constitutively active genomic-structural elements. Such elements may be hypothesized to perform functions that are similar to the role of CTCF in vertebrates[56] or Beaf-32 in *Drosophila melanogaster*[57,58]. In summary, we discovered the existence of distal elements of *M. leidyi* with strong sequence specificity and a potential role as enhancers and/or chromosomal organizers. Similar analysis could not detect any evidence for distal regulatory elements in *T. adhaerens*.

Discussion

Using whole-organism single-cell RNA-seq and a combination of sequence and chromatin analysis, we mapped differentiated transcriptional states and linked them with putative cell types in three representatives of the earliest-branching animal lineages. The unbiased approach we employed provides the first systematic insight into early animal cell type regulatory programmes, revealing distinct cell type repertoires in adult and larval sponges, a surprisingly high diversity of cell types in *M. leidyi*, and the existence of multiple specialized peptidergic cell types in *T. adhaerens*. A combination of these cell type transcriptional atlases with chromatin and sequence analyses indicates the existence of some key differences between sponge, placozoan and ctenophore cell-type-specific transcriptional control schemes. On the one hand, *A. queenslandica* and *T. adhaerens* have fewer cell types and show remarkably specific promoter sequence motifs. Moreover, *T. adhaerens* shows no evidence of regulation by distal enhancer elements. On the other hand, *M. leidyi* has higher cell type diversity, expresses fewer specific transcription factors per cell type, and shows lower information content in gene promoters. Moreover, *M. leidyi* shows strong evidence for distal regulatory elements. We suggest that the ctenophore mechanistic solution for defining and stabilizing cell type programmes might be more similar to the bilaterian solution, employing multiple layers of control to supplement the transcription factor combinatorics. We hypothesize that this elaborate regulation might be necessary to specify large repertoires of cell types embedded in a complex body plan such as that of ctenophores. By contrast, placozoans demonstrate the feasibility of defining and regulating multiple cell types without such strong layered architecture, but simply using a combination of transcription factors and proximal promoter regulatory elements, similarly to what is observed in unicellular eukaryotes and unlike the animal species studied to date. We expect the methodology we introduce here will facilitate multiple studies for mapping cell type regulation in diverse species in the coming years, resulting in increasingly dense phylogenetic coverage of cellular behaviours across the animal tree of life. The integrative analysis of this data will further allow a comprehensive and principled analysis of the evolutionary mechanisms leading to animal multicellularity and the genomic determinants of multifaceted transcriptional control schemes.

Methods

Animal sources, specimen dissociation and cell sorting. *A. queenslandica* adults and larvae were collected from Heron Island Reef, Great Barrier Reef, Queensland, Australia. Adult specimens were dissociated by placing them in a syringe and squeezing them through a 60 µm nylon mesh (fused to the end of the syringe) into calcium/magnesium-free seawater (CMFSW). Larvae were dissociated by gentle pipetting with gelatin-coated tips.

M. leidyi adults originated from L. Friis-Møller, Kristineberg, Sweden. They were maintained in the laboratory in filtered seawater, with small adult specimens (~20 mm) used for dissociation. Specimens were starved for 2–3 days, with daily changes of seawater. They were relaxed briefly in 7% magnesium chloride, then rinsed twice in CMFSW. For dissociation, they were incubated in 0.25% chymotrypsin (MP Biomedicals) in CMFSW for 20 min at room temperature with constant rocking and gentle pipetting. Cells were collected by centrifugation for 10 min at 1,000 *g* at 16 °C.

T. adhaerens (Grell strain[59]) was cultured in the laboratory at room temperature using artificial seawater (ASW) and feeding them with the cryptophyte algae *Pyrenomonas helgolandii* (strain SAG 28.87). Algae were obtained from the University of Gottingen algae culture collection (SAG), and cultured at room temperature in 250 ml flasks using PROV50 medium (#MKPROV50L; NCMA) and a long-wavelength fluorescent lamp. For dissociation, 30–40 animals were first transferred to a small plastic dish and, after they attached, cleaned three times with ASW. Then, ASW was replaced by CMFSW plus 10 mM ethylenediaminetetraacetic acid (EDTA) and the animals were dissociated by gentle pipetting with gelatin-coated tips.

In all cases, cells were distributed into 384-well capture plates (all coming from the same production batch) containing 2 µl of lysis solution using a FACSARIA III cell sorter. The lysis solution contained 0.2% Triton and RNAse inhibitors plus barcoded poly(T) reverse-transcription primers for single-cell RNA-seq. Non-cellular particles were discriminated by selecting only DRAQ5-positive cells (25 µM DRAQ5 staining; Thermo #62251), and cell doublet/multiplet exclusion was performed using forward scatter width (FSC-W) versus forward scatter height (FSC-H). Fresh cell dissociates were prepared every 2 h, and sorted plates were immediately spun down to ensure cell immersion into the lysis solution, then frozen at −80 °C until further processing.

MARS-seq. Single-cell libraries were prepared as previously described[22]. For each species, all single-cell libraries were prepared in parallel: 8,832 libraries for

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

A. queenslandica (13 plates for adult sponges and 10 for larvae), 6,144 for *M. leidyi* (16 plates) and 4,224 for *T. adhaerens* (12 plates). That is, we employed exactly the same conditions (incubation times, temperatures and so on) and reagents in order to minimize technical factors. First, using a Bravo automated liquid handling platform (Agilent), messenger RNA was converted into complimentary DNA with an oligo containing both the UMIs and cell barcodes. Unused oligonucleotides were removed by Exonuclease I treatment. Complimentary DNA was pooled (each pool representing half of the original 384-well MARS-seq plate) and linearly amplified using T7 in vitro transcription, and the resulting RNA was fragmented and ligated to an oligo containing the pool barcode and Illumina sequences, using T4 single stranded DNA:RNA ligase. Finally, RNA was reverse transcribed into DNA and amplified by polymerase chain reaction (PCR). Resulting libraries were tested for amplification using quantitative PCR, and the size distribution and concentration were calculated using TapeStation (Agilent) and Qubit (Invitrogen). For each species, all single-cell RNA-seq libraries were pooled at equimolar concentrations and sequenced to saturation (≥4 reads per UMI) using an Illumina NextSeq 500 sequencer and mid-output 75 cycles V2 kit (Illumina). For adult *A. queenslandica*, we obtained a total of 430 million reads, with an average depth of 53,000 reads per cell, and 6 reads per UMI on average (Supplementary Table 1). For *A. queenslandica* larvae, we obtained a total of 67 million reads, with an average depth of 11,000 reads per cell, and 5 reads per UMI on average. For *M. leidyi*, we obtained a total of 506 million reads, with an average depth of 36,000 reads per cell, and 5 reads per UMI on average. In the case of *T. adhaerens*, we obtained a total of 85 million reads, with an average depth of 14,000 reads per cell, and 7 reads per UMI on average.

Processing and filtering of MARS-seq reads. Reads were mapped into *A. queenslandica*, *T. adhaerens* and *M. leidyi* genomes using Bowtie2 (with the parameters -D 200 -R 3 -N 1 -L 20 -i S,1,0.50) and associated with gene intervals. For each species, we extended gene intervals up to 2 kb downstream or until the next gene in the same strand was found. This accounts for the poor 3′ untranslated region annotation of these species, which causes many of the MARS-seq (a 3′ biased RNA-seq method) reads to map outside genes. Additionally, to account for putative unannotated genes, we defined 500-bp bins (not covered by our gene intervals) genome-wide. We retained those with ≥10 uniquely mapping reads and used them in the cell clustering process (see below).

Mapped reads were further processed and filtered as previously described[22]. UMI filtering included two components—one eliminating spurious UMIs resulting from synthesis and sequencing errors, and the other eliminating artefacts involving unlikely in vitro transcription (IVT) product distributions that were probably a consequence of second-strand synthesis or IVT errors. The minim false discovery rate (FDR) Q value required for filtering was 0.2.

Metacell and clustering analysis. We used the MetaCell package (Supplementary Appendix 1) to select gene features, construct gene modules and create projected visualization of the data, using parameters as described below. We applied preliminary cell filtering based on total UMI counts using a permissive threshold of 100 UMIs (50 UMIs in the case of *A. queenslandica* larva, to account for the very different molecule count distributions in this sample). For gene selection, we used a normalized depth scaling correlation threshold of −0.1 (−0.05 in *A. queenslandica* larva and *T. adhaerens*) and a total UMI count of more than 100 molecules (the empirical median marker UMI count was 2,723 for the sponge, 1,013 for the ctenophore and 1,075 for the placozoan). For metacell construction, we used K = 150, a minimum module size of 30 and automatic filtering of background noise using an initial epsilon value of 0.03. Bootstrapping was performed using 1,000 iterations of resampling 75% of the cells, leading to an estimation of co-clustering between all pairs of single cells and the identification of robust clusters based on single or grouped metacells. For 2D projections, in the *A. queenslandica* adult dataset, we used a *k*-nearest neighbours constant of 50 and restricted the module graph degree by at most 10 (*A. queenslandica* larva, $k = 30$/max degree = 3; *M. leidyi*, $k = 30$/max degree = 7; *T. adhaerens*, $k = 30$/max degree = 8).

We performed manual validation and adjustment of the automatic module covers in Fig. 1 and Supplementary Fig. 4 as follows. We filtered metacells that were not enriched by at least three genes at over threefold over the median of the entire populations. Additionally, module-specific transcriptional enrichment was tested for each metacell by identifying a set of module-specific genes (top 50 genes with a fold change of ≥2) and computing the top 1% of their total expression across all non-module cells (also excluding cells in the two most similar modules). Given this top percentile as a threshold, the fraction of cells in the module that expressed the module's genes over the threshold was computed, and additional module filtering was applied if this value was lower than 30%. We also filtered out metacells with fewer than 10,000 total molecules. We note that cells that were filtered during this combined scheme may be part of additional undetected states, or may represent a weaker signal that is, in fact, part of other, more robust modules, but for our goals in the analysis here, robustness of the reported transcriptional states and the subsequent genomic analyses was key. Overall, this resulted in filtering 862 cells in the sponge, 785 cells in the ctenophore and 188 cells in the placozoan. Finally, we merged metacells with >20% of shared cells co-clustering in our 1,000 bootstrap replicates, resulting in the metacells presented in Figs. 2 and 3 and supported by bootstrap analysis in Supplementary Fig. 1.

iChIP. For the iChIP experiments, *M. leidyi* and *T. adhaerens* cells (dissociated as described above) were cross-linked in 1% formaldehyde for 10 min at room temperature. Cross-linking was quenched with 0.125 M glycine for 5 min at room temperature. Cross-linked cells were pelleted and stored at −80 °C. Chromatin was sonicated in a Bioruptor sonicator (Diagenode), distributing 1 M cells per 100 µl tube and using 45 sonication cycles (30” ON/30” OFF; High mode). Then, chromatin was immobilized onto anti-H3 antibody (#ab1791; Abcam)-coated Protein G Beads (Invitrogen). After 3 washes with 10 mM Tris pH8 plus protease inhibitors, immobilized chromatin was indexed with Illumina Y-shaped adaptors as described in ref. [55]. After barcoding, indexed chromatin was pooled and released from Ab-ProtG bead immunocomplexes by incubating for 30 min at 37 °C in a buffer containing 50 mM EDTA, 2% SDS, 2% deoxycholic acid and 1 M NaCl. After the incubation, the chromatin was separated from the magnetic beads using a magnet and the released indexed chromatin was transferred to another tube and diluted 1 to 20 in a buffer of 10 mM Tris-Cl, 10 mM NaCl and 1 mM EDTA. A small fraction of this dilution (60 µl) was separated to be sequenced as input. The remaining diluted indexed chromatin (approximately 10 ml) was concentrated to 200 µl using a 50 kDa centricon (Ambion), and the volume was brought to 400 µl with radioimmunoprecipitation assay (RIPA) buffer plus protease inhibitors. The 400 µl pool was divided into 2 to perform 2 ChIP assays—1 for H3K4me2 and another for H3K4me3. The specific ChIP reaction was carried out at this stage by incubating the 200 µl extract of the indexed chromatin pool with 4 µl of anti-H3K4me2 antibody (#ab3236; Abcam) or 2.5 µl of anti-H3K4me3 antibody (#07-473; Millipore) at 4 °C with rotation. After 10 h of incubation, 40 µl of pre-washed ProtG beads was added and incubated for 1 h to capture the antibody–chromatin complexes. Immunocomplexes were then washed 5 times with RIPA (150 mM NaCl, 0.1% SDS, 0.1% deoxycholate, 1% Tx-100 and 1 mM EDTA), twice with RIPA-500 (500 mM NaCl, 0.1% SDS, 0.1% deoxycholate, 1% Tx-100 and 1 mM EDTA), twice with LiCl buffer (250 mM LiCl, 0.5% NP-40, 0.5% deoxycholate and 1 mM EDTA) and twice with TE buffer, and resuspended in 50 µl of Chromatin Elution Buffer (0.4% SDS, 250 mM NaCl, 5 mM EDTA and 10 mM Tris-Cl pH 8) plus 2.5 µl of Proteinase K (NEB) and incubated for 2 h at 37 °C and 6 h at 65 °C. ChIPped DNA was purified with AMPure beads with a ratio of 2.5× and eluted in 23 µl of Elution Buffer (10 mM Tris pH8). To amplify the ChIPped barcoded DNA, 12 cycles of PCR were performed using 25 µl of 2× KAPA HiFi Master Mix and 2 µl of primer master mix.

iChIP libraries were sequenced using an Illumina NextSeq 500 sequencer. For *M. leidyi*, the total number of reads was 21 million for H3K4me2, 12 million for H3K4me3 and 10 million for input. For *T. adhaerens*, the read number totals were 24 million for H3K4me2, 14 million for H3K4me3 and 11 million for input.

iChIP analysis and enhancer definition. iChIP reads were trimmed to 37 nucleotides and then mapped into the corresponding reference genome using Bowtie version 1.1.1 (ref. [60]) with the parameters -v 3 -m 1. Duplicate reads were removed using SAMtools version 1.1 (ref. [61]). Mapped reads were extended to 200 bp (iChIP libraries fragment size), and 1-bp-resolution coverage statistics over each of the genomes were computed.

To control for ChIP-sequencing coverage and variable ChIP-sequencing specificity, we transformed raw coverage values to quantile values. H3K4me3 and H3K4me2 peaks were defined as regions with coverage quantiles over 0.97 (in *M. leidyi*) or 0.94 (in *T. adhaerens*), merging peaks located at <200 bp. To account for mappability and assembly problems (for example, repetitive regions), we defined ‘peaks’ using input data and excluded those regions from our H3K4me3/2 peaks. In downstream analysis, iChIP coverage is indicated as $-\log_2(1 - \text{coverage quantile})$, in a way that, for example, a normalized value of 9 indicates that coverage is in the top $1-2^{-9}$ quantile (that is, the top 1/512th of the distribution).

H3K4me3 is associated with promoter elements, while H3K4me2 is associated with both promoters and enhancers[55]. We used this property to search for distal enhancer elements in *M. leidyi* and *T. adhaerens*, by asking for H3K4me2 peaks that are ≥2 kb from any H3K4me3 or TSS (of an expressed gene; ≥5 total UMIs detected).

Sequence motif analysis. We extracted promoter sequences using −200 or + 50 bp from annotated TSSs and associated sequences with metacells whenever their gene was at least twofold overexpressed in the module compared with the background. We then performed de novo motif enrichment analysis for the regulatory sequences associated with each gene list, using the HOMER tool findMotifsGenome.pl (with default parameters, searching for 25 motifs and with a constant fragment size of 250 bp)[62]. For each species, we grouped all the resulting de novo motifs and used the HOMER tool compareMotif.pl to filter the motifs (minimum P value $< 1 \times 10^{-10}$; minimum number of hits in target sequences ≥10) and then merge redundant motifs (>0.8 similarity threshold). Additionally, in the case of *M. leidyi*, we searched for enriched motifs in all enhancers (1,157) versus the entire genome, using a HOMER fragment size of 600 bp.

For a comparison of de novo motifs with the database, we used data from Jolma et al.[63,64], the HOCOMOCO database[62,65], the JASPAR database, *Drosophila* DMMPMM database, the plant AthaMap database, and *Saccharomyces* motif collections from Harbinson et al. and MacIsaac et al. We computed similarities between motifs (Fig. 6 and Supplementary Fig. 7) using the motifSimilarity

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

function of the PWMEnrich R library, which computes the normalized sum of correlations between motif position frequency matrices.

As a result of the de novo motif finding, filtering and merging, we obtained a single set of motifs per species. We then analysed the over-representation of specific motifs in promoters associated with metacell-specific gene modules. For a short sequence element $s[1..k]=s_1,...,s_k$ and a position weight matrix (PWM) $w_i[c]$, the standard local probability model is defined by multiplication: $\log(P(s))=\sum_i \log(w_i[s_i])$ and the binding energy for a larger sequence element can be approximated[66] by $E(s[1...n])=\log(\sum_{j=1:(n-k)} P(s[j:(j+k)]))$. For each PWM, the 0.98 quantiles of genome-wide binding energies in windows of 250 bp (same size as promoters) were determined. These quantile values were then used as thresholds to determine the motif occurrence for each PWM at each element. The enrichment level of each PWM–metacell pair was computed as the fold change between the frequency of occurrence of a motif in the metacell promoters and the frequency in the background gene set (all other genes detected in this study). Enrichments were assessed statistically using a hypergeometric test. We account for multiple testing by performing 100 random permutations of the promoter motif energy matrix, computing *P* values for each permutation and using the resulting distribution to derive FDR values on the empirical enrichments. An FDR threshold of 0.02 was used for the motif enrichment visualization. Additionally, only motifs with a fold-change enrichment over 1.5 in at least one metacell, and a minimum foreground count of 5 (that is, at least 5 genes in the metacell gene set with the motif in their promoters) and background count of 100 were considered.

Finally, we performed a cross-validation analysis by dividing expressed genes into 5 blocks and, for each of them, running the whole de novo motif discovery pipeline with the other 80% of the genes (training set). Using the glmnet R package, we built a LASSO (least absolute shrinkage and selection operator) regularized linear model based on the promoter motif energies and gene expression values of the training set (80%). We then employed this model to predict the expression values of the gene test set (20%) based on the motif energies in their promoters. We did this for each of the five blocks, resulting in predicted expression values for all expressed genes in our dataset. Receiver operating characteristic curves and AUC values were computed using the pROC R package.

Gene functional annotation. We used BLASTp (with the parameters -evalue 1×10^{-5} and -max_target_seqs 1) to find the most similar, if any, human, fruit fly and yeast homologues (retrieved from UniProt) for each protein of the predicted *A. queenslandica, M. leidyi* and *T. adhaerens* predicted proteomes. Additionally, we predicted for each protein the Pfam domain composition using PfamScan[67] with the default curated gathering threshold. Transcription factors were identified using univocal Pfam domains for each structural transcription factor family[68]. In the case of multiple transcription factor families (Homeobox, Fox, bHLH, bZIP, DM, Smad, Myb, NR, RFX, RHD, SRF, Ets, T-box and Sox), we used phylogenetic analyses for each family to classify them into specific subfamilies (together with the complete transcription factor sets of an additional ten animal species, including *Homo sapiens* and *D. melanogaster* for reference annotation). Briefly, sequences were aligned using MAFFT[69], the resulting analysis were manually edited, ProtTest[70] was used to define the best-fit aminoacidic substitution model in each case, and then phylogenies were computed using RAxML[71] and PhyloBayes[72], for maximum likelihood and Bayesian inference, respectively. We used a similar strategy to build a phylogeny of *A. queenslandica* aspzincins (Supplementary Fig. 2e), extending our search for aspzincins to other eukaryotic and bacterial species. To this end, we used the presence of the Aspzincin_M35 domain (PF14521; Pfam) to identify aspzincins in different species.

Phylogenetic distribution and gene-age estimation. We used the complete predicted proteomes of 39 species (Supplementary Table 6) at key phylogenetic positions to compute orthogroups, including an extensive set of 11 ctenophore species (*Beroe abyssicola, Bolynopsis infundibulum, Coeloplana astericola, Coeloplana meteoris, Dryodora glandiformis, Euplokamis dunlapae, Mertensiidae* species, *Vallicula multiformis, Lampea pancerina, Pleurobrachia bachei* and *Mnemiopsis leyidi*)[4,12], 10 poriferan species (*Clathrina coriacea, Grantia compressa, Leuconia nivea, Sycon ciliatum, Plakina jani, Oscarella carmela, Pleraplysilla spinifera, A. queenslandica, Eunapius carteri* and *Ephydatia muelleri*)[4,47], the placozoan *T. adhaerens*, 8 cnidarian and bilaterian species (*H. sapiens, Branchiostoma floridae, D. melanogaster, Tribolium castaneum, Capitella teleta, Lottia gigantea, Acropora digitifera* and *Nematostella vectensis*) and 8 non-metazoan eukaryotes (*Salpingoca rosetta, Capsaspora owczarzaki, Creolimax fragrantissima, Saccharomyces cerevisiae, Spizellomyces punctatus, Dictyostelium discoideum, Arabidopsis thaliana* and *Naegleria gruberi*). We computed reciprocal BLAST results between all complete proteomes, with a fixed database size and an e-value threshold of 1×10^{-4}. Based on these reciprocal BLAST results, orthogroups were computed using the orthoMCL algorithm[73] with an inflation value (I parameter) of 1.3. We parsed these orthogroups using a parsimony criterion to generate an age estimation for each *A. queenslandica*, *M. leidyi* and *T. adhaerens* gene.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The analysis code is available on our group website at http://compgenomics.weizmann.ac.il/tanay/?page_id=99.

Data availability. All data was deposited in Gene Expression Omnibus with the accession number GSE111068. The MetaCell package, UMI tables and annotation files are available on our group website at http://compgenomics.weizmann.ac.il/tanay/?page_id=99.

Received: 8 March 2018; Accepted: 9 May 2018;
Published online: 25 June 2018

References

1. Arendt, D. et al. The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).
2. Sebé-Pedrós, A., Degnan, B. M. & Ruiz-Trillo, I. The origin of Metazoa: a unicellular perspective. *Nat. Rev. Genet.* **18**, 498–512 (2017).
3. Whelan, N. V. et al. Ctenophore relationships and their placement as the sister group to all other animals. *Nat. Ecol. Evol.* **1**, 1737–1746 (2017).
4. Simion, P. et al. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr. Biol.* **27**, 958–967 (2017).
5. Hejnol, A. et al. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc. R. Soc. B* **276**, 4261–4270 (2009).
6. Dunn, C. W. et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745–749 (2008).
7. Valentine, J. W. in *Keywords and Concepts in Evolutionary Developmental Biology* (eds. Hall, B. & Olson, W.) 35–53 (Harvard Univ. Press, Cambridge, MA, 2003).
8. Cunningham, J. A., Liu, A. G., Bengtson, S. & Donoghue, P. C. J. The origin of animals: can molecular clocks and the fossil record be reconciled? *BioEssays* **39**, e201600120 (2017).
9. Dunn, C. W., Leys, S. P. & Haddock, S. H. D. The hidden biology of sponges and ctenophores. *Trends Ecol. Evol.* **30**, 282–291 (2015).
10. Jager, M. & Manuel, M. Ctenophores: an evolutionary-developmental perspective. *Curr. Opin. Genet. Dev.* **39**, 85–92 (2016).
11. Jékely, G., Paps, J. & Nielsen, C. The phylogenetic position of ctenophores and the origin(s) of nervous systems. *EvoDevo* **6**, 1 (2015).
12. Moroz, L. L. et al. The ctenophore genome and the evolutionary origins of neural systems. *Nature* **510**, 109–114 (2014).
13. Simpson, T. L. *The Cell Biology of Sponges* (Springer, New York, NY, 1984).
14. Schierwater, B. & DeSalle, R. Placozoa. *Curr. Biol.* **28**, R97–R98 (2018).
15. Srivastava, M. et al. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* **466**, 720–726 (2010).
16. 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).
17. Ryan, J. F. et al. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science* **342**, 1242592 (2013).
18. Srivastava, M. et al. The *Trichoplax* genome and the nature of placozoans. *Nature* **454**, 955–960 (2008).
19. Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317**, 86–94 (2007).
20. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
21. Sebé-Pedrós, A. et al. Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-Seq. *Cell* **173**, 1520–1534 (2018).
22. Jaitin, D. A. et al. Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
23. Gonobobleva, E. & Maldonado, M. Choanocyte ultrastructure in *Halisarca dujardini* (Demospongiae, Halisarcida). *J. Morphol.* **270**, 615–627 (2009).
24. Funayama, N., Nakatsukasa, M., Hayashi, T. & Agata, K. Isolation of the choanocyte in the fresh water sponge, *Ephydatia fluviatilis* and its lineage marker, Ef annexin. *Dev. Growth Differ.* **47**, 243–253 (2005).
25. Nickel, M., Scheer, C., Hammel, J. U., Herzen, J. & Beckmann, F. The contractile sponge epithelium sensu lato—body contraction of the demosponge *Tethya wilhelma* is mediated by the pinacoderm. *J. Exp. Biol.* **214**, 1692–1698 (2011).
26. Nichols, S. A., Roberts, B. W., Richter, D. J., Fairclough, S. R. & King, N. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex. *Proc. Natl Acad. Sci. USA* **109**, 13046–13051 (2012).
27. Adamska, M. et al. The evolutionary origin of hedgehog proteins. *Curr. Biol.* **17**, 836–837 (2007).
28. Nakanishi, N., Sogabe, S. & Degnan, B. M. Evolutionary origin of gastrulation: insights from sponge development. *BMC Biol.* **12**, 26 (2014).
29. Müller, W. E. G. The stem cell concept in sponges (Porifera): metazoan traits. *Semin. Cell Dev. Biol.* **17**, 481–491 (2006).
30. Alié, A. et al. The ancestral gene repertoire of animal stem cells. *Proc. Natl Acad. Sci. USA* **112**, E7093–E7100 (2015).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

31. Rieger, R. M. The biphasic life cycle—a central theme of metazoan evolution. *Am. Zool.* **34**, 484–491 (1994).
32. Degnan, S. M. & Degnan, B. M. The origin of the pelagobenthic metazoan life cycle: what's sex got to do with it? *Integr. Comp. Biol.* **46**, 683–690 (2006).
33. Nakanishi, N., Stoupin, D., Degnan, S. M. & Degnan, B. M. Sensory flask cells in sponge larvae regulate metamorphosis via calcium signaling. *Integr. Comp. Biol.* **55**, 1018–1027 (2015).
34. Adamska, M. et al. Wnt and TGF-β expression in the sponge *Amphimedon queenslandica* and the origin of metazoan embryonic patterning. *PLoS ONE* **2**, e1031 (2007).
35. Philippe, H. et al. Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **19**, 706–712 (2009).
36. Liebeskind, B. J., Hofmann, H. A., Hillis, D. M. & Zakon, H. H. Evolution of animal neural systems. *Annu. Rev. Ecol. Evol. Syst.* **48**, 377–398 (2017).
37. Schnitzler, C. E. et al. Genomic organization, evolution, and expression of photoprotein and opsin genes in *Mnemiopsis leidyi*: a new view of ctenophore photocytes. *BMC Biol.* **10**, 107 (2012).
38. Satterlie, R. & Case, J. Gap junctions suggest epithelial conduction within the comb plates of the ctenophore *Pleurobrachia bachei*. *Cell Tissue Res.* **193**, 87–91 (1978).
39. Steinmetz, P. R. H. et al. Independent evolution of striated muscles in cnidarians and bilaterians. *Nature* **487**, 231–234 (2012).
40. Sebé-Pedrós, A. et al. Insights into the origin of metazoan filopodia and microvilli. *Mol. Biol. Evol.* **30**, 2013–2023 (2013).
41. Tudor, J. E., Pallaghy, P. K., Pennington, M. W. & Norton, R. S. Solution structure of ShK toxin, a novel potassium channel inhibitor from a sea anemone. *Nat. Struct. Biol.* **3**, 317–320 (1996).
42. Marlow, H. & Arendt, D. Evolution: ctenophore genomes and the origin of neurons. *Curr. Biol.* **24**, R757–R761 (2014).
43. Smith, C. L. et al. Novel cell types, neurosecretory cells, and body plan of the early-diverging metazoan *Trichoplax adhaerens*. *Curr. Biol.* **24**, 1565–1572 (2014).
44. Ganz, T. Defensins: antimicrobial peptides of innate immunity. *Nat. Rev. Immunol.* **3**, 710–720 (2003).
45. Senatore, A., Reese, T. S. & Smith, C. L. Neuropeptidergic integration of behavior in *Trichoplax adhaerens*, an animal without synapses. *J. Exp. Biol.* **220**, 3381–3390 (2017).
46. Nikitin, M. Bioinformatic prediction of *Trichoplax adhaerens* regulatory peptides. *Gen. Comp. Endocrinol.* **212**, 145–155 (2015).
47. Riesgo, A., Farrar, N., Windsor, P. J., Giribet, G. & Leys, S. P. The analysis of eight transcriptomes from all poriferan classes reveals surprising genetic complexity in sponges. *Mol. Biol. Evol.* **31**, 1102–1120 (2014).
48. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
49. Sebé-Pedrós, A. et al. High-throughput proteomics reveals the unicellular roots of animal phosphosignaling and cell differentiation. *Dev. Cell* **39**, 186–197 (2016).
50. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
51. Piasecki, B. P., Burghoorn, J. & Swoboda, P. Regulatory Factor X (RFX)-mediated transcriptional rewiring of ciliary genes in animals. *Proc. Natl Acad. Sci. USA* **107**, 12969–12974 (2010).
52. Wang, S. & Samakovlis, C. Grainy head and its target genes in epithelial morphogenesis and wound healing. *Curr. Top. Dev. Biol.* **98**, 35–63 (2012).
53. Peter, I. S. & Davidson, E. H. Evolution of gene regulatory networks controlling body plan development. *Cell* **144**, 970–985 (2011).
54. Sebé-Pedrós, A. et al. The dynamic regulatory genome of Capsaspora and the origin of animal multicellularity. *Cell* **165**, 1224–1237 (2016).
55. Lara-Astiaso, D. et al. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
56. Nora, E. P. et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944 (2017).
57. Wang, Q., Sun, Q., Czajkowsky, D. M. & Shao, Z. Sub-kb Hi-C in *D. melanogaster* reveals conserved characteristics of TADs between insect and mammalian cells. *Nat. Commun.* **9**, 188 (2018).
58. Ramírez, F. et al. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).
59. Grell, K. G. & Benwitz, G. Ultrastruktur von *Trichoplax adhaerens* F.E. Schulze. *Cytobiologie* **4**, 216–240 (1971).
60. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
61. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
62. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
63. Jolma, A. et al. DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
64. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**, 384–388 (2015).
65. Kulakovskiy, I. V. et al. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* **41**, D195–D202 (2013).
66. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972 (2006).
67. Punta, M. et al. The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
68. De Mendoza, A. et al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc. Natl Acad. Sci. USA* **110**, E4858–E4866 (2013).
69. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).
70. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
71. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
72. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).
73. Li, L., Stoeckert, C. J. Jr & Roos, D. S. D. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
74. Gaiti, F. et al. Landscape of histone modifications in a sponge reveals the origin of animal *cis*-regulatory complexity. *eLife* **6**, e22194 (2017).
75. Booth, D. S. & King, N. Evolution: gene regulation in transition. *Nature* **534**, 482–483 (2016).
76. Vij, S. et al. Evolutionarily ancient association of the FoxJ1 transcription factor with the motile ciliogenic program. *PLoS Genet.* **8**, e1003019 (2012).

Acknowledgements

We thank all members of the Tanay laboratory for comments and discussion, X. Grau-Bové for genome statistics in different species, A. Furu for help with *M. leidyi*, and H.-J. Osigus and B. Schierwater for providing *T. adharens* starting culture. Research in A.H.'s group was supported by the European Research Council Community's Framework Program Horizon 2020 (2014–2020) ERC grant agreement 648861 and an NSF IRFP Postdoctoral Fellowship (1158629) to K.P. Research by B.M.D. is supported by the Australian Research Council. A.S.-P. was supported by an EMBO Long-Term Fellowship (ALTF 841-2014). Research in A.T.'s group was supported by the European Research Council Community's Framework Program Horizon 2020 (2014–2020) ERC grant agreement 724824. A.T. is a Kimmel investigator.

Author contributions

A.S.-P. and A.T. conceived the project. K.P., A.H., B.M.D. and F.G. provided animal specimens and chromatin material. Z.M. and E.C. assisted with experimental setup and analysis tools. I.A. assisted with iChIP and MARS-seq setup and reagents. A.S.-P. performed the MARS-seq experiments. A.S.-P. and D.L.-A. performed the iChIP experiments. A.S.-P. and A.T. analysed the data and wrote the manuscript. All authors discussed and commented on the data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0575-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to A.S. or A.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

natureresearch

Corresponding author(s): Amos Tanay
Arnau Sebe-Pedros

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.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
□	☒	An indication of 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 statistics 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
☒	□	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	NA
Data analysis	Link to the code and usage instructions are provided, as well as detailed description of the algorithm (Appendix S1).

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 upon request. 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

All data was deposited in GEO with the accession number GSE111068. The MetaCell package, UMI tables and annotation files are available on our group website: http://compgenomics.weizmann.ac.il/tanay/?page_id=99

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	Sequencing depth and number of libraries were defined to allow support for the paper's main conclusions (there is no "sample size" in this paper).
Data exclusions	No data were excluded from the analysis.
Replication	NA
Randomization	NA
Blinding	NA

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used	We employed antibodies against histone H3 and specific modifications of histone H3: anti-H3 antibody (Abcam, #ab1791) anti-H3K4me2 antibody (Abcam, #ab3236) anti-H3K4me3 antibody (Millipore, #07-473)
Validation	These antibodies have a wide species spectrum (paneukaryotic) and have been extensively used and validated in iChIP studies.

Animals and other organisms

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

Laboratory animals	Placozoan (Trichoplax adhaerens) specimens were cultured in the lab.
Wild animals	Sponge (Amphimedon queenslandica) specimens were collected from Heron Island Reef, Great Barrier Reef, Queensland, Australia. Ctenophore (Mnemiopsis leidyi) specimens originated from L. Friis-Møller, Kristineberg, Sweden.
Field-collected samples	Sponge and ctenophore specimens were mantained in filtered artificial sea water until further processing (dissociation for single-cell analysis or for chromatin extraction).

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	GSE111068
Files in database submission	Mnemiopsis_genes.bed Mnemiopsis_genome_sequence.fasta Mnemiopsis_original_scaffolds_edges.bed Trichoplax_genes.bed Trichoplax_genome_sequence.fasta Trichoplax_original_scaffolds_edges.bed Mnemiopsis_H3K4me2_peaks.bed Mnemiopsis_H3K4me3_peaks.bed Trichoplax_H3K4me2_peaks.bed Trichoplax_H3K4me3_peaks.bed Mnemiopsis_input_all_RPM.bw Mnemiopsis_me2_all_RPM.bw Mnemiopsis_me3_all_RPM.bw Trichoplax_input_all_RPM.bw Trichoplax_me2_all_RPM.bw Trichoplax_me3_all_RPM.bw Mnemiopsis_iChIP1_input_R1.fastq.gz Mnemiopsis_iChIP1_input_R2.fastq.gz Mnemiopsis_iChIP1_K4me2_R1.fastq.gz Mnemiopsis_iChIP1_K4me2_R2.fastq.gz Mnemiopsis_iChIP1_K4me3_R1.fastq.gz Mnemiopsis_iChIP1_K4me3_R2.fastq.gz Mnemiopsis_iChIP2_input_R1.fastq.gz Mnemiopsis_iChIP2_input_R2.fastq.gz Mnemiopsis_iChIP2_K4me2_R1.fastq.gz Mnemiopsis_iChIP2_K4me2_R2.fastq.gz Mnemiopsis_iChIP2_K4me3_R1.fastq.gz Mnemiopsis_iChIP2_K4me3_R2.fastq.gz Trichoplax_iChIP1_input_R1.fastq.gz Trichoplax_iChIP1_input_R2.fastq.gz Trichoplax_iChIP1_K4me2_R1.fastq.gz Trichoplax_iChIP1_K4me2_R2.fastq.gz Trichoplax_iChIP1_K4me3_R1.fastq.gz Trichoplax_iChIP1_K4me3_R2.fastq.gz Trichoplax_iChIP2_input_R1.fastq.gz Trichoplax_iChIP2_input_R2.fastq.gz Trichoplax_iChIP2_K4me2_R1.fastq.gz Trichoplax_iChIP2_K4me2_R2.fastq.gz Trichoplax_iChIP2_K4me3_R1.fastq.gz Trichoplax_iChIP2_K4me3_R2.fastq.gz
Genome browser session (e.g. UCSC)	NA

Methodology

Replicates	Two replicates of iChIP experiments. Reads were pooled for downstream analysis.
Sequencing depth	12 cycles of library PCR. 37nt Paired-End Reads. For M.leidyi, the total number of reads was: 21M (H3K4me2), 12M (H3K4me3) and 10M (input). For T.adhaerens, the total number of reads was: 24M (H3K4me2), 14M (H3K4me3), and 11M (input). For T.adhaerens, % of single-mapping reads ranged 65-68%. For M.leidyi, % of single-mapping reads ranged 40-42%.
Antibodies	anti-H3 antibody (Abcam, #ab1791)

Antibodies

anti-H3K4me2 antibody (Abcam, #ab3236)
anti-H3K4me3 antibody (Millipore, #07-473)

Peak calling parameters

We transformed raw coverage values to quantile values. H3K4me3 and H3K4me2 peaks were defined as regions with coverage quantiles over 0.97 (in M.leidyi) or 0.94 (in T.adhaerens), merging peaks located at <200bp. To account for mappability/assembly problems, we defined “peaks” using input data and excluded those regions from our H3K4me3/me2 peaks.

Data quality

NA

Software

Reads mapped using bowtie v1.1.1 with parameters -m1 -v3. Duplicated reads removed using SAMtools v1.1.