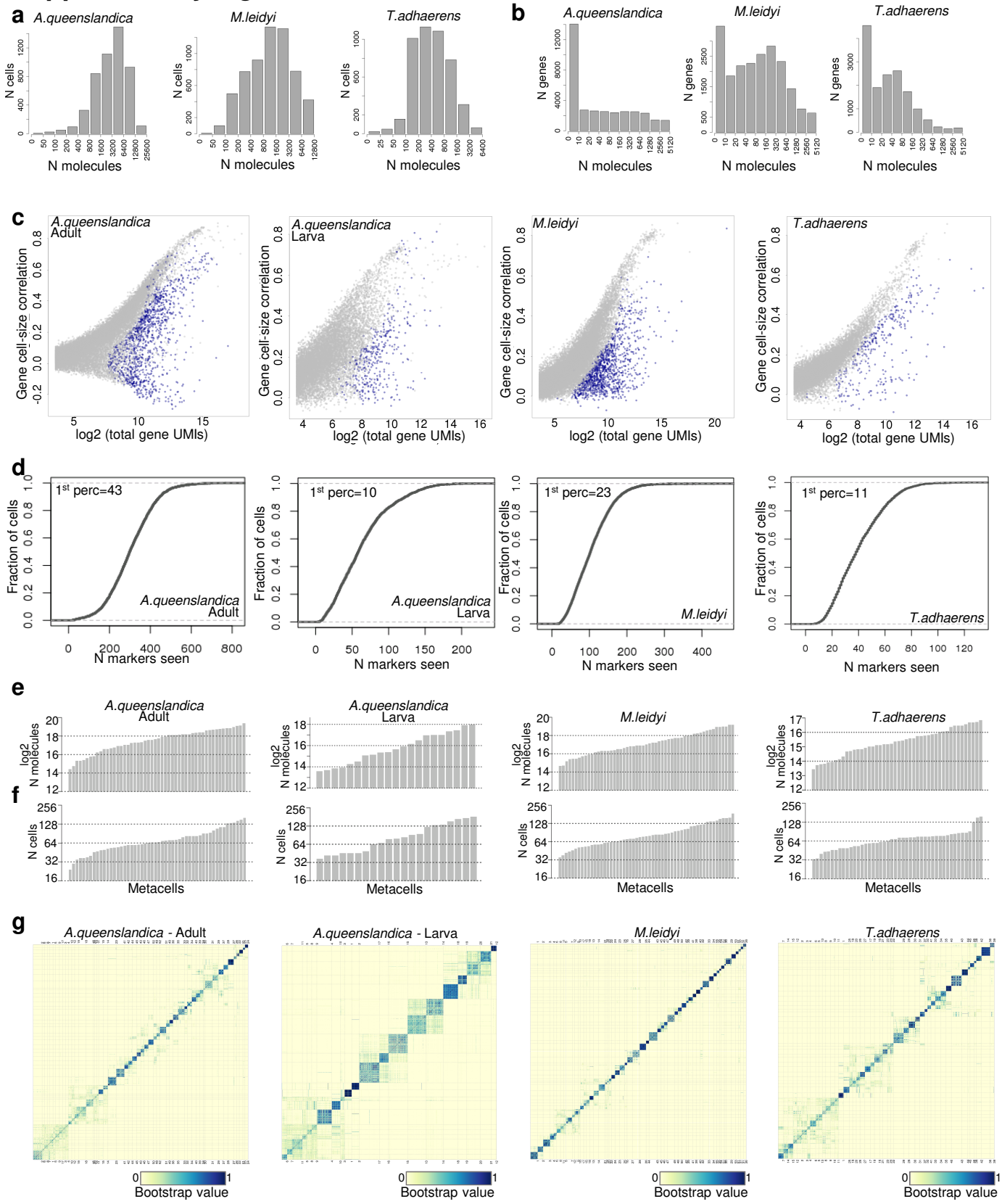
Early metazoan cell type diversity and the evolution of multicellular gene regulation

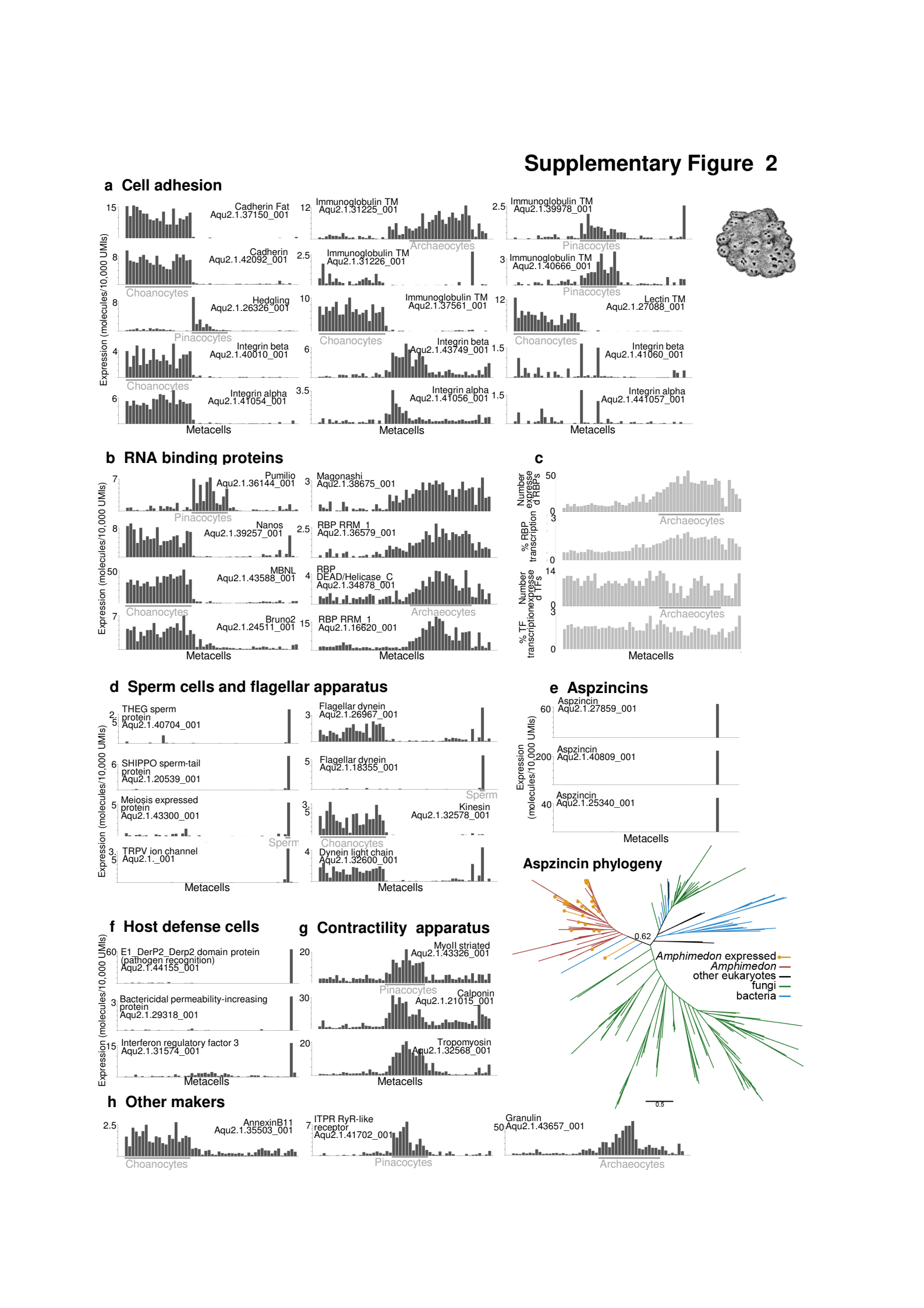
Sebé-Pedrós, Arnau; Chomsky, Elad; Pang, Kevin; Lara-Astiaso, David; Gaiti, Federico; Mukamel, Zohar; Amit, Ido; Hejnol, Andreas; Degnan, Bernard M.; Tanay, Amos

Supplementary Figure 1



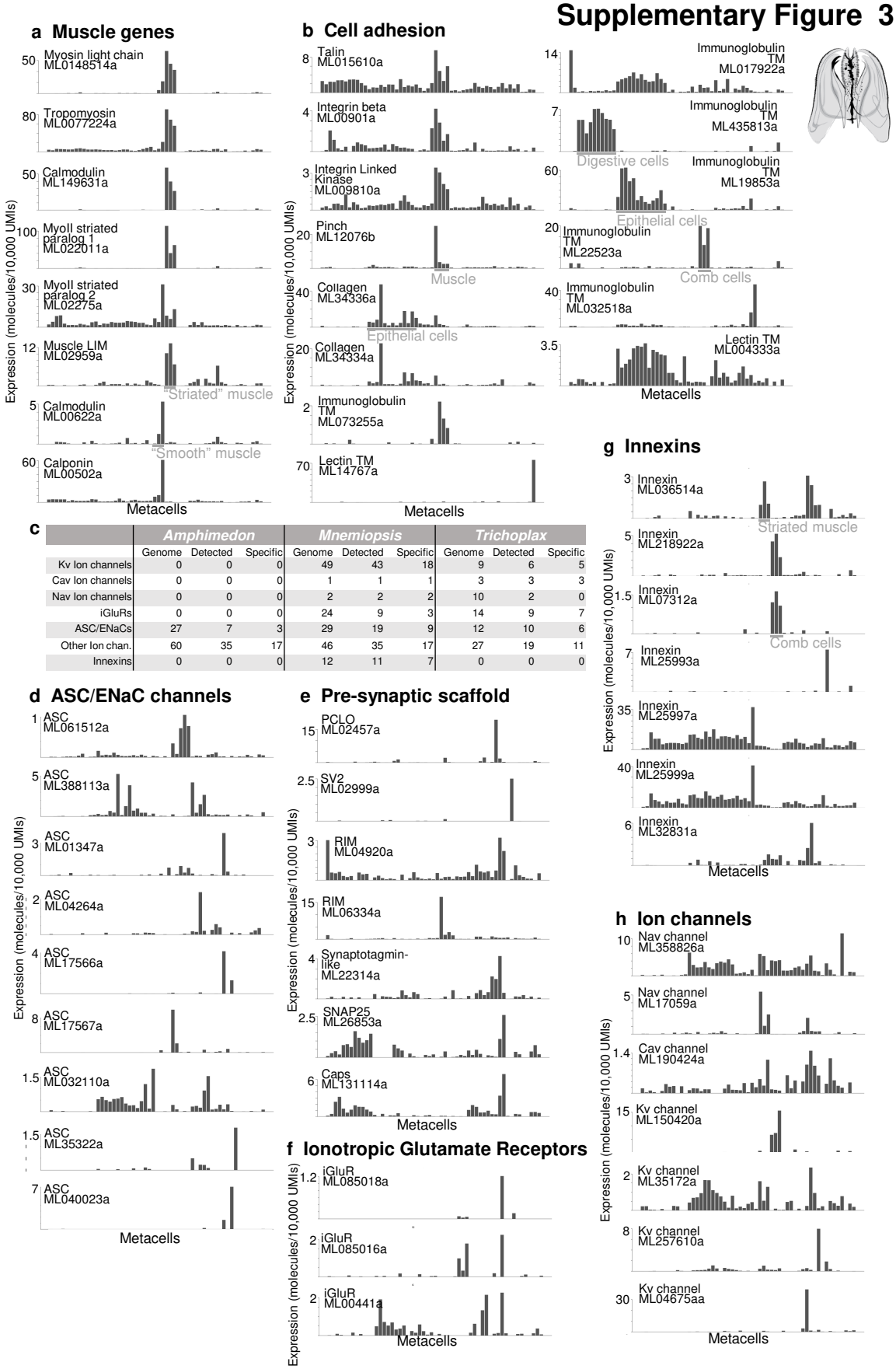
Supplementary Figure S1. UMI statistics and cell clustering analysis. a, Distribution of total RNA molecules per cell. **b,** Distribution of total RNA molecules per gene. **c,** Relationship between gene total expression (x-axis) and the correlation between gene expression and total RNA molecules per cell (y-axis). Marker genes selected for cell clustering are shown in blue. **d,** Cumulative distribution of number of marker genes detected per single cell. **e,** Total number of molecules per metacell. **f,** Total number of cells per metacell. **g,** Bootstrap analysis. Heatmap representing the frequency of cell-to-cell association in 1,000 bootstrap subsamplings.

Supplementary Figure 2



Supplementary Figure 2. ***Amphimedon queenslandica*** **gene expression profiles. a,** Expression profiles of selected cell adhesion genes. Multiple transmembrane (TM) adhesion proteins mediate cell type-specific adhesion. Notice the existence of three pairs of co-expressed integrin α/β paralogs (rows 4 and 5). **b,** Expression profiles of selected RNA binding proteins (RBPs). **c,** Number of RBPs and percentage of transcription (N molecules RBPs/total molecules in metacell) dedicated to RBPs in each metacell (top). For a reference, the same values are shown for TFs. Notice that RBP expression is prominent in archaeocytes ($p<<0.001$, chi-square test), in line with previous reports[30]. **d,** Expression profiles of genes related to sperm and flagellar functions. Flagellar genes (right column) are expressed both in choanocytes and in sperm cells. **e,** Expression profiles of 3 aspzincin metallo-endopeptidase paralogs (top). Bayesian phylogenetic analysis of aspzincin proteins, color-coded taxonomically as indicated in the legend (bottom). *A.queenslandica* aspzincins likely emerged from an horizontal gene transfer event from bacteria, followed by a massive expansion that generated *A.queenslandica* aspzincin paralogs. Many of these paralogs are expressed in a specific cell type, providing a striking example of HGT-based gene innovation linked to the emergence of a new cell type. **f,** Expression of selected markers genes associated to host defense cells. **g,** Expression of genes associated to actin cytoskeleton and cell contractility. These are co-expressed in pinacocyte cells. **h,** Expression of selected genes associated to broad cell clusters. AnnexinB11 has been shown to be associated to choanocytes in other sponge species[24]. All expression values are shown as molecules per 10,000 UMIs.

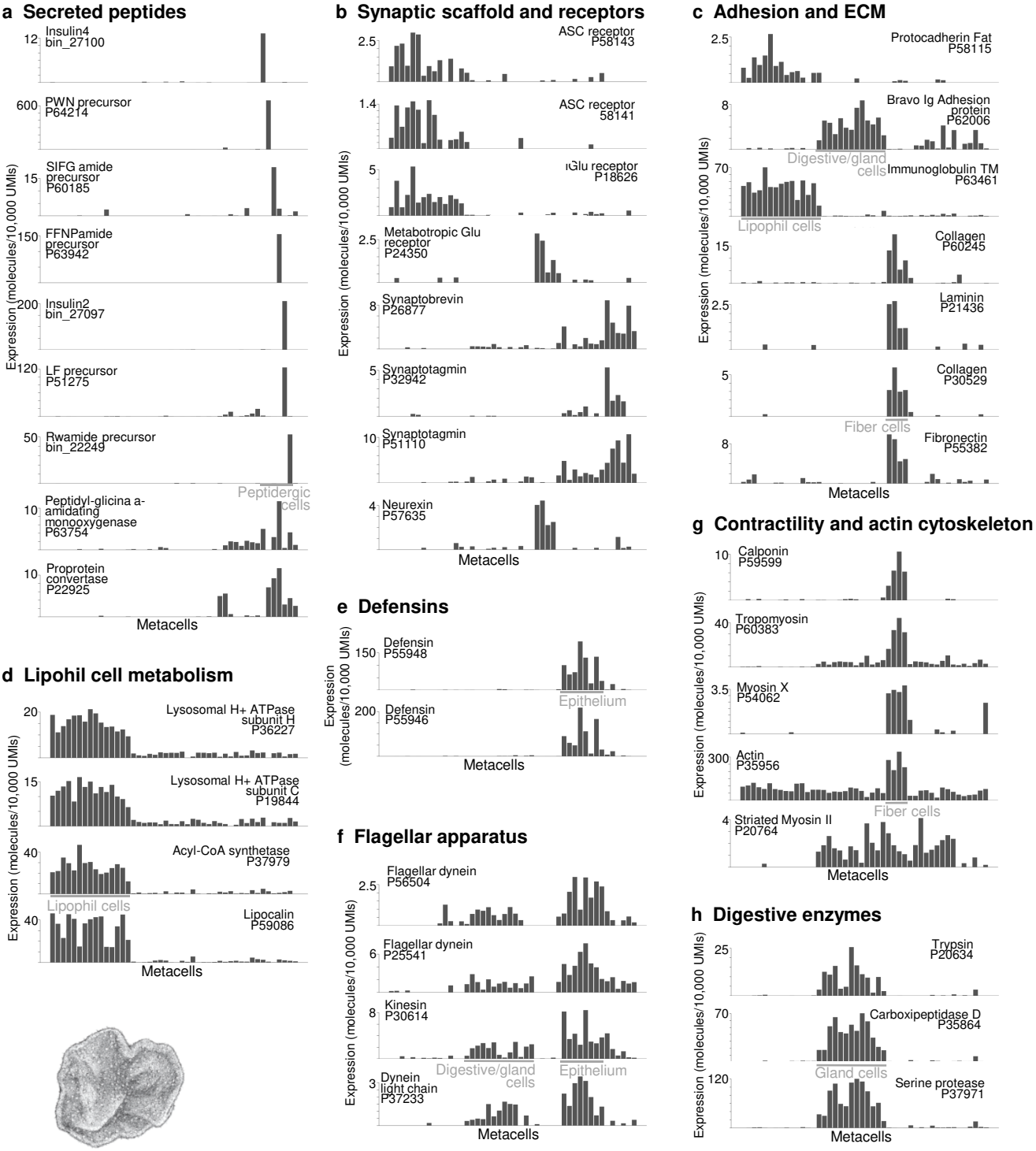
Supplementary Figure 3



	Amphimedon			*Mnemiopsis*			*Trichoplax*		
	Genome	Detected	Specific	Genome	Detected	Specific	Genome	Detected	Specific
Kv Ion channels	0	0	0	49	43	18	9	6	5
Cav Ion channels	0	0	0	1	1	1	3	3	3
Nav Ion channels	0	0	0	2	2	2	10	2	0
iGluRs	0	0	0	24	9	3	14	9	7
ASC/ENaCs	27	7	3	29	19	9	12	10	6
Other Ion chan.	60	35	17	46	35	17	27	19	11
Innexins	0	0	0	12	11	7	0	0	0

Supplementary Figure 3. ***Mnemiopsis leidyi*** **gene expression profiles. a,** Expression profiles of genes associated to muscle functions[39]. **b,** Expression profiles of selected cell adhesion genes. Multiple transmembrane (TM) adhesion proteins mediate cell type-specific adhesion, including numerous transmembrane lectins and immunoglobulins. **c,** Comparison of number of ion channels and innexins encoded in the genome, detected in our scRNAseq dataset and expressed in a cell type-specific manner (FC>2 in at least one metacell) in each species. Notice the large number of Kv ion channels and innexins expressed in the ctenophore *M.leidyi*. (Kv, potassium voltage-gated; Cav, calcium voltage-gated; Nav, sodium voltage-gated; iGluR, ionotropic glutamate receptors; ASC, amiloride-sensitive sodium channel). **d,** Expression profiles of amiloride-sensitive sodium channels (ASC). **e,** Expression profiles of genes associated to pre-synaptic scaffold in bilaterian animals. Notice the lack of co-expression in any particular metacell. **f,** Expression of ionotropic glutamate receptors. **g,** Expression of innexins. Notice that diverse cell clusters express different innexins, suggesting the existence of specialized electrical synapses in multiple cell types. **h,** Expression of selected voltage-gated ion channels. As in the case of innexins, ASC and iGluRs, multiple specific cell types express different ion channels. All expression values are shown as molecules per 10,000 UMIs.

Supplementary Figure 4



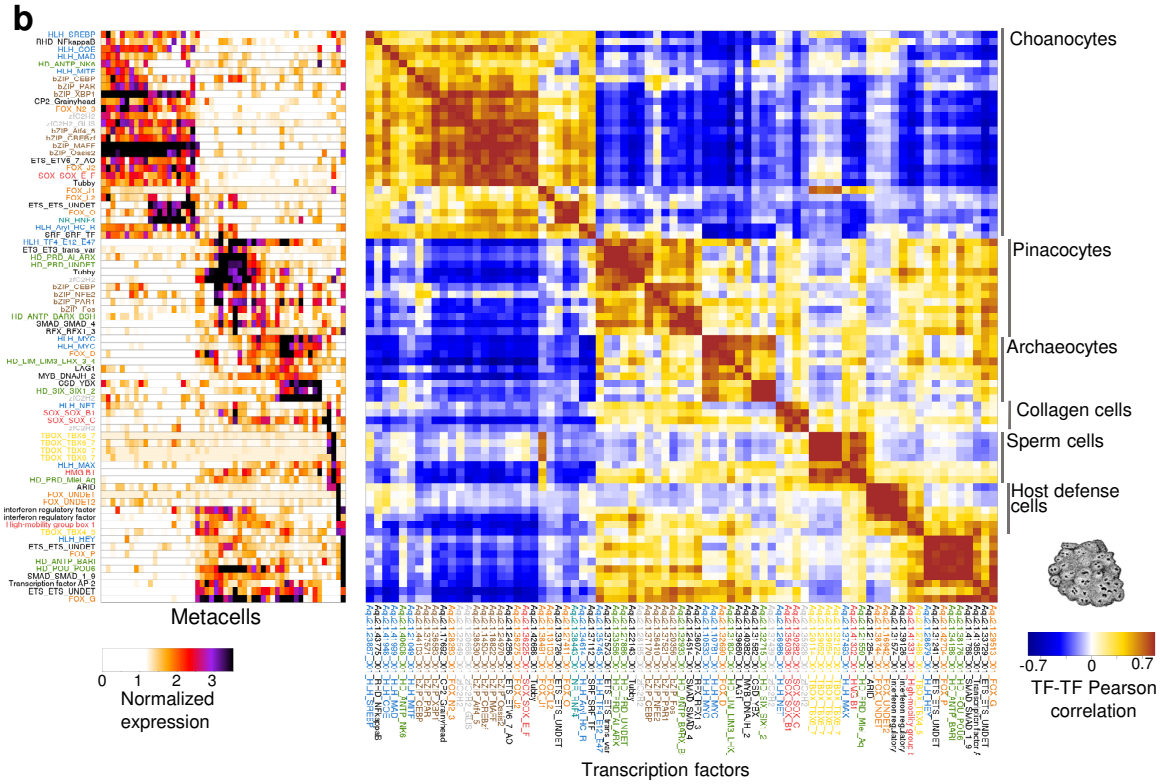
Supplementary Figure 4. ***Trichoplax adhaerens*** **gene expression profiles. a,** Expression profiles of regulatory peptides (as predicted by[45]) and peptide-processing enzymes. **b,** Expression profiles of pre-synaptic scaffold components and selected ion channels. **c,** Expression profiles of cell adhesion genes and extracellular matrix components. Notice that fiber cells express a large variety of ECM components, in accordance to its known biology[43]. **d,** Expression profiles of genes related to lipid metabolism and lysosomes. **e,** Expression profiles of defensin genes. These are host defense peptides expressed often in epithelial cells[59]. **f,** Expression profiles of ciliary components. Both epithelial cells and digestive/gland cells express these markers. In accordance with its known ultrastructural features[43], this suggests that these two cell types are ciliated. **g,** Expression profiles of actin cytoskeleton components, mostly expressed in fiber cells. **h,** Expression profiles of selected digestive enzymes. All expression values are shown as molecules per 10,000 UMIs.

Supplementary Figure 5

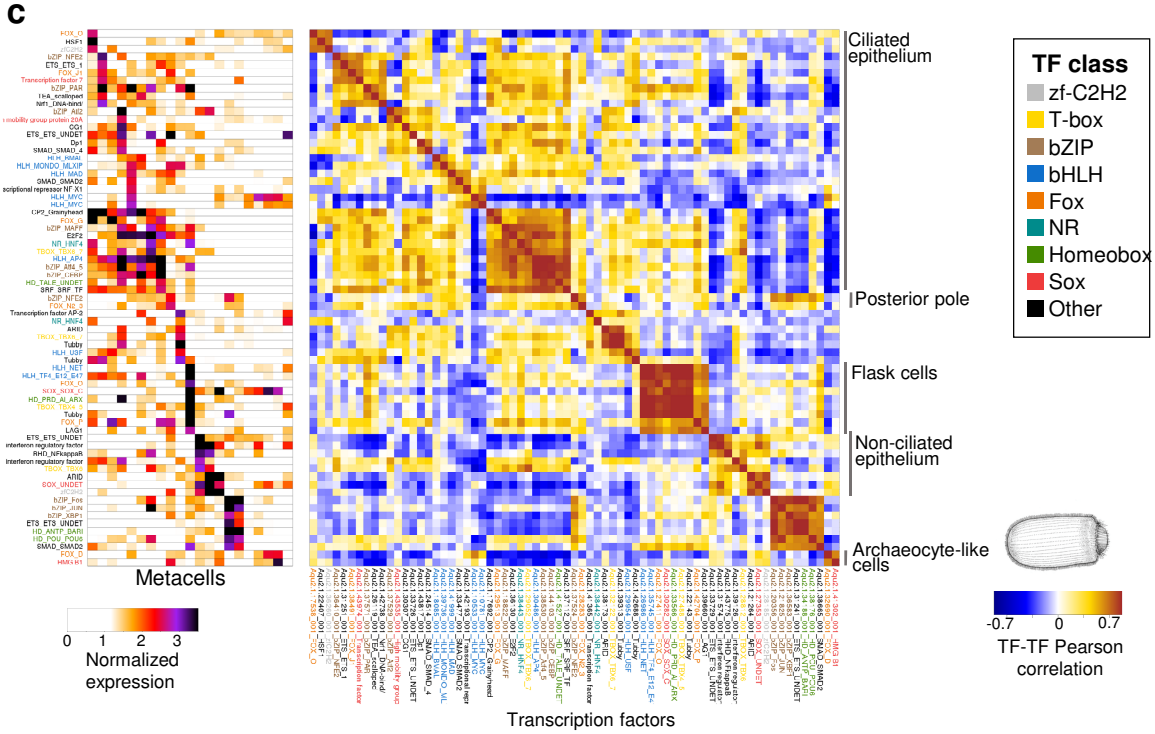
a

	Amphimedon	*Mnemiopsis*	*Trichoplax*
N TFs genome	231	281	209
N detected TFs	168	231	129
N specific TFs	87	82	86
% specific TFs	**51.7**	**35.5**	**66.7**

b

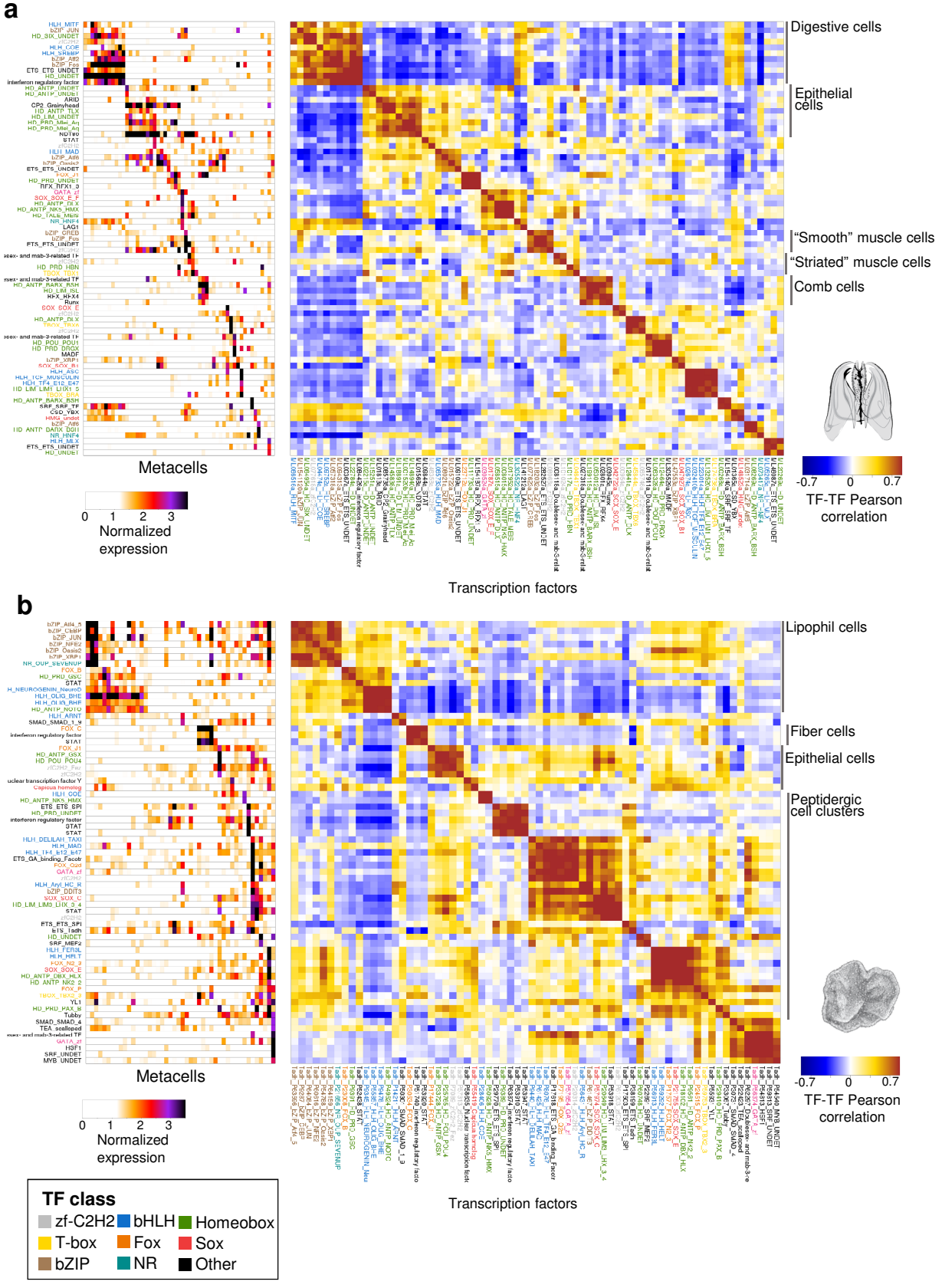


c



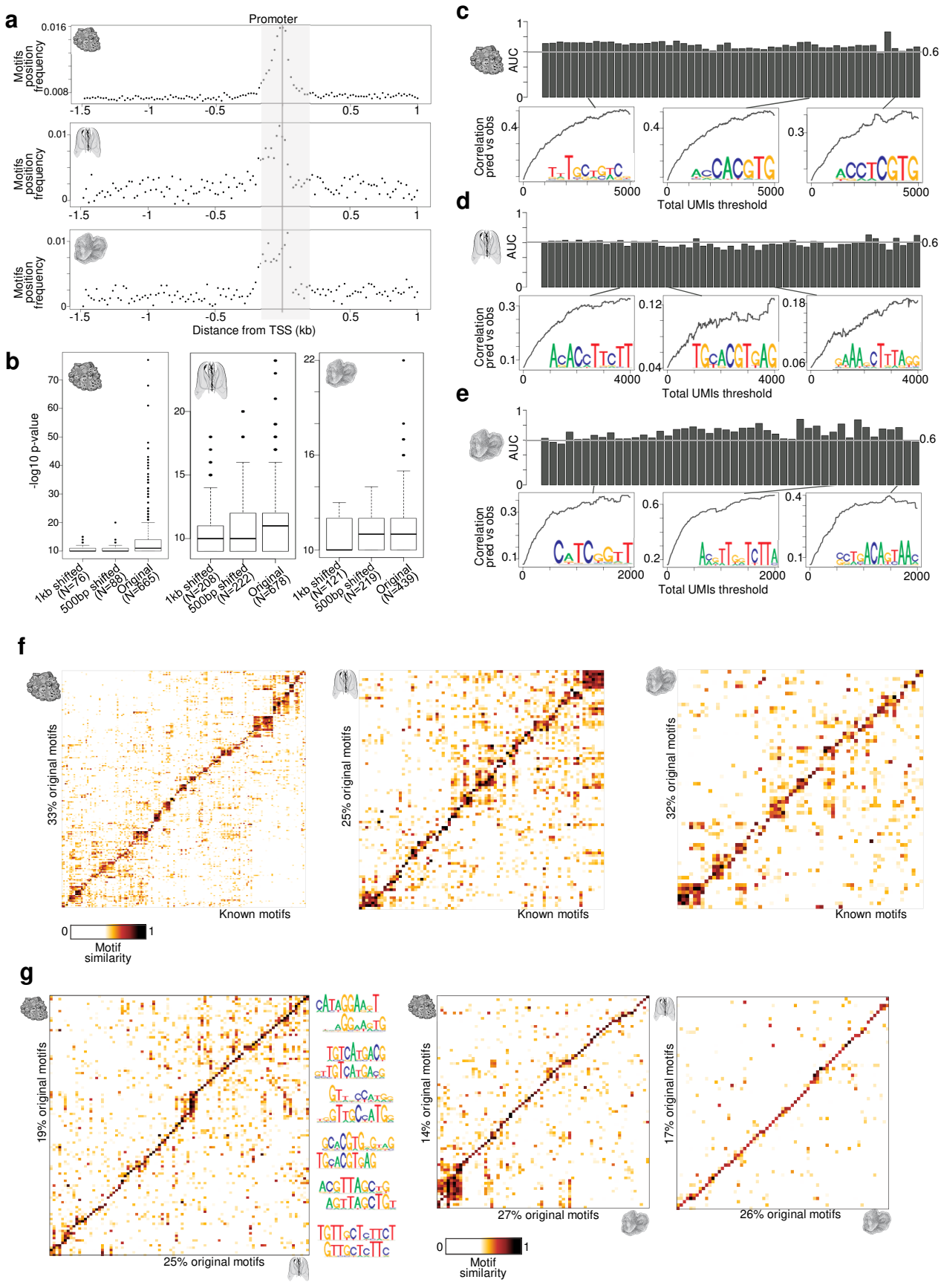
Supplementary Figure 5. Transcription factor regulation in *A.queenlandica*. a, Table showing, for each species, the number of TFs encoded in the genome, the number of TFs detected and the number of cell type-specific TFs (showing max fold-change > 2 in at least one metacell). Also indicated are the percentage of cell type-specific TFs (compared to the total detected TFs) and, for comparison, the percentage of cell type-specific genes. Notice that TFs are significantly more cell type-specific than the general gene population ($p<<0.001$, chi-square test) in all three species. **b,** *A.queenslandica* general TF map with names and gene IDs, showing TF expression profiles across metacells (left heatmap) and TF-TF correlation based on expression profile across metacells (right heatmap). TF names and IDs are indicated and color-coded according to TF structural class. **c,** Same as **b** for *A.queenslandica* larva.

Supplementary Figure 6



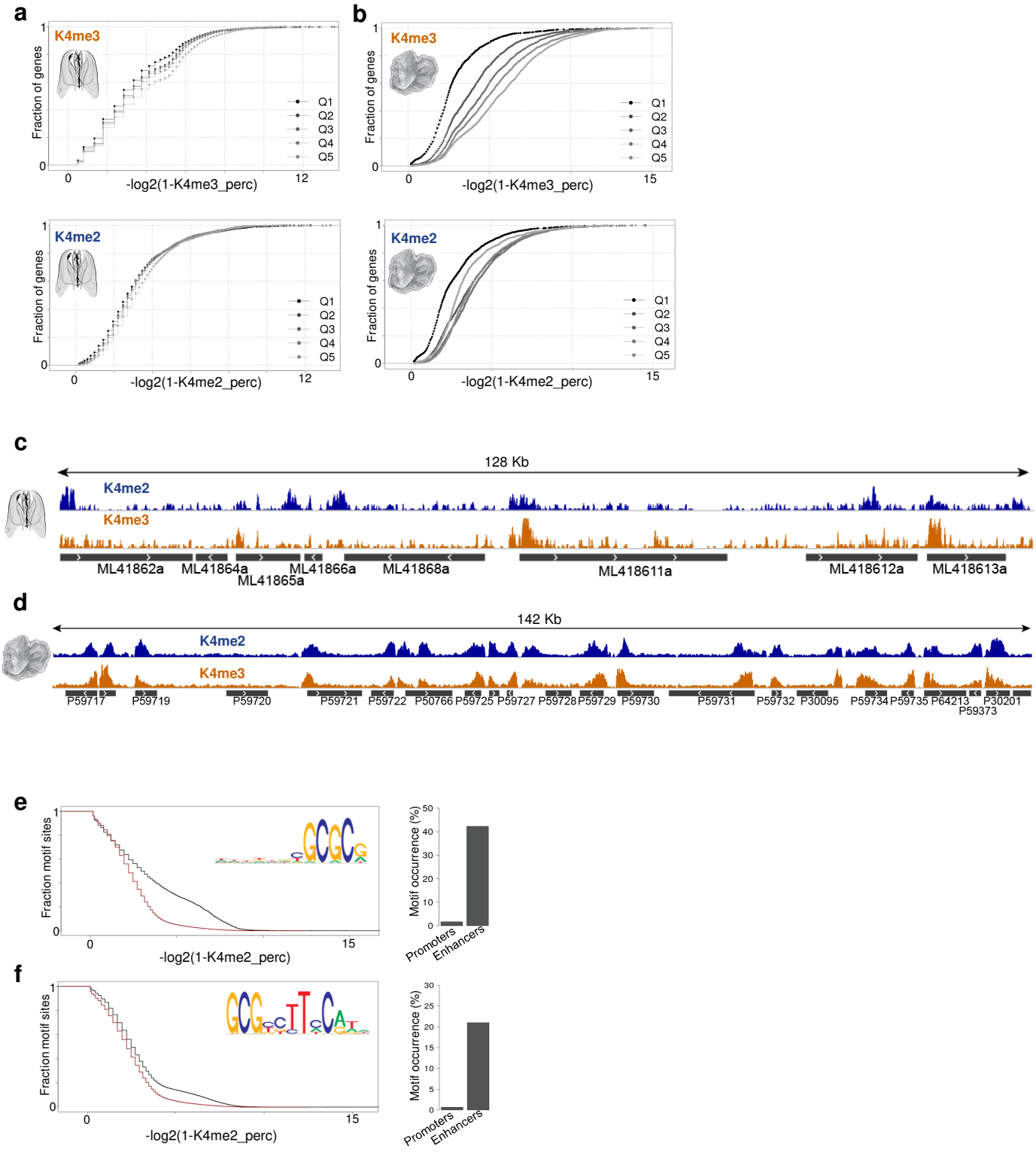
Supplementary Figure 6. Transcription factor regulation in *M.leidyi* and *T.adhaerens*. a, *M.leidyi* general TF map with names and gene IDs, showing TF expression profiles across metacells (left heatmap) and TF-TF correlation based on expression profile across metacells (right heatmap). TF names and IDs are indicated and color-coded according to TF structural class. **b,** Same as **a** for *T.adhaerens*.

Supplementary Figure 7



Supplementary Figure 7. Motif enrichment supplementary analyses. a, Promoter motif positional distribution. The plots show, for each species, the averaged maximum position in 20bp bins of all motifs (shown in Fig. 3A) around the TSS. **b,** Distribution of p-values obtained in *de novo* motif enrichment analysis for each species. Values are shown for the original promoter set and, as a control, for promoters shifted 500bp upstream and for promoters shifted 1Kb upstream. **c,** Cross-validation analysis for *A.queenslandica* promoter motifs. The barplot shows, for each metacell, the area under the curve (AUC) value of gene expression values predicted using a linear model. This model was trained with 80% of the dataset (training set) and applied to the other 20% of the genes (test set), repeated 5 times to cover all genes in test sets. The plots below show 3 examples of the relationship between observed/predicted gene expression correlation and total gene molecule count thresholding. Notice that the accuracy of the prediction increase as we filter out lowly expressed genes. In each case, the motif with the highest coefficient in the linear model is indicated. **d, e,** Same as **c** for *M.leidyi* and *T.adhaerens*. **f,** Database motif similarities. Left, heatmap showing the similarity between *A.queenslandica de novo* predicted motifs and known motifs found in public databases. Only motifs with similarity >0.7 with any motif in the other species are shown. Middle, heatmap showing similarity between *M.leidyi* and known motifs. Right, heatmap showing similarity between *T.adharens* and known motifs. **g,** Cross-species motif similarities. Left, heatmap showing the similarity between *A.queenslandica de novo* predicted motifs and *M.leidyi de novo* predicted motifs. Only motifs with similarity >0.7 with any motif in the other species are shown. Examples of similar motifs are shown next to the heatmap (top motif, *A.queenslandica;* bottom motif, *M.leidyi).* Middle, heatmap showing similarity between *A.queenslandica* and *T.adharens* motifs. Right, heatmap showing similarity between *M.leidyi* and *T.adharens* motifs.

Supplementary Figure 8



Supplementary Figure 8. H3K4me2/me3 iChIP supplementary analysis. a, Relationship between fraction of cells expressing a gene and K4me3/me2 iChIP signal in the gene promoter for *M.leidyi*. Genes are grouped into five quantiles, from smaller (Q1) to higher (Q5) cell fraction. **b,** Same as **a** for *T.adhaerens*. Notice that the H3K4me2/3 promoter signal increases for genes expressed in a larger fraction of cells ($p<<0.001$, Wilcoxon rank-sum test). This shows, as expected, that whole-organism iChIP preferentially detects regulatory features present in all or at least the most abundant cell types. **c,** Example *M.leidyi* genomic region showing normalized H3K4me2 and H3K4me3 iChIP coverage. **d,** Example *T.adhaerens* genomic region showing normalized H3K4me2 and H3K4me3 iChIP coverage. **e,** *M.leidyi* global enhancer motif enrichment. Left, cumulative distribution of H3K4me2 iChIP signal in non-promoter genomic regions (red) and in non-promoter genomic regions with the presence of *M.leidyi* enhancer-motif 1 (represented in the logo). This suggests strong co-localization of this motif and non-promoter H3K4me2. Right, barplot showing the frequency of *M.leidyi* enhancer-motif 1 occurrence in promoters and enhancers. Notice that >40% of *M.leidyi* enhancers present this motif. **f,** Same as **e** for *M.leidyi* enhancer-motif 2. In this case, the motif is present in >20% of enhancers.