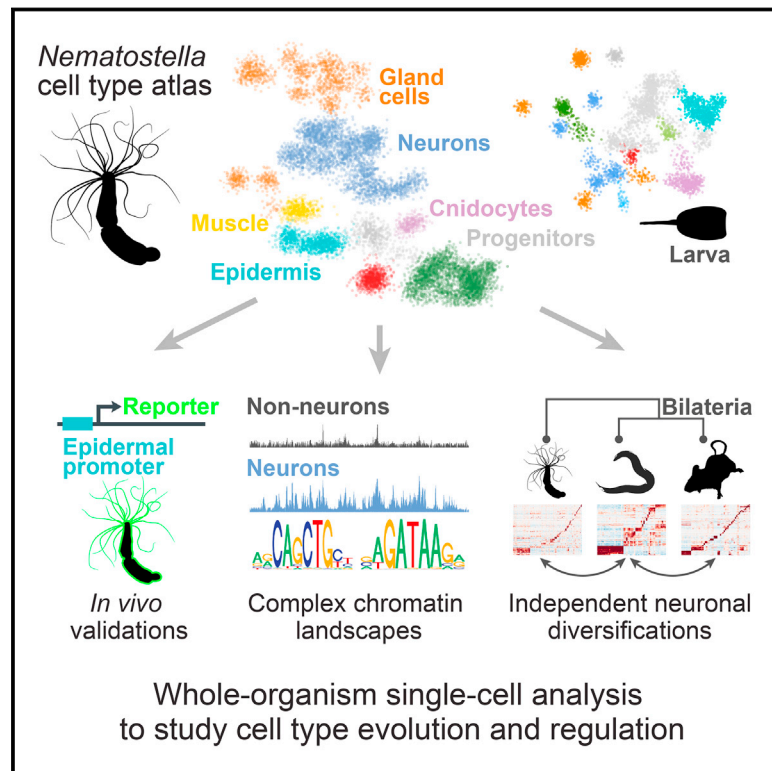


Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq

Graphical Abstract



Authors

Arнау Seb e-Pedr s,
Baptiste Saudemont, Elad Chomsky, ...,
Fran ois Spitz, Amos Tanay,
Heather Marlow

Correspondence

amos.tanay@weizmann.ac.il (A.T.),
heather.marlow@pasteur.fr (H.M.)

In Brief

An integrative whole-organism single-cell analysis in *Nematostella vectensis* reveals cnidarian cell type complexity and provides insights into the mechanisms of animal cell-specific genomic regulation and evolution.

Highlights

- Integrative single-cell analysis of whole-organism cell type regulatory programs
- *Nematostella vectensis* adult and larval cell atlas reveals high cell type diversity
- Independent diversification of neuronal cell type repertoires in Cnidaria/Bilateria
- Pre-bilaterian complex *cis-regulatory* landscapes linked to cell type specification



Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq

Arnaud Seb -Pedr s,¹ Baptiste Saudemont,^{2,3} Elad Chomsky,¹ Flora Plessier,^{2,3,4} Marie-Pierre Mailh ,^{2,3} Justine Renno,^{2,3} Yann Loe-Mie,^{2,3} Aviezer Lifshitz,¹ Zohar Mukamel,¹ Sandrine Schmutz,⁵ Sophie Novault,⁵ Patrick R.H. Steinmetz,⁶ Fran ois Spitz,^{2,3} Amos Tanay,^{1,7,*} and Heather Marlow^{2,3,*}

¹Department of Computer Science and Applied Mathematics and Department of Biological Regulation, Weizmann Institute of Science, 76100 Rehovot, Israel

²(Epi)genomics of Animal Development Unit, Department of Developmental and Stem Cell Biology, Institut Pasteur, 75015 Paris, France

³CNRS, UMR3738, 25 Rue du Dr Roux, 75015 Paris, France

⁴D partement de Biologie,  cole Normale Sup rieure de Lyon, 46 All e d'Italie, 69364 Lyon Cedex 07, France

⁵Cytometry & Biomarkers UtechS, Cytometry Platform, Institut Pasteur, 75015 Paris, France

⁶Sars International Centre for Marine Molecular Biology, University of Bergen, Thorm hlensgate 55, Bergen 5006, Norway

⁷Lead Contact

*Correspondence: amos.tanay@weizmann.ac.il (A.T.), heather.marlow@pasteur.fr (H.M.)

<https://doi.org/10.1016/j.cell.2018.05.019>

SUMMARY

The emergence and diversification of cell types is a leading factor in animal evolution. So far, systematic characterization of the gene regulatory programs associated with cell type specificity was limited to few cell types and few species. Here, we perform whole-organism single-cell transcriptomics to map adult and larval cell types in the cnidarian *Nematostella vectensis*, a non-bilaterian animal with complex tissue-level body-plan organization. We uncover eight broad cell classes in *Nematostella*, including neurons, cnidocytes, and digestive cells. Each class comprises different subtypes defined by the expression of multiple specific markers. In particular, we characterize a surprisingly diverse repertoire of neurons, which comparative analysis suggests are the result of lineage-specific diversification. By integrating transcription factor expression, chromatin profiling, and sequence motif analysis, we identify the regulatory codes that underlie *Nematostella* cell-specific expression. Our study reveals cnidarian cell type complexity and provides insights into the evolution of animal cell-specific genomic regulation.

INTRODUCTION

Non-bilaterian animal lineages, including cnidarians, ctenophores, sponges, and placozoans, have simple body plans and have been historically considered to contain limited numbers of cell types. In cnidarians, these cells include a small number of morphologically distinct neurons, gland cells, muscle cells, epidermis, and gut (Frank and Bleakney, 1976). This presumed simplicity in the number of cnidarian cell types stands in marked

contrast to their genomic complexity. Indeed, multiple cnidarian genomic features, such as the gene repertoire, syntenic gene blocks, and intronic structure, are more similar to vertebrates than to model bilaterian invertebrates like *Drosophila melanogaster* or *Caenorhabditis elegans* (Putnam et al., 2007). Recently, it was shown that these similarities extend to the regulatory landscape of *Nematostella* genes (Schwaiger et al., 2014). However, the extent to which these genomic features participate in regulating cell type hierarchies remains largely unknown.

Gene expression profiling by *in situ* hybridization (ISH) has allowed for comparative study of cell types and tissue organization in different species (Steinmetz et al., 2012). However, these approaches require *a priori* selection of candidate gene markers; they are difficult to scale toward multiple expressed genes simultaneously, and they are not readily applicable to all species or life stages—in particular, adult specimens. On the other hand, techniques for genome-wide profiling of gene expression were so far dependent on established staging and tissue dissection procedures, which are not possible in animals that lack defined organs, such as cnidarians. Single-cell RNA sequencing (scRNA-seq) is rapidly emerging as a powerful approach for unbiased *de novo* discovery and detailed molecular characterization of transcriptional states and cell types in mammalian tissues (Jaitin et al., 2014; Tasic et al., 2016) and even whole organisms (Cao et al., 2017), opening the way to the systematic molecular characterization of cell types and regulatory programs in poorly sampled metazoan lineages.

The phylogenetic position of cnidarians as sister group of Bilateria makes them a particularly interesting lineage to study the evolution of cell type regulation in animals. *Nematostella vectensis* has emerged as a laboratory model species for cnidarians (Rentzsch and Technau, 2016). Here, we generated over 17,000 adult and larval scRNA-seq profiles to derive a detailed map of *Nematostella* cell types. We combine this map with reporter assays, chromatin profiling, and motif analysis to characterize the transcription factor (TF) combinatorics and distal regulatory element usage underlying *Nematostella* cell type



hierarchy. Together with an initial cross-species comparative analysis, these data offer insights into the molecular paths that have allowed cell type diversification during animal evolution.

RESULTS

An Atlas of Adult *Nematostella vectensis* Cell Types

In order to generate a comprehensive map of cell-state transcriptional profiles in *Nematostella* (Figure 1A), we applied MARS-seq (Jaitin et al., 2014) to whole adults using fluorescence-activated cell sorting (FACS) to distribute live cells (calcein positive, propidium-iodide negative) into 384-well plates and performed labeling, amplification, sequencing, and de-noising as previously described (Jaitin et al., 2014). Because of the diversity and overall small size of cells composing an adult *Nematostella*, this whole-organism analysis involved dramatic variability in cell size and total RNA content within the data cohort (Figure S1A). We therefore lowered the coverage threshold on single cells to 100 molecules and developed an analytic pipeline that combines non-parametric construction of k-nearest-neighbor graphs, detection of “metacells” (see STAR Methods), and multiple layers of validation. We retained for analysis 11,888 cells that showed sufficiently high quality control metrics, with a median number of 541 RNA molecules per cell (Figures S1A and S1B and Table S1). Using a combined genome annotation scheme, we mapped 81% of the MARS-seq sequences to the genome (Table S1) and detected expression for 16,853 genes (66% of the total predicted genes).

To identify genes with potential cell-type-specific expression while accounting for variability in cell size, we analyzed expression variance across cells (Figure S1C) and the correlation of individual gene expression with the total RNA content per cell (Figure S1D). We selected 700 genes with high overall expression but low correlation with total RNA content, providing features that sensitively define similarity between sparse single-cell profiles. We observed that 99.5% of the cells are expressing at least seven of the selected markers (Figure S1G). Using these gene features, we identified coherent groups of single cells (or metacells; see STAR Methods) and pooled RNA data within each of them to derive 104 rich transcriptional signatures representing the diversity of *Nematostella* cellular programs (Figures 1B, 1C, S1I, and S1J and Table S2). We performed bootstrap and subsampling analysis to quantify metacell robustness (Figures S1K and S1L). This demonstrated that despite the broad distribution of the number of molecules per single cell in our data (Figure S1A), single cells retain high information content and could be grouped robustly into a large number of distinct molecular states.

Broad Classification of *Nematostella* Adult Cell Types

Based on a graph-based 2D projection plot (Figure 1B) and direct visualization of gene distributions over cells (Figure 1C), we organized the complex transcriptional landscape in *Nematostella* into eight broad metacell groups (Figure S1M), including cnidocytes, gland cells, neurons, retractor muscles, gastrodermis, and digestive filaments (Frank and Bleakney, 1976; Jähnel et al., 2014; Steinmetz et al., 2017). We linked three of these groups with specific phenotypic classes previously described

in anthozoan cnidarians, such as *Nematostella*. The first are cnidocyte cells, highly specialized evolutionary novelties that enable prey capture and defense, which can be easily recognized by the co-expression of multiple venom and capsule proteins, such as minicollagens, NEP proteins, and nematogalectins (Moran et al., 2013) (Figure 1D). Gland/secretory cells can be identified by expression of toxins, proteases, and digestive enzymes, linking together otherwise highly diverse transcriptional states based on functional similarity (Steinmetz et al., 2017). Finally, neuronal cell types represent a rich and diverse group of metacells that are all marked by the previously described neuronal marker *elav* (Nakanishi et al., 2012), together with multiple additional neuronal markers (Figures 1D and 1E).

Three additional metacell groups represent tissue-grade assemblages of regionalized ectodermal and endodermal germ-layer derivatives. The molecular signatures of these groups provide valuable new insights into the biology of these cell assemblages. For example, in line with fate-mapping results (Steinmetz et al., 2017), our scRNA-seq data indicate that the inner epithelial lining and mesenterial folds (gastroderm) can be subdivided into two distinct clusters that we refer to as “digestive filaments,” where prey is held and digested, and the gastrodermal lining that is defined here as gastrodermis. The digestive filaments include epithelial and cilioglandular structures of the mesenteries and show expression of the cilia-associated TF *rfx*, *hedgehog*, a transient receptor potential (TRP) channel, and the MPP5 MAGUK protein (Figure 1D and Table S2). In contrast, the gastrodermis includes the parietal and ring epitheliomuscular cells and corresponds to signatures of genes associated with musculature, including myosin heavy and light chains, and several collagens. A sixth metacell group represents a subset of cells enriched in *striated-type myosin II*, previously described in *Nematostella* as distinct “retractor muscles” and “tentacle longitudinal muscles” (Frank and Bleakney, 1976; Jähnel et al., 2014).

Two additional metacell groups were defined by specific transcriptional states but did not obviously match any previously characterized *Nematostella* cell type. The first group showed expression of adhesion molecules such as claudins and cadherins, as well as multiple uncharacterized proteins specific to cnidarian/anthozoan/*Nematostella* (Figure 1D). By examining these markers through ISH, we found that they labeled the tentacle and body wall epidermal cells (Figure 1F). To better characterize these cells, we generated a transgenic reporter construct using 1.64 kb upstream of the TSS of *ep2a* gene (Figure 1G). Reporter expression was observed in the developing epidermis of the planula and was maintained at the polyp stage in contiguous patches of the outer epithelium (Figure 1H). Single-cell expression data (Figure 1D), corroborated by the spatial localization of specific marker genes (Figure 1F) and *ep2a* reporter-injected animals (Figure 1H), identify these metacells as non-neuronal epithelial cells forming the epidermis.

The second uncharacterized group of metacells was much less cohesively defined by co-expression signatures than other clusters (Figure S1M) but showed expression of several factors associated with multipotency or progenitor states in other species, including *nanos1*, two *myc* paralogs, *mago nashi*, *soxB2a*,

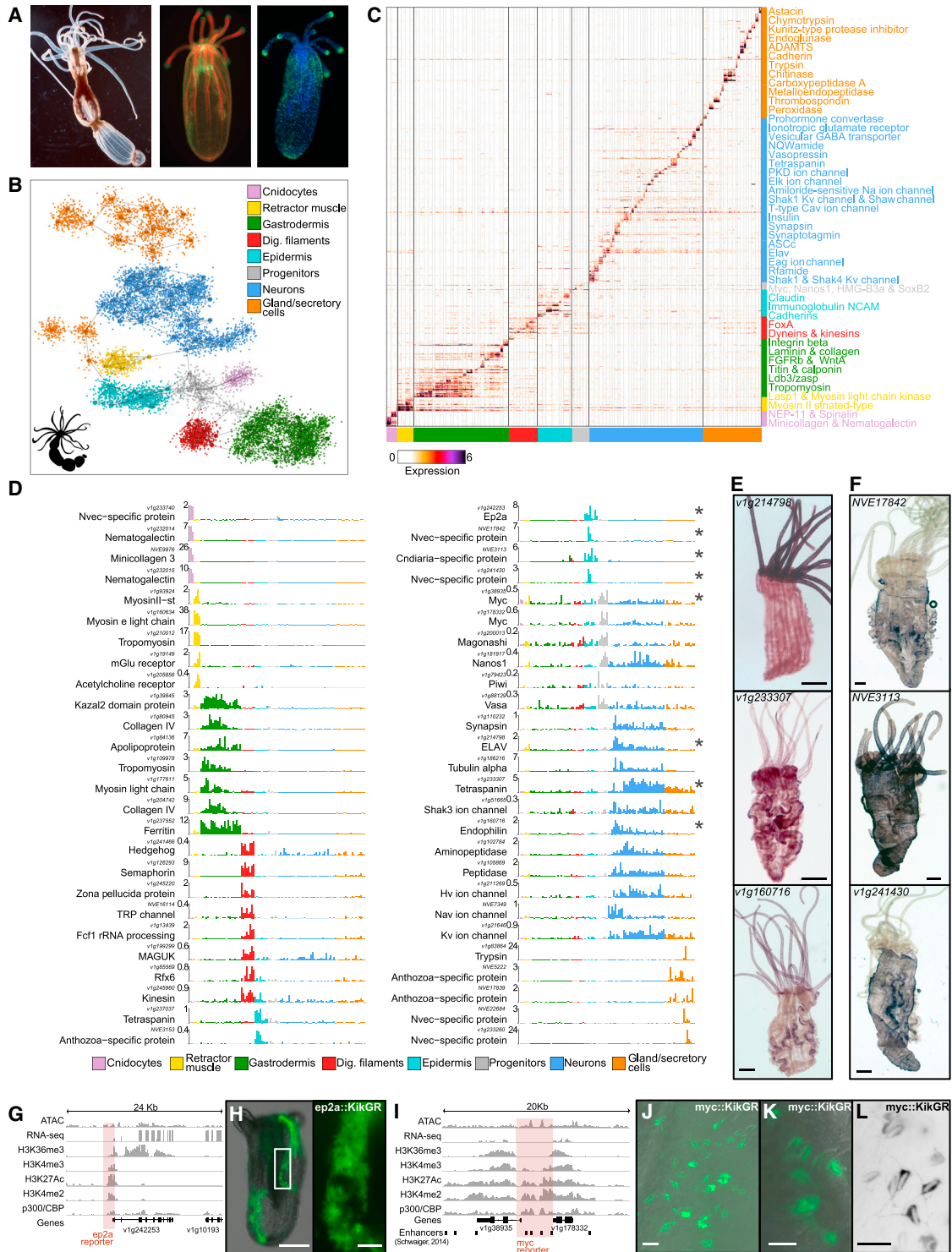


Figure 1. Whole-Organism Dissection of Adult *Nematostella* Cell Types

(A) Adult *Nematostella* polyp (left). Juvenile polyp stained against F-actin (red) and acetylated tubulin (green) (middle). Juvenile polyp stained against nuclei (blue) and nematocysts (green) (right).

(B) 2D projection of *Nematostella* 104 metacells and 10,460 cells.

(C) Expression of 904 variable genes (rows) across 10,460 cells sorted by metacell and broad cluster association.

(D) Expression (molecules/1,000 UMIs) of selected genes in each metacell (*, genes in E–L). Scale bars: 500 μ m.

(legend continued on next page)

and *hgmb3a* (Figure 1D). To gain initial insights into this population, we generated a reporter construct consisting of 2.85 kb upstream of the TSS of a *myc* paralog expressed in these metacells (*myc_v1g38935*), including several enhancer elements (Figure 1I). *myc* TFs have been shown to serve as critical regulators of cell proliferation and growth across animals, including in the model cnidarian *Hydra* (Hartl et al., 2010). In *Nematostella*, we found that the *myc_v1g38935* reporter labels several populations of cells located in the outer epithelium of the animal, as well as large numbers of cnidocytes (Figures 1J–1L). These initial observations and the markers reported here open the way to further detailed characterization of the potential progenitor developmental dynamics in these *Nematostella* adult cell populations.

Characterization of Larval Cell Types

Many invertebrates, including some cnidarian species like *Nematostella*, progress through a planktonic dispersal life history phase known as the larval form. Larval forms are typically characterized by ciliary swimming and the presence of the apical organ or tuft—an enigmatic ciliated structure thought to be of neural identity and involved in settlement and metamorphosis (Nielsen, 2005). In *Nematostella*, the transition from the larval planula to the adult polyp has been termed “minimally indirect development” to highlight the gradual transition between these stages (Marlow et al., 2014). However, so far, it was unclear to what extent larval cell types persist during the transition from larval to polyp stages. To examine the relationship between the larval and adult cell types, we sampled 5,281 single-cell transcriptomes from larva 2, 4, and 7 days post fertilization, corresponding to gastrula, early planula, and planula stages, respectively. We computationally pooled and analyzed cells from all larval developmental stages together following the same strategy employed for the adult data to organize single cells into metacells (Figures 2A and S2A–S2K). Each larval metacell was supported by hundreds of specific gene markers consistently expressed across single cells (Figure 2B and Table S3) and strong single-cell co-associations in bootstrap analysis (Figure S2H).

In order to compare adult and larval transcriptional states, we defined for each larval metacell its maximum correlation with an adult metacell. We then compared the larva-adult correlation to the overall larval transcriptional specificity (Figure 2C). We observed two clusters (metacells 23 and 37) with high gene-expression specificity but low similarity to any adult cell cluster, indicating that these states represent larval-specific cell types. The first of these metacells (37) expresses the TFs *foxQ2b*, *vsx*, *ASCc*, and *soxB2a*, together with cyclic nucleotide-gated

(CNG), TrpA, polycystic kidney disease (PKD), and two shaker ion channels, all implicated in putative neuronal functions. This suggests that this larval transcriptional state represents a previously uncharacterized larval-specific neuronal cell type. The second larval-specific metacell (23) is defined by previously identified markers of the apical organ/tuft, such as *fgf* ligand (Sinigaglia et al., 2015) (Figures 2D and S2I), and specific TFs, such as *coe*, *nk3*, and an unclassified paired-class homeobox (Figure 2D). Interestingly, we observe lack of co-expression of any recognizable neuronal effector genes (e.g., ion channels, GPCRs, synaptic scaffold proteins, neurotransmitter-related enzymes, and neuropeptides) in this cell cluster. In fact, we found 0 genes with fold change (FC) ≥ 2 in the apical organ metacell out of 122 potential neuronal effector genes differentially expressed (maxFC ≥ 2) in the larval dataset ($p < 0.001$, hypergeometric test). Thus, our data molecularly characterize apical organ cells and suggest that they are not neuronal cells, contrary to what was previously assumed (Marlow et al., 2014).

A second group of larval metacells shows low similarity to adult cell types, but because these lack strong effector gene signatures (Figure 2C), we hypothesized that these metacells correspond to undifferentiated states. Supporting a precursor/progenitor identity, these metacells fall into two subgroups defined by co-expression of TFs previously associated with precursor/progenitor populations. A first group of metacells shows co-expression of two *hes* genes and *soxB2* and is enriched for gastrula-stage cells (Figure 2E). Another group of progenitor/precursor cells is characterized by the expression of *soxB2a*, two *myc* paralogs, and *nanos2* and is more frequent in post-gastrula stages (Figure 2E).

In addition to the larval-specific cell types and undifferentiated cell populations, we found that 63.4% of the larval cells show strong affinities with adult cell clusters (Figures 2C–2F). These affinities are demonstrated by the expression profiles of key TFs across larva and adult stages (Figure 2G). For example, expression of *paxA* and *TEA/scalloped* mark both larva and adult cnidocytes, while expression of *tbx1/10-1* links gastrodermal cells in both larva and adult. Two matched larva-adult gland/secretory metacells are defined by expression of the *oasis1* TF and *rx* homeobox TF (with both gland/secretory metacells expressing the *rfx4* TF) (Figure 2G). We did not identify larval retractor/tentacle muscle cells, as these cells differentiate at later stages (Jahnel et al., 2014). In conclusion, *Nematostella* larval- and adult-differentiated cell type transcriptional programs are remarkably related, at the level of both effector and regulatory genes. This indicates that limited rearrangement of cell types occurs during metamorphosis in *Nematostella*.

(E) ISH for three genes associated with neuronal metacells.

(F) ISH for three genes associated with epidermal metacells.

(G) Chromatin features around the *ep2a* gene. Reporter region in red.

(H) Polyp with chimeric expression of a reporter for the *ep2a* gene in epidermal patches. Scale bars: 100 μm (20 μm , inset).

(I) Chromatin features around *myc_v1g38935* gene. Reporter region in red.

(J) *myc_v1g38935* reporter expression. Cnidocytes with developing capsules and rounded cells within the outer epithelium are visible.

(K) Cnidocytes and cells in the epithelium labeled with the *myc_v1g38935* transgene.

(L) Labeled cnidocytes with visible projections in the epithelium. Scale bars: 20 μm .

See also Figure S1 and Tables S1 and S2.

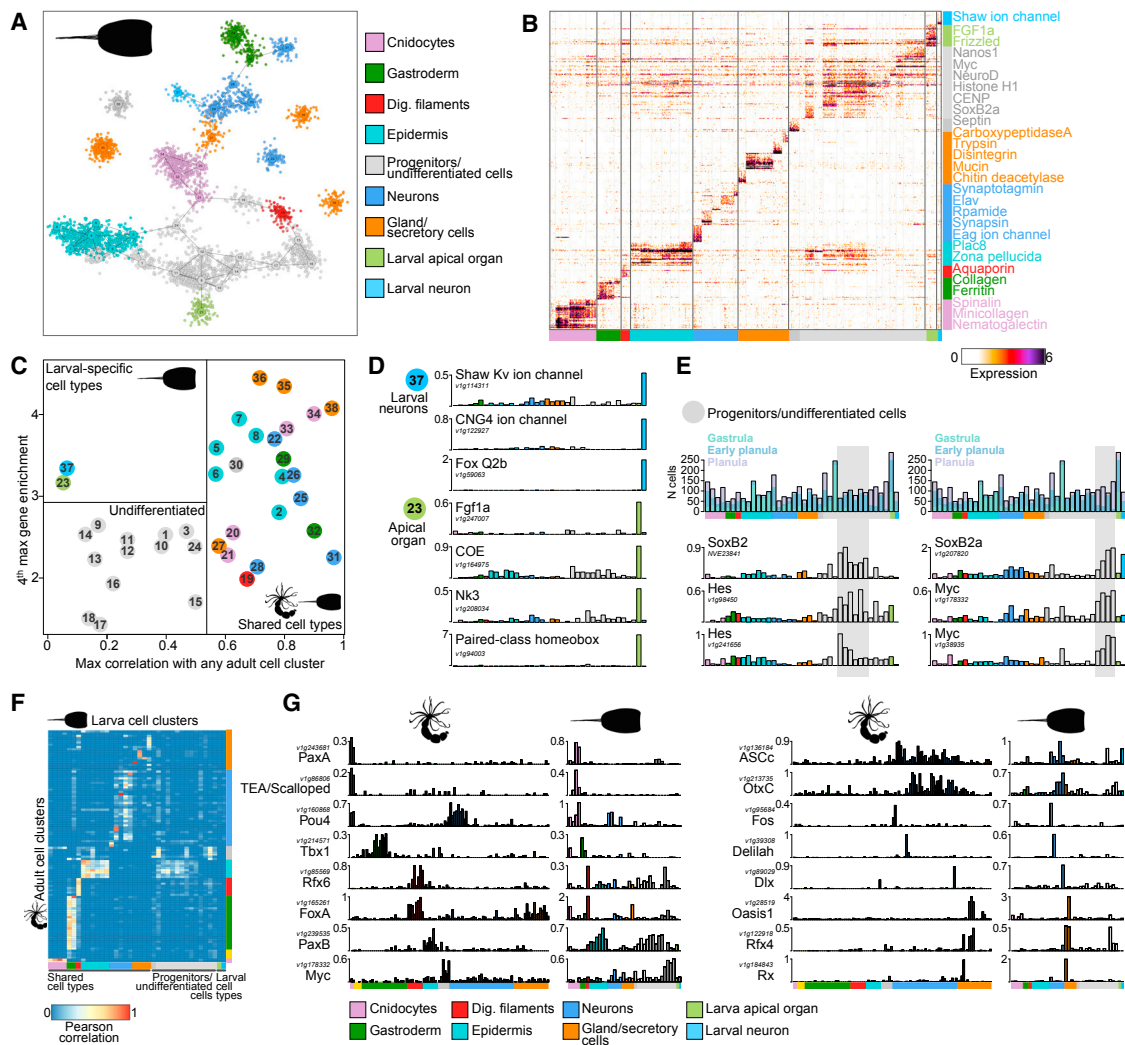


Figure 2. *Nematostella* Larval Cell Types

(A) 2D projection of *Nematostella* larva 38 metacells and 4,146 cells.

(B) Expression of 367 variable genes (rows) across 4,146 cells.

(C) Larval-adult metacell maximum correlation versus fourth maximum observed gene enrichment.

(D) Expression (molecules/1,000 UMIs) of selected markers for larval-specific cell types.

(E) (Top) Number of cells and larval stage composition of each metacell. (Bottom) Expression of selected markers for larval progenitors.

(F) Correlation between adult (rows) and larval (columns) metacells.

(G) Expression of selected TFs across adult and larva metacells.

See also [Figure S2](#) and [Tables S1](#) and [S3](#).

Comparative Analysis of Cell-Type-Specific Gene Repertoires

To facilitate comparative analysis of the *Nematostella* cell type map, we embedded each predicted *Nematostella* protein within orthology groups, including proteomes of species spanning a wide phylogenetic spectrum ([Figure 3A](#)). We compared the depth of protein conservation to the estimated cell type specificity of the gene encoding it, which we defined as the maximal fold enrichment observed across the metacell model. We found that deeply conserved genes generally show low cell type specificity, while new genes coding for Cnidaria-specific proteins are strongly enriched for tissue-specific expression

([Figure 3A](#); $p < 0.0001$ Wilcoxon rank-sum). Complementarily, genes that are expressed specifically in each of the eight broad metacell groups described above showed specific gene age distributions, reflecting, for example, enrichment of cnidarian-specific genes in cnidocytes ([Figure 3B](#)). Importantly, neurons are significantly enriched in genes that originated at multiple evolutionary times within Metazoa, suggesting a stepwise assembly and specialization of the neuronal toolkit from ancestral pan-metazoan genes to lineage-specific gene acquisitions ([Liebeskind et al., 2017](#)).

To gain further insight into the functional affinities of *Nematostella* broad cell classes, we compared the correlations of their

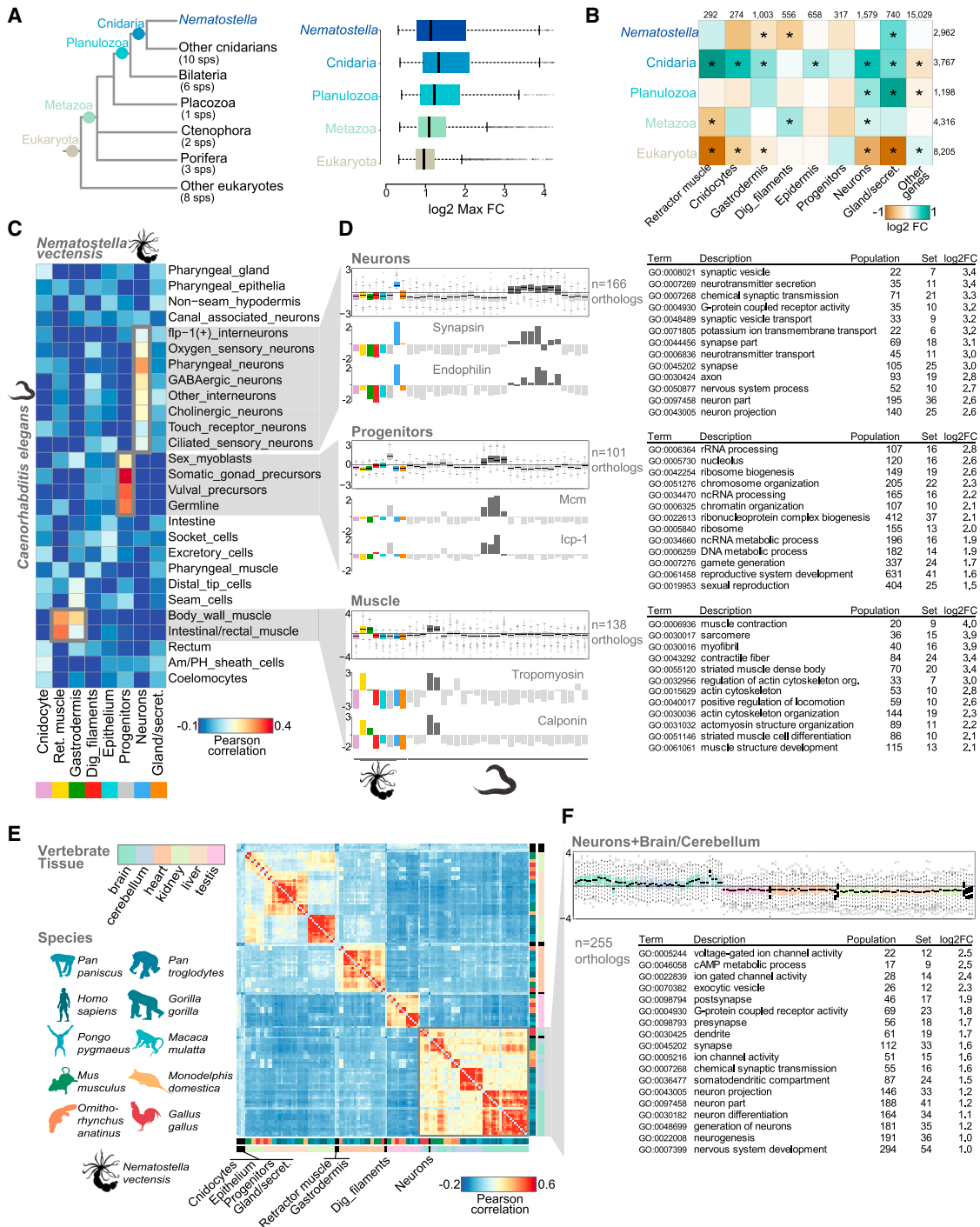


Figure 3. Phylogenetic Patterns of Cell-Type-Specific Gene Repertoires

- (A) Expression variability stratified by gene age.
 (B) Gene age enrichment/depletion in gene sets specific to each broad cluster (* $q < 0.001$; chi-squared test).
 (C) Correlation between *Nematostella* (columns) and *C. elegans* cell clusters. *C. elegans* data from Cao et al., (2017).
 (D) (Left) Expression of orthologs driving specific cross-species similarities. (Right) Enriched Gene Ontology (GO) terms among co-expressed orthologs.
 (E) Correlation between organs/tissues of different vertebrate species and *Nematostella* broad cell clusters. Vertebrate data from Brawand et al., (2011).
 (F) (Top) Expression orthologs driving neuron and brain/cerebellum similarity. (Bottom) Enriched GO terms among neuron/brain co-expressed orthologs.

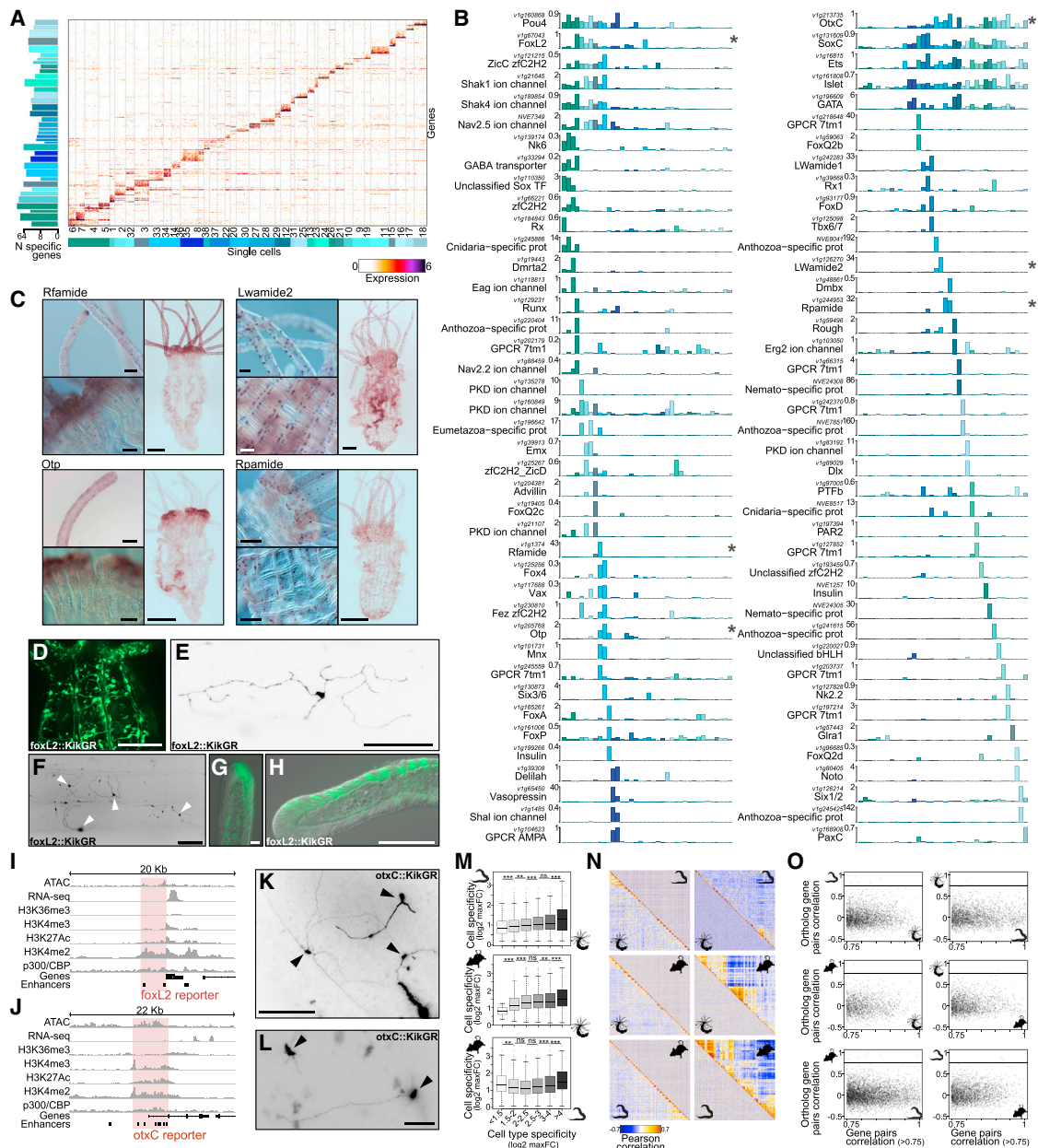


Figure 4. *Nematostella* Neuronal Cell-Type Diversity

(A) Expression of 631 variable genes (rows) across 3,485 cells sorted by metacell association.

(B) Expression (molecules/1,000 UMIs) of selected markers (*, genes in C–L).

(C) ISH for three neuropeptides and *otp* TF. Scale bars: 500 μ m (100 μ m, insets).

(D–F) Expression of *foxL2* transgenic reporter in neurons of the body column and tentacles (D), a single tripolar neuron (E), and several neurons (F) in a different region of the body column of the animal in (E) (arrowheads, cell bodies).

(G and H) *foxL2* reporter expression in cnidocytes (G) and neurons (H) in the tentacle.

(I) Chromatin features around *foxL2* gene.

(J) Chromatin features around *otxC* gene.

(K) Neurons in a polyp expressing *otxC* reporter (arrowheads, cell bodies).

(L) Neurons in the aboral pole of a polyp expressing the *otxC* reporter.

(G–L) Scale bars: 100 μ m (20 μ m, G and L).

(M) Neuronal gene specificity values in one species (y axis) stratified by the specificity of the orthologous genes in the other species (x axis) (***) $p < 0.001$; ** $p < 0.01$; ns, non-significant; Wilcoxon rank-sum test).

(legend continued on next page)

transcriptional states with orthologous expression profiles acquired from the model nematode *Caenorhabditis elegans* (Cao et al., 2017) (Figure 3C). Three cell type groups showed strong similarities between *Nematostella* and *C. elegans*. First, *Nematostella* retractor muscle and gastrodermal cells are associated with *C. elegans* intestinal and body wall muscles, a similarity driven by 138 orthologs enriched in muscle- and actin-based contractility functions—for example, tropomyosin and calponin (Figure 3D). Second, *Nematostella* precursor cells show similarities with *C. elegans* germline and precursor cells, supported by 101 shared orthologs. Some of these genes are associated with chromosomal organization functions. Specific examples include inner centromere protein (*icp1*), a key regulator of mitosis, and multiple minichromosome maintenance (*mcm*) helicase orthologs, which are associated with genome replication and cell proliferation. Finally, we observed 166 orthologs co-enriched in *Nematostella* and *C. elegans* neuronal types, including synapse structural components, ion channels, and GPCR receptors. However, despite conservation in cell-type-specific effector gene expression, we could not identify correspondingly conserved TFs. Further comparative analysis of published vertebrate-organ-specific transcriptomes (Brawand et al., 2011) again revealed similarities between *Nematostella* broad neuronal cluster and vertebrate brain/cerebellum tissues (Figure 3E), identifying 255 co-expressed orthologs (Figure 3F) strongly enriched in neuronal functions. In summary, despite the considerable evolutionary divergence of the cnidarian, nematode, and vertebrate lineages, we can identify correlated signatures of orthologous gene co-expression and support the broad conservation of cell-type-specific expression for some key cellular functions.

Diversity of Cnidarian Neuronal Cell Types

The richness of the scRNA-seq profile prompted us to examine transcriptional programs at high resolution within the broad and heterogeneous group of neuronal metacells. In accordance with previous reports, we confirm that *elav* and *ASCc* were broadly expressed across neuronal cells (Layden et al., 2012; Nakanishi et al., 2012) (Figure 1D). In addition, we identify a large diversity of neuronal-associated genes in *Nematostella*, including members of multiple ion channel families, *synapsin*, and peptide-processing enzymes (Figure 1D and Table S2). To improve the resolution of our neuronal cell map, we sorted and processed over 3,000 additional adult single cells using two existing *Nematostella* neuronal reporter transgenic lines driving mOrange expression under the control of *elav* promoter (Nakanishi et al., 2012) (Figures S3A, S3B, S3E, and S3G) and *soxB2a* promoter (Richards and Rentzsch, 2014) (Figures S3C, S3D, S3F, and S3H).

By combining *in vivo* and *in silico* sorted datasets, we assembled 4,775 adult and larva neuronal cells for in-depth analysis. We organized these cells into 32 metacells (Figures 4A and S4A) and found hundreds of marker genes with neuronal type-specific expression (Figure 4A and Table S4). In general, we

identified a wide diversity of ion channels (57) to be differentially expressed within the *Nematostella* neurons, thus highlighting the likely functional diversity of these 32 metacells. This includes 10 PKD proteins, which have been found to be employed in sensory functions in invertebrate nervous systems (Barr and Sternberg, 1999). Among these 32 cell clusters, we recognized two broad subsets of neuronal transcriptional states. The first group consists of neurons marked by expression of the TFs *foxL2* and *pou4* and the ion channels *NvShaker1*, *NvShaker4*, and *NvNav2.5* (Figure 4B; left) and includes specific types such as neurons expressing *otp* TF (Mazza et al., 2010) and *Rfamide* neuropeptide (Marlow et al., 2009) (Figure 4C).

The second broad group of neurons shares expression of *gata*, *otxC* and *islet* (Figure 4B; right). A recent study showed expression of *gata* in cells of the endodermal nerve net (Steinmetz et al., 2017). Among this group of neurons, we identify several peptidergic cell clusters, three of which express *Lwamide1* (Figure 4B), which has been shown by transgenic reporter assays to be expressed in adult neurons (Havrilak et al., 2017). We also identified two cell clusters expressing *LWamide2* and three distinct clusters expressing *RPamide* (Figure 4B). We performed ISH in adult polyps and found that *LWamide2* is expressed in both the tentacles and neural tracks running along the mesenteries, while *RPamide* is expressed in the body wall and is largely absent from the tentacles (Figure 4C). Another metacell belonging to the second broad class of neurons was identifiable by expression of the TF *foxQ2d*, which has previously been shown to be expressed in progenitor cells that give rise to putative sensory neurons (Busengdal and Rentzsch, 2017). We found *foxQ2d* expression to be linked with specific expression of additional TFs, including *noto1*, *unc4*, *NvNR1*, and *PTFb* (Figure 4B and Table S4).

In order to gain further insights into the localization of neurons within the two broad groups, we generated two reporter transgenic constructs for *foxL2* and *otxC* TFs (Figures 4B, 4I, and 4J). We observed strong and specific expression of the *foxL2* reporter in neurons of diverse morphologies, particularly in tentacles and the basiepithelial nerve net (Figures 4D–4H). For *otxC*, reporter expression could be visualized in bi-polar and tri-polar neurons in the basiepithelial nerve net of the body wall (Figures 4K and 4L).

Following our observation of broad cross-lineage similarities in neuronal molecular profiles (Figure 3), we asked whether the specific neuronal types we characterize in *Nematostella* have any cognates in mice and/or *C. elegans* (Cao et al., 2017; Tasic et al., 2016). We examined the expression profiles of gene orthologs across neuronal cell types in these three species and observed that similar groups of orthologs tend to be neuron-type specific in all pairwise comparisons (Figure 4M). Then, we analyzed the composition of neuronal gene modules based on gene-gene co-expression but could not identify shared gene modules in any of the cross-species pairwise comparisons (Figure 4N). Moreover, we could also not identify

(N) Gene-gene correlation values for gene modules in one species (left panels, lower triangle) compared to the correlation of gene orthologs in the other species. The right panels show the reciprocal analysis.

(O) Neuronal gene pair correlations in the focus species (x axis) compared to the orthologous gene pair correlations in the other species (y axis). See also Figures S3 and S4 and Table S4.

conserved pairs of co-expressed orthologs between species (Figure 4O). These results suggest that, despite broad similarities between neurons across species (Figure 3), the diversification of specific neuronal cell types is largely lineage specific—at least at the evolutionary distances considered here—and that different neuronal-type radiations assembled specific gene modules.

Characterization of Muscle, Cnidocyte, and Gland/Secretory Cell-Type Repertoires

We next performed in-depth analysis of the transcriptional states represented by cnidocyte, muscle/gastrodermis, and gland/secretory metacells. Two cnidocyte types, denoted as spirocytes and nematocytes, have been described in *Nematostella* (Frank and Bleakney, 1976). Spirocytes are characterized by the absence of spindle and the lack of *spinalin*, NEP toxins, and *minicollagen1* expression (Zenkert et al., 2011). Our analysis found that spirocytes (metacell 1) (Figures 5A and S4B and Table S5) lack expression of *spinalin* and NEPs, and instead, they express *minicollagen5* and *minicollagen v1g819141*, as well as the TF *foxL2* (Figure 5C; also supported by the *foxL2* reporter line, Figure 4G). Nematocytes (metacells 3–5) showed high levels of *spinalin*, multiple minicollagens (Zenkert et al., 2011), and several NEP toxins, including *NEP-19* (Figure 5D). Moreover, and unlike spirocytes, nematocytes are abundant in larval stages (Figure 5B). Interestingly, our results suggest some degree of specialization between nematocytes—e.g., with distinct *NEP-13* and *NEP-2/3* levels and a specific carboxypeptidase found in only one of nematocyte metacells (Figure 5D). In addition to nematocytes and spirocytes, we also identified intermediate states in cnidocyte biogenesis (metacells 8, 2, 6, and 7), showing weak cnidocyte effector gene signatures and instead specific expression of *nanos1*, *nanos2*, and *soxB2a* (Richards and Rentzsch, 2014).

Next, we performed in-depth analysis of gastrodermis and muscle metacells (Figures 5E–5H and S4C and Table S6). We identify metacells (1–5) as tentacle and longitudinal retractor myocytes (Renfer et al., 2010). Accordingly, these cells express high levels of *striated-type myosin II* (Steinmetz et al., 2012), *muscle LIM protein*, and a specific set of single regulatory and essential light chains (Figure 5G). ISH analysis for three of these markers (asterisks in Figure 5G) show specific expression in the myocytes along the longitudinal and tentacle retractor markers, as expected, and no enrichment in the epitheliomuscular cells of the gastrodermis (Figure 5H). In addition, we found that two TFs, *soxE1* and an unclassified bHLH *nem64*, and the RNA-binding protein *nova* to be highly enriched in these cells, forming a characteristic regulatory signature for these muscle cells.

Gastrodermal cells are characterized by shared expression a specific *tropomyosin* paralog and *non-muscle myosin II* (Figure 5G). Re-clustering helped to resolve the broad gastrodermal cluster into separate groups representing two morphologically distinct regions: “somatic gonad” and “epitheliomuscular” regions. We identify a group of six metacells (31, 24, 23, 22, 25, and 32) with enrichment of *foxC*, *nkx3/bagpipe*, *six4/5*, and *lysosomal lipase-2* genes (Figure 5G) previously shown to demarcate the somatic gonad, a nutrient storage tissue (Steinmetz et al., 2017). This cluster is clearly distinct from the other group of gastrodermal metacells (8–21), which broadly shares muscle

structural markers (distinct regulatory and essential light chains) and TFs (*hand*, *tbx20.1*, *tbx1/10-1*, and *tbx1/10-2*) (Figure 5G) previously shown to label parietal and circular epitheliomuscular cells (Steinmetz et al., 2017).

Finally, re-analysis of the extremely heterogeneous gland/secretory cells revealed multiple cell types expressing unique combinations of digestive enzymes and venom proteins (Figures 5I–5L and S4D and Table S7), at least four of which are shared between adults and larvae (Figure 5J). Supporting our classification scheme, previous observations coincide with cell behaviors identified here, such as co-expression of *trypsins A, B*, and *C* and co-expression of *chitinases A, B*, and *C*, but mutually exclusive expression of *pancreatic lipases 1* and *2* (*PanLip1,2*) (Steinmetz et al., 2017) (Figure 5K). Moreover, additional ISH analysis for three enzymes with restricted expression (asterisks in Figure 5K) shows expression in specific cells of the gut (Figure 5L). In summary, targeted re-analysis of specific cell groups allowed us to explore the fine-grained structure of the *Nematostella* cell type hierarchy, both covering previously described cell behaviors and discovering potentially new cell types.

Transcription Factor Modules Underlying *Nematostella* Cell-Type Hierarchy

TFs directly link regulatory information encoded in enhancers and promoters to transcriptional output and show high combinatorial specificity, making them key building blocks in gene regulatory networks (Davidson and Erwin, 2006). To test if the observed diversity of transcriptional programs is accompanied by expression of a comparably rich TF repertoire in *Nematostella*, we assessed the expression profiles of the 409 TFs detected in our single-cell cohort. Classification of these 409 *Nematostella* TFs into 38 structural families (Figure 6A) showed that the homeobox TF class constitutes 26% of the total TFs expressed in *Nematostella*. Moreover, analysis of evolutionary conservation showed that 85% of the identified TFs are metazoan innovations, with candidates for cnidarian or *Nematostella* novelties enriched for zfc2H2 TFs ($p < 0.0001$, chi-squared). 51% (207) of the expressed TFs showed evidence for metacell-specific expression (compared to 35% overall for all genes; $p < 0.0001$, chi-squared) and amount to 1.7% (median) of the total transcripts (Figure 6B). This rate of TF expression was, however, tissue specific, with 2-fold less TF expression in cnidocytes, muscle, and gastrodermis compared to neuronal and progenitor metacells ($p < 0.0001$, chi-squared).

Clustering of the TF-TF correlations among 207 TFs with metacell-specific expression (Figures 6C, S5A, and S5B) revealed a rich combinatorial structure, identifying 24 groups of correlated TFs consisting of 2–21 TFs each. Analysis of the TF expression profiles across metacells (Figure 6D) highlighted several TFs that are widely expressed within one of the eight broad clusters identified above, as well as many TFs with more restricted profiles. The detection of broadly expressed TFs in the cnidocytes, muscle cells, digestive filaments, epidermis, and neurons suggests the existence of tissue-specific hierarchy in *Nematostella* gene regulation in addition to the finer-grained cell-type-specific transcriptional signatures. We used ISH to morphologically validate some of these observations of restricted, cell-type-specific TF expression in adult animals.

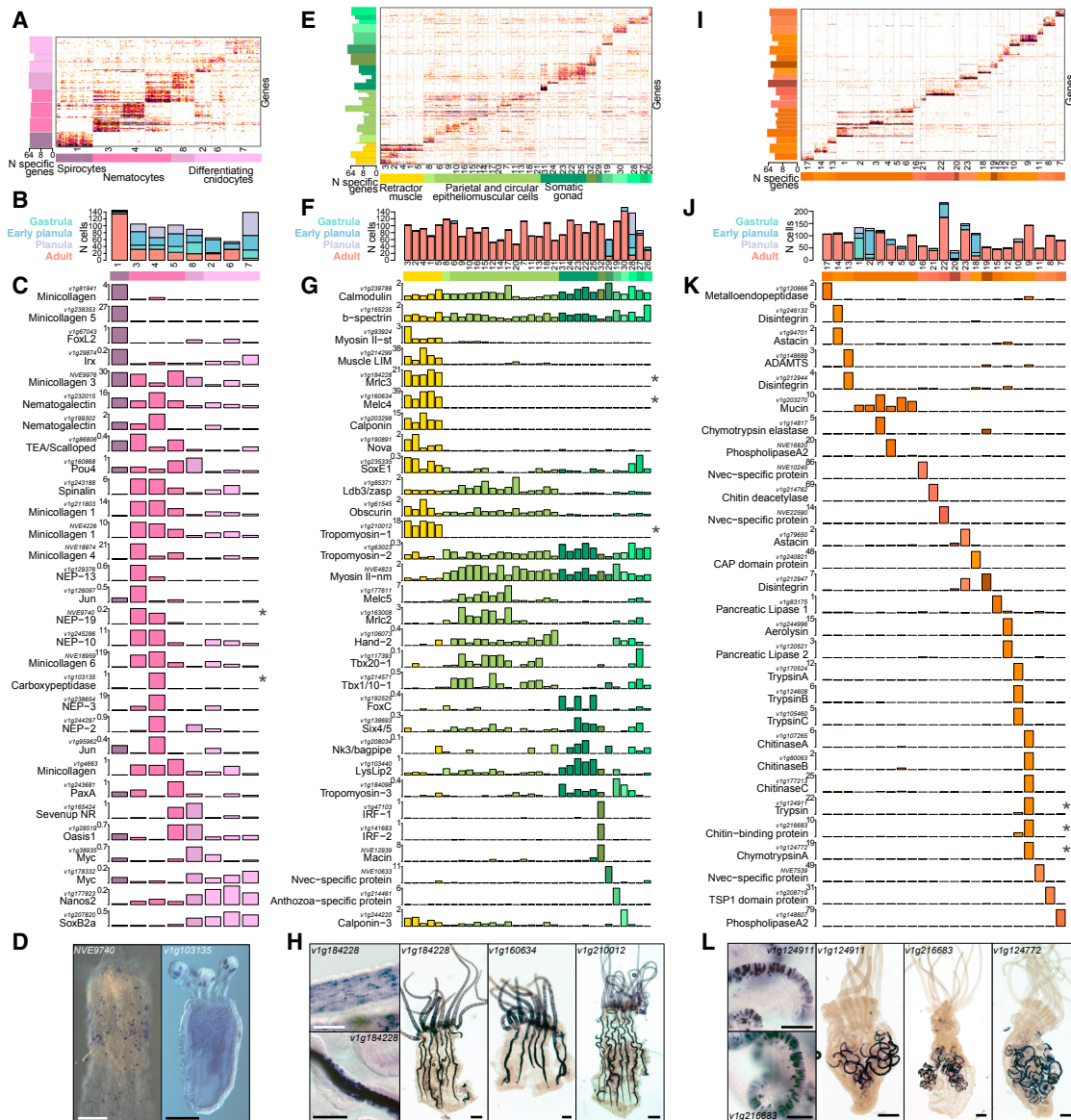


Figure 5. Characterization of Cnidocyte, Muscular, and Gland/Secretory Cell Types

(A) Expression of 178 genes (rows) across 796 cnidocyte cells sorted by metacell association.

(B) Number of cells and the distribution of stages per metacell.

(C) Expression of selected markers (*, genes in D).

(D) ISH for two genes associated with nematocyte metacells.

(E) Expression of 342 genes (rows) across 2,858 muscle/gastrodermis cells.

(F) Number of cells and the distribution of stages per metacell.

(G) Expression of selected markers (*, genes in H).

(H) ISH for three genes associated with tentacle and retractor muscle cells.

(I) Expression of 397 genes (rows) across 2,167 gland/secretory cells.

(J) Number of cells and the distribution of stages per metacell.

(K) Expression of selected markers (*, genes in L).

(L) ISH for three genes associated with digestive system gland cells.

All expression values are molecules/1,000 UMIs. Scale bars: 500 μ m (100 μ m, D and insets). See also Figure S4 and Tables S5, S6, and S7.

For example, the cnidocyte-specific *paxA* expression observed in the scRNA-seq data correlated with cnidocytes in the epidermis and tentacle tips (Babonis and Martindale, 2017) (Fig-

ure 6E). Similarly, *zfc2h2 fez* TF showed expression in a specific population of neurons (Figure 6F). Finally, the bilaterian gut marker *foxA* (Martindale et al., 2004) showed expression across

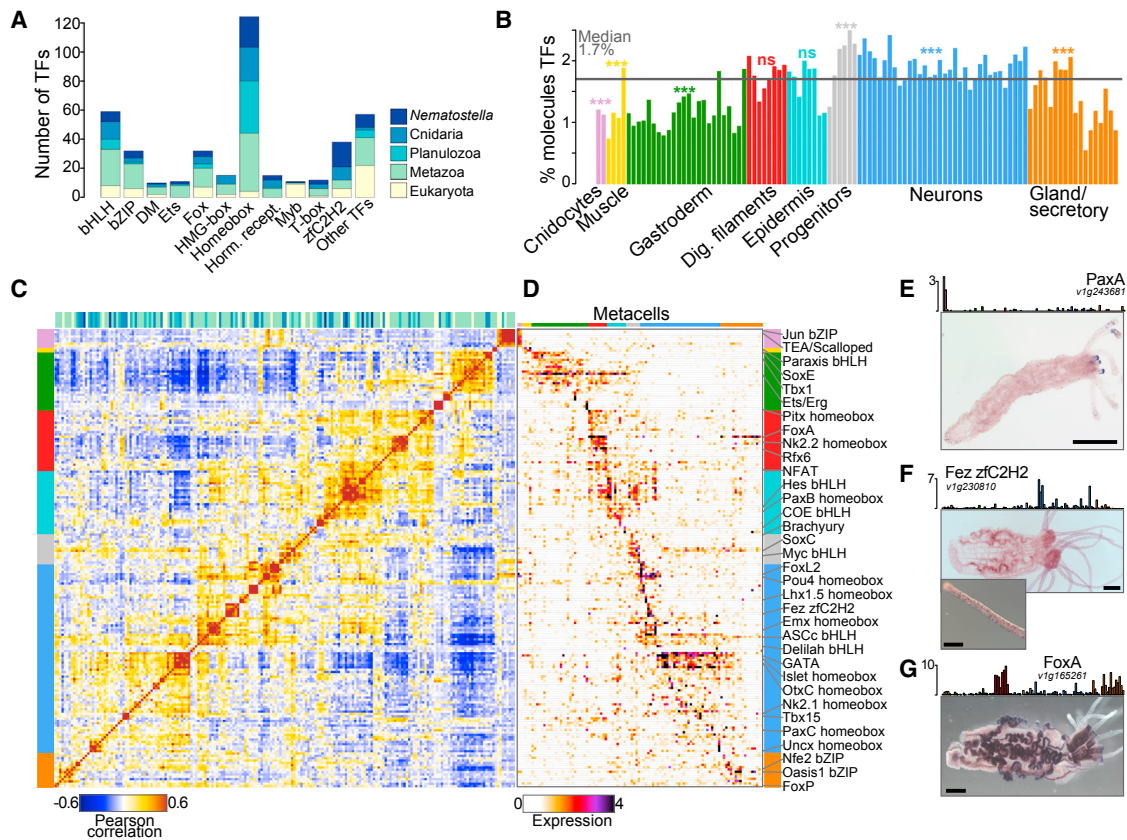


Figure 6. Transcription Factor Regulatory Programs in *Nematostella*

(A) Structural class and evolutionary age of 409 detected *Nematostella* TFs.

(B) Proportion of TF RNA molecules in each metacell (* $p < 0.0001$, chi-squared test).

(C) TF-TF correlation.

(D) TF expression across metacells.

(E) Expression and ISH of *Nematostella paxA* TF.

(F) Expression and ISH of *Nematostella zfC2H2 fez* TF.

(G) Expression and ISH of *Nematostella foxA* TF.

All expression values are molecules/10,000 UMIs. Scale bars: 500 μm (200 μm , Fez inset). See also Figure S5.

digestive filament and gland/secretory cell clusters (Figure 6G). Overall, we uncover a rich combinatorial map of TFs underlying specific cell type transcriptional identities in *Nematostella*.

Genomic Embedding of *Nematostella* Cell-Type Programs

TFs control downstream genes by binding to their regulatory elements, namely promoters and enhancers, through sequence-specific TF-DNA interactions. The combination of expressed TFs and associated regulatory elements defines the blueprint of the underlying cell-type-specific gene regulatory networks. To identify putative regulatory elements and link them to cell-type-specific genes—and since sequence specificities of *Nematostella* TFs are not characterized—we predicted TF affinities from sequence homology. This provided us with predicted binding motifs for 255 TFs (Weirauch et al., 2014) (Figure S6A). We then extracted the promoter sequence elements (−200/+50bp around annotated TSS) associated to the genes specifically upregulated in each of the 104 adult metacells introduced

above. We detected binding motifs enrichments (false discovery rate [FDR] < 0.02) in all of these gene sets (Figure 6A) and identified cases in which the expression profile of a TF was correlated with the enrichment profile of its putative binding motif, as exemplified by the ASCc neuronal profile, *foxA*-gland cells/digestive filaments profile, *erg* gastrodermal profile, and *pou4* neuronal/cnidocyte profile (Figures 7B and S6D–S6F). Analysis of shuffled controls confirmed the specificity of this result (Figure S6B).

We then sought to extend this analysis to enhancers, which play an essential role in tissue-specific gene expression in vertebrates and invertebrates (Javierre et al., 2016; Kvon et al., 2014). To examine if this is the case in *Nematostella*, we performed motif sequence enrichment analysis on 5,747 potential enhancer elements (Schwaiger et al., 2014), grouping them by proximity to the genes specifically upregulated in each of the metacells. This gave rise to another set of putative links between TFs and tissue-specific regulation (Figure 7C), as exemplified by *jun* and *gata* TFs associated with cnidocyte and neuronal cell clusters,

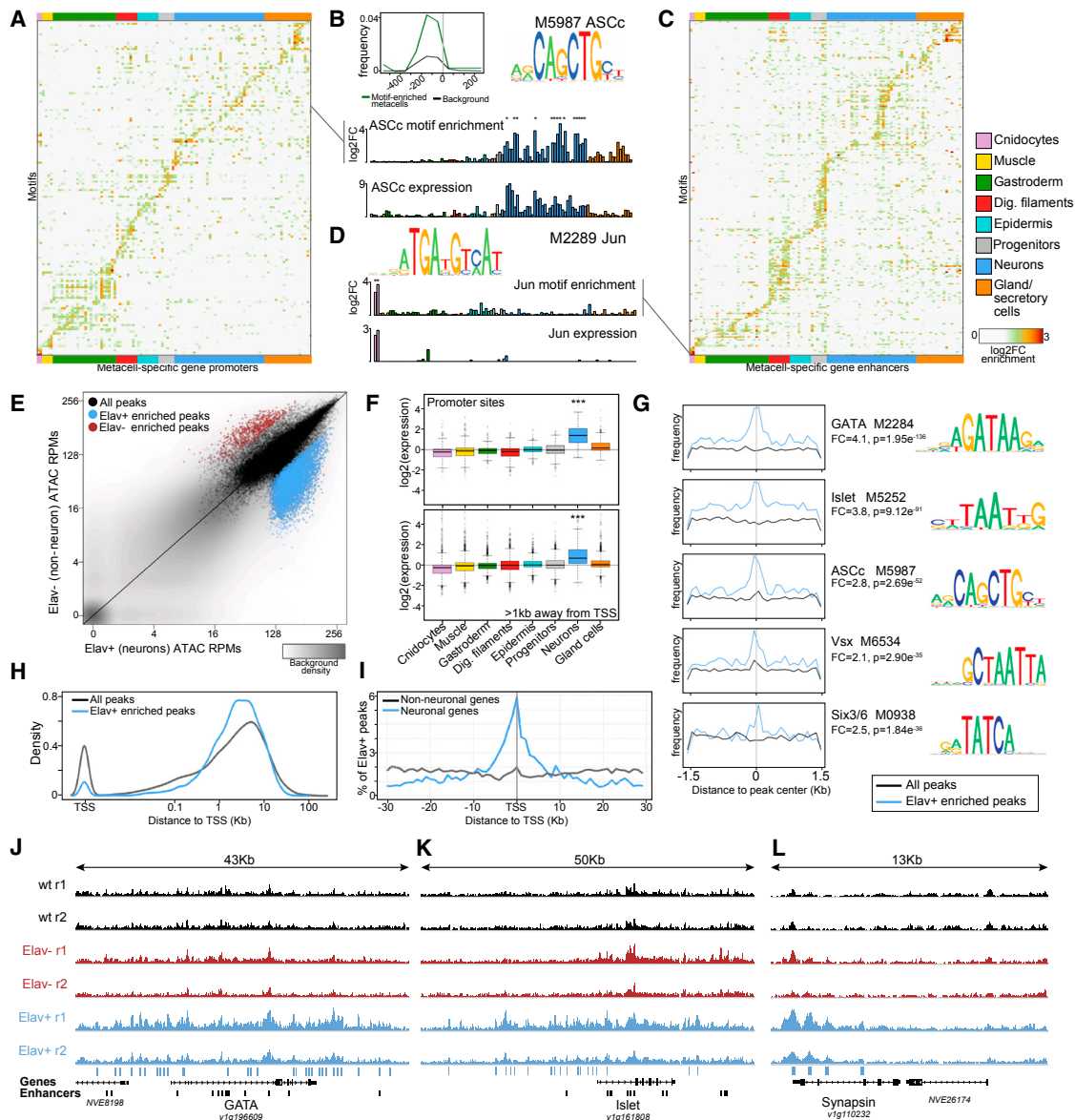


Figure 7. Genomic Regulation of *Nematostella* Cell-Type Programs

- (A) Motif enrichment in the promoters of metacell-specific gene sets.
 (B) ASCc TF motif frequency distribution around the TSS (top-left). Promoter ASCc motif enrichment (*FDR < 0.001) (middle). ASCc expression (bottom).
 (C) Motif enrichment in the enhancers associated with metacell-specific gene sets.
 (D) Same as (B) for *jun* TF in enhancers (*FDR < 0.001).
 (E) ATAC-seq signal in neuronal cells (*elav*⁺) versus non-neuronal cells (*elav*⁻) for genome-wide 140-bps bins (background density) and for defined regulatory sites (black dots).
 (F) Expression across broad cell types for genes associated with neuronal-enriched regulatory sites near the TSS (top) and >1 Kb away from the TSS (bottom).
 (G) Top five enriched motifs in neuronal-specific ATAC peaks. Line plots show the motif position distribution from the center of the peaks.
 (H) Distribution of ATAC peaks distance to the closest TSS.
 (I) Frequency of *elav*⁺ peaks around the TSS.
 (J) Chromatin accessibility landscape associated to the neuronal TF *gata*. Blue bars indicate significantly enriched neuronal peaks.
 (K and L) Same as (J) for *islet* TF (K) and *synapsin* (L).
 All expression values are molecules/1,000 UMIs. See also Figures S6 and S7.

respectively (Figures 7D and S6G). Importantly, this association was distinct from the preferences associated with the promoter sets (Figure S6C). Together, this analysis indicates that *Nema-*

tostella TF sequence specificities are sufficiently high to allow modeling and association of these specificities with putative regulatory mechanisms.

In order to gain further insights into the genomic encoding of *Nematostella* cell type programs, we profiled specific chromatin accessibility in neuronal cells. We first generated whole-adult *Nematostella* ATAC-seq to obtain a genome-wide reference of regulatory sites (Figures S7A–S7C). We found that the accessibility in gene promoters was proportional to the fraction of cells in which the gene is expressed, as expected for a whole-organism sampling (Figure S7D). We then used an *elav::mOrange* reporter line (described above) (Nakanishi et al., 2012) to enrich for neuronal cells (Figure S3) and interrogate their genome accessibility patterns. Comparison of the normalized ATAC profiles in *elav+* (neurons) and *elav-* (non-neurons) allowed for robust detection of constitutive and neuron-specific accessible sites (Figures 7E and S7E). We identified 5,456 sites with neuronal-specific regulatory activity in *Nematostella*. Analysis of neuron-specific accessible TSSs indicated significant enrichment for neuronal gene expression (Figure 7F; $p < 0.0001$, Wilcoxon rank-sum). Interestingly, such enrichment was also observed when analyzing accessible sites that are distal to the TSS (>1 Kb, $p < 0.0001$, Wilcoxon rank-sum), indicating the identification of putative neuronal enhancers. Analysis of sequence motifs at loci with neuronal-specific accessibility showed strong enrichment for motifs linked to neuronal TFs (e.g., *gata*, *islet*, and ASCc) (Figure 7G), corroborating the statistical association of these motifs with cell-type-specific genes as shown above.

The genome-wide neuronal ATAC-seq profile allowed for unbiased analysis of the usage and distribution of cell-type-specific regulatory elements around neuronal genes. We observed a strong over-representation of distal sites among neuronal-specific regulatory regions (Figure 7H; $p < 0.0001$, Wilcoxon rank-sum). Moreover, when examining the distribution of neuronal-specific regulatory sites around neuronally expressed genes, we found enrichment of distal sites up to 10 kb from the TSS (Figure 7I). In specific instances, such as the neuronal TFs *gata*, *islet*, and *ets/erg*, these regulatory landscapes were even broader, and we found arrays of neuron-specific regulatory sites spanning more than 40 Kb around the TSS (Figures 7J, 7K, and S7F). In contrast to these neuronal TFs, important neuronal effector genes, such as *synapsin*, showed less complex landscapes (Figures 7L and S7G). Overall, these results indicate that cell-type-specific regulation in *Nematostella* involves a high frequency of distal regulatory elements and complex regulatory landscapes, particularly around key mediators of cell type specificity.

DISCUSSION

In this work, we used whole-organism scRNA-seq augmented by transgenic reporters, ISH, genome sequence analysis, and chromatin-profiling assays to decompose transcriptional programs at single-cell resolution in the cnidarian *Nematostella vectensis*. This map provides a comprehensive view of the diversity of cell types present in this animal and support it with detailed gene expression signatures. For example, in neurons, broad expression patterns of pan-neuronal genes, such as *elav* or ASCc, are divided into distinct neuronal subtypes expressing specific ion channels, neuropeptides, and TFs, as well as multiple cnidarian-specific proteins. The apparent hierarchical organization of *Nematostella* cell types suggests, but does

not necessarily imply, a corresponding ontogenetic hierarchy. Cell-lineage tracking and functional assays will be needed to determine if this organization reflects the progressive specification and differentiation of these cells.

We can associate the *Nematostella* cell type hierarchy with highly specific co-expression patterns of a large TF repertoire, confirming the potential for elaborated combinatorial control schemes that are encoded into hundreds of TFs in *Nematostella*. This finding emphasizes the functional relevance of the TF diversification that occurred concomitantly with the evolution of metazoan multicellularity (de Mendoza et al., 2013). In addition to characterizing cell-type-specific TF modules, we identify a corresponding motif lexicon encoded in promoters and enhancers associated with cell-type-specific genes. The detection of this rich pool of tissue-specific TFs and motif signatures in *Nematostella* suggests that highly organized tissue-specific gene expression programs are not an exclusive hallmark of bilaterian animals but may have already existed in the cnidarian-bilaterian ancestor. We further investigated the epigenomic interface for such transcriptional diversity and dissected neuronal-specific genomic regulation through targeted chromatin accessibility analysis. This led to the identification of rich gene regulatory landscapes and demonstrated that many cell-type-specific genes are associated with dozens of putative enhancers packed into regulatory domains spanning up to 40 kb. This indicates the existence of combinatorial distal regulation mediated by chromatin looping or other long-range control schemes in this non-bilaterian lineage, a feature that has been suggested to be an important novelty linked to the emergence of animal multicellularity (Sebé-Pedrós et al., 2016). However, it remains to be determined if and to which extent cnidarians implement these cell identity programs in a similar way to vertebrates or other bilaterian models (Sexton et al., 2012), particularly given the lack of characterized cnidarian insulators (Schwaiger et al., 2014), which are key effectors of long-range gene regulation in bilaterian species (Cubefías-Potts et al., 2017).

The *Nematostella* single-cell atlas also allows us to initiate a systematic comparative study of cell types across species. The comparison of molecular profiles between *Nematostella*, *C. elegans*, and vertebrates reveals extensive conservation in neuronal effector gene repertoires. This conserved neuronal-specific transcription for many effector genes suggests that their neuronal association may have been a feature of a putative neuronal module in the cnidarian-bilaterian ancestor. This would support the existence in the cnidarian-bilaterian ancestor of a cell type with neuronal characteristics, including receptors that transduce an external signal into an electrical signal, fast propagation of this electrical signal through ion channels, and excitation-coupled secretion of neurotransmitters through a specialized synaptic scaffold (Liebeskind et al., 2017). In contrast to this general conservation of neuronal gene co-expression, comparative analysis of neuronal subtypes between *C. elegans*, mice, and *Nematostella* does not detect any cross-species similarities. More generally, we observe a lack of ortholog pairs co-expression; i.e., no pair of co-regulated genes within neuronal subtypes of one species have a conserved pair that is also co-regulated in another species. This result may be

explained if the core set of neuronal components was independently assembled into specific neuronal subtypes in different lineages, and this parallel diversification was accompanied by the emergence of new neuronal components—for example unique regulatory peptides and new ion channels and GPCR receptor paralog families. What we cannot infer from our current comparisons is whether the cnidarian-bilaterian ancestor harbored a highly diverse neuronal-type repertoire, which would have been extensively remodeled in the diverging cnidarian and bilaterian lineages, or, in contrast, if this ancestor had a very limited set of neuronal types. Systematic characterization of other cnidarian species, of early-branching bilaterians such as *Xenoturbella* and acoels, and of Cnidaria-Bilateria outgroup species can shed light into this question.

As exemplified here, we envision that applications of whole-organism scRNA-seq can rapidly allow the development of new insights into the evolution of metazoan cell type and tissue-specific genome regulation. On the one hand, the methodology can be applied to poorly characterized species and genomes, opening the way to massive expansion in the phylogenetic coverage and overall quality of models for genome regulation. On the other hand, and in a highly complementary way, comprehensive maps of cell type repertoires can be used as the basis for detailed mechanistic exploration of the many possible ways by which evolution solved the multicellularity challenge.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **CONTACT FOR REAGENT AND RESOURCE SHARING**
- **EXPERIMENTAL MODEL AND SUBJECT DETAILS**
 - *Nematostella* culture and cell sorting
- **METHOD DETAILS**
 - Single-cell RNA-seq
 - ATAC-seq
 - Whole-mount In Situ Hybridization and Imaging
 - Expression of transgenic reporter constructs
 - Staining of nematocyte capsules, tubulin and actin
- **QUANTIFICATION AND STATISTICAL ANALYSIS**
 - scRNA-seq analysis
 - Gene functional annotation
 - Phylogenetic distribution estimation
 - Cross-species transcriptome comparison
 - Motif analysis
 - ATAC-seq analysis
- **DATA AND SOFTWARE AVAILABILITY**

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at <https://doi.org/10.1016/j.cell.2018.05.019>.

ACKNOWLEDGMENTS

We thank Yehu Moran for help in the interpretation of the cell clusters, David Lara-Astiaso for critical comments on the manuscript, Manuel Irimia for dis-

cussion on cross-species clustering, and all the members of the Tanay and Spitz labs for comments and discussion. We thank Dr. Malte Paulsen for assistance with initial cell-sorting trials. We thank Dr. Fabian Rentzsch for providing the Elav and SoxB2a reporter lines and Dr. Detlev Arendt for allowing the use of the Fez, Lwamide2, and RPamide clones and riboprobes, as well as for constructive discussions. We thank Elodie Brient-Litzler for tremendous support developing the Pasteur Single Cell Initiative. A.S.-P. was supported by an EMBO long-term fellowship (ALTF 841-2014). Research in the Unit Epigenomics of Animal Development (GEAD) group was supported by the Region Ile de France (program SESAME 2016 “Paris Single Cell Centre”) and Pasteur Citech (“Single cell genomics”). Research in the A.T. group was supported by the European Research Council (724824). A.T. is a Kimmel investigator.

AUTHOR CONTRIBUTIONS

A.S.-P., F.S., A.T., and H.M. conceived the project. A.S.-P., H.M., B.S., F.P., E.C., S.S., S.N., and Z.M. performed cell sorting and single-cell RNA sequencing experiments. F.P. performed *in vivo* reporter experiments. M.-P.M. and J.R. performed ISH experiments. H.M. and B.S. performed ATAC-seq experiments. A.S.-P., A.L., and Y.L.-M. processed the raw data and performed preliminary analyses. A.S.-P. and A.T. conducted the data analysis. A.S.-P., F.S., P.R.H.S., A.T., and H.M. wrote the paper. All authors read and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 11, 2017

Revised: March 22, 2018

Accepted: May 9, 2018

Published: May 31, 2018

REFERENCES

- Babonis, L.S., and Martindale, M.Q. (2017). *PaxA*, but not *PaxC*, is required for cnidocyte development in the sea anemone *Nematostella vectensis*. *EvoDevo* 8, 14.
- Barr, M.M., and Sternberg, P.W. (1999). A polycystic kidney-disease gene homologue required for male mating behaviour in *C. elegans*. *Nature* 401, 386–389.
- Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature* 478, 343–348.
- Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218.
- Busengdal, H., and Rentzsch, F. (2017). Unipotent progenitors contribute to the generation of sensory cell types in the nervous system of the cnidarian *Nematostella vectensis*. *Dev. Biol.* 437, 59–68.
- Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661–667.
- Cubeñas-Potts, C., Rowley, M.J., Lyu, X., Li, G., Lei, E.P., and Corces, V.G. (2017). Different enhancer classes in *Drosophila* bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Res.* 45, 1714–1730.
- Davidson, E.H., and Erwin, D.H. (2006). Gene regulatory networks and the evolution of animal body plans. *Science* 311, 796–800.
- de Mendoza, A., Sebé-Pedrós, A., Šestak, M.S., Matejčić, M., Torruella, G., Domazet-Lošo, T., and Ruiz-Trillo, I. (2013). Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. *Proc. Natl. Acad. Sci. USA* 110, E4858–E4866.

- Frank, P., and Bleakney, J.S. (1976). Histology and sexual reproduction of the anemone *Nematostella vectensis* Stephenson 1935. *J. Nat. Hist.* *10*, 441–449.
- Fritzenwanker, J.H., and Technau, U. (2002). Induction of gametogenesis in the basal cnidarian *Nematostella vectensis* (Anthozoa). *Dev. Genes Evol.* *212*, 99–103.
- Hartl, M., Mitterstiller, A.-M., Valovka, T., Breuker, K., Hobmayer, B., and Bister, K. (2010). Stem cell-specific activation of an ancestral myc protooncogene with conserved basic functions in the early metazoan *Hydra*. *Proc. Natl. Acad. Sci. USA* *107*, 4051–4056.
- Havrilak, J.A., Faltine-Gonzalez, D., Wen, Y., Fodera, D., Simpson, A.C., Magie, C.R., and Layden, M.J. (2017). Characterization of NvLWamide-like neurons reveals stereotypy in *Nematostella* nerve net development. *Dev. Biol.* *431*, 336–346.
- Jahnel, S.M., Walzl, M., and Technau, U. (2014). Development and epithelial organisation of muscle cells in the sea anemone *Nematostella vectensis*. *Front. Zool.* *11*, 44.
- Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* *343*, 776–779.
- Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., et al.; BLUEPRINT Consortium (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* *167*, 1369–1384.
- Kvon, E.Z., Kazmar, T., Stampfel, G., Yáñez-Cuna, J.O., Pagani, M., Schernhuber, K., Dickson, B.J., and Stark, A. (2014). Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo. *Nature* *512*, 91–95.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* *9*, 357–359.
- Layden, M.J., Boekhout, M., and Martindale, M.Q. (2012). *Nematostella vectensis* achaete-scute homolog NvashA regulates embryonic ectodermal neurogenesis and represents an ancient component of the metazoan neural specification pathway. *Development* *139*, 1013–1022.
- Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* *13*, 2178–2189.
- Liebeskind, B.J., Hofmann, H.A., Hillis, D.M., and Zakon, H.H. (2017). Evolution of Animal Neural Systems. *Annu. Rev. Ecol. Syst.* *48*, 377–398.
- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* *15*, 509.
- Marlow, H.Q., Srivastava, M., Matus, D.Q., Rokhsar, D., and Martindale, M.Q. (2009). Anatomy and development of the nervous system of *Nematostella vectensis*, an anthozoan cnidarian. *Dev. Neurobiol.* *69*, 235–254.
- Marlow, H., Tosches, M.A., Tomer, R., Steinmetz, P.R., Lauri, A., Larsson, T., and Arendt, D. (2014). Larval body patterning and apical organs are conserved in animal evolution. *BMC Biol.* *12*, 7.
- Martindale, M.Q., Pang, K., and Finnerty, J.R. (2004). Investigating the origins of triploblasty: ‘mesodermal’ gene expression in a diploblastic animal, the sea anemone *Nematostella vectensis* (phylum, Cnidaria; class, Anthozoa). *Development* *131*, 2463–2474.
- Matus, D.Q., Pang, K., Daly, M., and Martindale, M.Q. (2007). Expression of Pax gene family members in the anthozoan cnidarian, *Nematostella vectensis*. *Evol. Dev.* *9*, 25–38.
- Mazza, M.E., Pang, K., Reitzel, A.M., Martindale, M.Q., and Finnerty, J.R. (2010). A conserved cluster of three PRD-class homeobox genes (homeobrain, rx and orthopedia) in the Cnidaria and Protostomia. *Evodevo* *1*, 3.
- Moran, Y., Praher, D., Schlesinger, A., Ayalon, A., Tal, Y., and Technau, U. (2013). Analysis of Soluble Protein Contents from the Nematocysts of a Model Sea Anemone Sheds Light on Venom Evolution. *Mar. Biotechnol.* *15*, 329–339.
- Nakanishi, N., Renfer, E., Technau, U., and Rentzsch, F. (2012). Nervous systems of the sea anemone *Nematostella vectensis* are generated by ectoderm and endoderm and shaped by distinct mechanisms. *Development* *139*, 347–357.
- Nielsen, C. (2005). Larval and adult brains. *Evol. Dev.* *7*, 483–489.
- Nowotschin, S., and Hadjantonakis, A.-K. (2009). Use of KikGR a photoconvertible green-to-red fluorescent protein for cell labeling and lineage analysis in ES cells and mouse embryos. *BMC Dev. Biol.* *9*, 49.
- Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V., et al. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* *317*, 86–94.
- Renfer, E., Amon-Hassenzahl, A., Steinmetz, P.R.H., and Technau, U. (2010). A muscle-specific transgenic reporter line of the sea anemone, *Nematostella vectensis*. *Proc. Natl. Acad. Sci. USA* *107*, 104–108.
- Rentzsch, F., and Technau, U. (2016). Genomics and development of *Nematostella vectensis* and other anthozoans. *Curr. Opin. Genet. Dev.* *39*, 63–70.
- Richards, G.S., and Rentzsch, F. (2014). Transgenic analysis of a SoxB gene reveals neural progenitor cells in the cnidarian *Nematostella vectensis*. *Development* *141*, 4681–4689.
- Schwaiger, M., Schönauer, A., Rendeiro, A.F., Pribitzer, C., Schauer, A., Gilles, A.F., Schinko, J.B., Renfer, E., Fredman, D., and Technau, U. (2014). Evolutionary conservation of the eumetazoan gene regulatory landscape. *Genome Res.* *24*, 639–650.
- Sebé-Pedrós, A., Ballaré, C., Parra-Acero, H., Chiva, C., Tena, J.J., Sabidó, E., Gómez-Skarmeta, J.L., Di Croce, L., and Ruiz-Trillo, I. (2016). The Dynamic Regulatory Genome of *Capsaspora* and the Origin of Animal Multicellularity. *Cell* *165*, 1224–1237.
- Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* *148*, 458–472.
- Sinaglia, C., Busengdal, H., Lerner, A., Oliveri, P., and Rentzsch, F. (2015). Molecular characterization of the apical organ of the anthozoan *Nematostella vectensis*. *Dev. Biol.* *398*, 120–133.
- Steinmetz, P.R.H., Kraus, J.E.M., Larroux, C., Hammel, J.U., Amon-Hassenzahl, A., Houliston, E., Wörheide, G., Nickel, M., Degnan, B.M., and Technau, U. (2012). Independent evolution of striated muscles in cnidarians and bilaterians. *Nature* *487*, 231–234.
- Steinmetz, P.R.H., Aman, A., Kraus, J.E.M., and Technau, U. (2017). Gut-like ectodermal tissue in a sea anemone challenges germ layer homology. *Nat. Ecol. Evol.* *1*, 1535–1542.
- Tasic, B., Menon, V., Nguyen, T.N., Kim, T.K., Jarsky, T., Yao, Z., Levi, B., Gray, L.T., Sorensen, S.A., Dolbeare, T., et al. (2016). Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* *19*, 335–346.
- Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* *158*, 1431–1443.
- Wolenski, F.S., Layden, M.J., Martindale, M.Q., Gilmore, T.D., and Finnerty, J.R. (2013). Characterizing the spatiotemporal expression of RNAs and proteins in the starlet sea anemone, *Nematostella vectensis*. *Nat. Protoc.* *8*, 900–915.
- Zenkert, C., Takahashi, T., Diesner, M.-O., and Özbek, S. (2011). Morphological and molecular analysis of the *Nematostella vectensis* cnidom. *PLoS ONE* *6*, e22725.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse Anti-acetylated tubulin antibody	Sigma	Cat#T6793
Alexa Fluor 488 goat anti-mouse secondary antibody	ThermoFisher Scientific	Cat#10656163
Biological Samples		
<i>Nematostella</i> adult, juvenile, and larval specimens	In-home culture.	N/A
Critical Commercial Assays		
Calcein Violet AM	ThermoFisher Scientific	Cat#C34858
Sytox Red	ThermoFisher Scientific	Cat#S34859
Nextera DNA Library Preparation Kit	Illumina	Cat#FC-121-1030
CalceinAM	ThermoFisher Scientific	Cat#C3100MP
Propidium Iodide	ThermoFisher Scientific	Cat#P3566
LiberaseTM	Roche	Cat#05401119001
Patent Blue VF	Sigma	Cat#198218
Deposited Data		
Raw and analyzed data	This paper.	GSE95723
Experimental Models: Organisms/Strains		
<i>Nematostella vectensis</i> wild type	In-home culture	N/A
<i>Nematostella NvElav1::mOrange</i> stable transgenic line	Nakanishi et al., 2012	N/A
<i>Nematostella SoxB2(2)::mOrange</i> stable transgenic line	Richards and Rentzsch, 2014	N/A
Oligonucleotides		
MARS-seq barcoded primers for mRNA capture	Jaitin et al., 2014	N/A
Primers for ISH probes (see STAR Methods)	This paper.	N/A
Primers to clone gene promoters for reporter plasmid construction (see STAR Methods)	This paper.	N/A
Recombinant DNA		
Plasmid pCAG:KikGR	Addgene	Cat# 32608
Plasmid pNvT-MHC::mCh	Addgene	Cat# 67943
Plasmid Ep2a::KikGR reporter	This paper.	N/A
Plasmid myc_v1g38935::KikGR reporter	This paper.	N/A
Plasmid OtxC::KikGR reporter	This paper.	N/A
Plasmid FoxL2::KikGR reporter	This paper.	N/A
Software and Algorithms		
Bowtie2	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
OrthoMCL	Li et al., 2003	http://orthomcl.org/common/downloads/software/
DESeq2	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
MetaCell	This paper.	http://compgenomics.weizmann.ac.il/tanay/?page_id=724
Other		
CisBP database	Weirauch et al., 2014	http://cisbp.cibr.utoronto.ca/
<i>Nematostella</i> JGI genome	Putnam et al., 2007	https://genome.jgi.doe.gov/Nemve1/Nemve1.home.html

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Amos Tanay (amos.tanay@weizmann.ac.il).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Nematostella culture and cell sorting

Nematostella vectensis polyps were spawned and reared as previously described to 11 days (“tentacle bud stage”), and five months (small adult polyps) (Fritzenwanker and Technau, 2002). For planula and gastrula stages, fertilized eggs were reared at 18.5°C until mid-gastrula, planula or late planula stages (respectively 2-days, 4-days and 7 days post-fertilization). As *Nematostella* development is not precisely synchronous, we staged each batch and sorted by hand all embryos to ensure the presence of developmental landmarks for each of the stages. For aboral ends dissections, two batches of 9-10 months old polyps were sequentially relaxed, their aboral ends dissected and dissociated into single cells. Polyps were relaxed and immobilized by adding a few drops of 7% MgCl₂ solution in their 3mL of 1/3X filtered seawater medium. Around 4mm of the most aboral region was cut perpendicularly to the main body axis, with no mesentery visible in the cut aboral end. 50 aboral ends were collected, washed in 1/3x Ca²⁺/Mg²⁺-free artificial seawater. Multiple polyps (~20), gastrula (~200) and planula (~200) from the same life stages per sample (adult or juvenile) were processed together, resulting in a unique single cell suspension for each life history stage from where cells were randomly sampled by FACS sorting, as described below. This process was repeated successively in order to ensure a short time between dissociation and cell capture (< 3h), even though cell viability was continuously monitored throughout the sorting process (using Calcein/PI staining, see below). The dissociation and sorting were done using the same reagents (enzymes, fluorescent sorting dyes, and media), in order to minimize technical factors and various batch effects. In order to dissociate the polyps, metamorphosing juvenile polyps of the “tentacle bud stage” or adult polyps were first placed in 1/3 strength Ca²⁺/Mg²⁺-free and EDTA-free artificial seawater (17mM TrisHCl, 165mM NaCl, 3.3mM KCl, 9mM NaHCO₃; final solution pH8). Immediately after that, multiple polyps were transferred to the above solution with LiberaseTM (Roche, #05401119001) at a concentration of 50 µg/ml. Dissociation was carried out at room temperature with occasional disruption using gelatine-coated pipettes for 20 minutes, and was subsequently stopped via addition of 1/10th volume 500mM EDTA solution. Cells were spun in a pre-chilled centrifuge at 500g in low binding eppendorf tubes and placed in fresh 1/3x Ca²⁺/Mg²⁺-free artificial seawater containing 2 µg/mL Calcein AM (Thermo, #C3100MP) and 1 µg/mL Propidium iodide (Thermo, #P3566). Cells were not washed out of the Calcein/PI staining solution. The same dissociation procedure was applied to dissected aboral ends of five-month polyps, to larval samples, and also to ELAV and SoxB2a reporter transgenic lines.

Cells were distributed into 384-wells capture plates (all coming from the same production batch) containing 2 µl of lysis solution using a FACSARIA III cell sorter. Lysis solution contains 0.2% Triton and RNase inhibitors plus barcoded poly(T) reverse-transcription (RT) primers for single cell RNA-seq. Live cells were selected by sorting only Calcein positive/PI negative cells, and doublet/multiplet exclusion was performed using FSC-W versus FSC-H. Additionally, in the case of ELAV and SoxB2a reporter transgenic lines, we selected mOrange-positive cells and, in this case, live/dead cells were stained using Calcein Violet AM (Thermo, #C34858) and Sytox Red (Thermo, #S34859). Fresh cell dissociates were prepared every 3h and sorted plates were immediately spun down, to ensure cell immersion into the lysis solution, and frozen at –80°C until further processing. Eight empty wells were kept in each plate as a control for data analysis.

METHOD DETAILS

Single-cell RNA-seq

Single cell libraries were prepared using the MARS-seq protocol, as previously described (Jaitin et al., 2014). All 17,664 single cell libraries (starting from 46 384-wells MARS-seq capture plates) were prepared with the same conditions (incubation times, temperatures, etc) and reagents. First, using a Bravo automated liquid handling platform (Agilent), mRNA was converted into cDNA with an oligonucleotide containing both the unique molecule identifiers (UMIs) and cell barcodes. Unused oligonucleotides were removed by Exonuclease I treatment. cDNAs were pooled (each pool representing half of the original 46 384-wells MARS-seq plate, 92 batches in total) 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 ssDNA:RNA ligase. Finally, RNA was reverse transcribed into DNA and PCR amplified. Resulting libraries were tested for amplification using qPCR and the size distribution and concentration were calculated using TapeStation (Agilent) and Qubit (Invitrogen). scRNA-seq libraries were pooled at equimolar concentration and sequenced using Illumina NextSeq 500 sequencer, in five sequencing runs, using high-output 75 cycles v2 kits (Illumina). We obtained 1,307M reads, of which 84% passed filtering, resulting in 1,098M reads with an average depth of 38,000 reads per cell (Table S1) and 6 reads/UMI on average.

ATAC-seq

Adult *NvElav1::mOrange* polyps were dissociated as described above and sorted on a BD FACS Aria using fluorescence profiles to discriminate mOrange+ (enriched for neurons) and mOrange- (depleted for neurons) populations. 200,000 cells from each population

were sorted for ATAC-seq. ATAC-seq library preparation was carried out according to (Buenrostro et al., 2013), with the following modifications: samples were spun at 2500g to recover cells and nuclei were spun at maximum speed for 2 minutes following cell lysis. For reference whole-organism ATAC experiments, wild-type young adult polyps were processed directly in lysis buffer (10mM Tris pH7.4, 10mM NaCl, 3mM MgCl₂, and 0.1% (v/v) IGEPAL CA-630), cells were first filtered across a 40µm mesh to remove debris and then spun at 2500g to pellet cells. 200,000 cells were used for wild-type ATAC experiments. A total of 14 cycles of PCR were performed to generate the libraries. Size selection was performed with SPRIselect beads (BeckmanCoulter, #B23317) to include fragments with a distribution from two hundred to seven hundred base pairs in length.

ATAC libraries were pooled at equimolar concentration and sequenced using an Illumina NextSeq 500 sequencer, in three sequencing runs, using high-output 75 cycles v2 kits (Illumina). We sequenced two replicates of each condition (wild-type, ELAV+, ELAV-) (Figure S7A), obtaining 404M reads and a total of 194M mapped and de-duplicated reads.

Whole-mount In Situ Hybridization and Imaging

Fixation of tentacle bud and adult stage *Nematostella* was carried out as previously described (Wolenski et al., 2013), with the following modifications: adult polyps were first micro-dissected by longitudinally opening the body wall along the oral-aboral axis using needles, were then fixed in 4% paraformaldehyde/0.2% glutaraldehyde for 90 s, followed by fixation at 4C overnight. *In situ* hybridization with digoxigenin-labeled riboprobes at a final concentration of 1 ng/µL was performed as previously described. Following *in situ* hybridization, polyps were sequentially cleared through a 40/60/70% glycerol series and imaged on a Zeiss Axiolmager DIC microscope or a Zeiss Discovery V16 microscope as single plane images or multi-plane stacks. Color balance, contrast and levels were adjusted across the image and images were cropped in Fiji and Adobe Photoshop. For multi-plane stacks, the Zeiss “extended focus” module was employed to generate a single image.

For *in situ* probes transcribed from PCR based templates, the primer sequences can be found in the table below. For *in situ* probes generated from templates produced via standard cloning and restriction digest, the fragment sequence is provided below. Finally, we used previously published *in situ* probes against *otp* TF (Mazza et al., 2010), *paxA* TF (Matus et al., 2007), *RFamide* and *elav* (Marlow et al., 2009).

Gene	Forward Primer	Reverse Primer
v1g233307	GCAATGCATGAAATATCTGCTC	GCGTAATACGACTCACTATAGGATAGAACGGAACAAGCACAAAGG
v1g160716	GAGGGTCTAATGGGAGATGTGA	GCGTAATACGACTCACTATAGGTTTGCATAGTTGCGATAACCTG
NVE17842	GCTTCAGCAGGAAAACAAGTG	GCGTAATACGACTCACTATAGGAAGAACTAAATTGGCACCAGACT
NVE3113	AAACAGCTTCAGGAAGACCTTG	GCGTAATACGACTCACTATAGGAGATTCTCTTGCTCGGCACTC
v1g241430	AGCGATCCGTTACAAGGTACAA	GCGTAATACGACTCACTATAGGTATGAAGGGAGCGCTTTTACTC
v1g184228	GCAAGAAAGGAAGCAAGAAGAA	GCGTAATACGACTCACTATAGGTAGGCAACGTTATGATCTCTGG
v1g160634	GTACCTGCCAAATCCATAGTCC	GCGTAATACGACTCACTATAGGTCGACACTCACAACCTCGGTAT
v1g210012	AGCGGAGCATCTTACAAAGGT	GCGTAATACGACTCACTATAGGATTCTCTTCTCTTCTCTCTGTC
v1g124911	CAGGCGATGTTACGGTATTACA	GCGTAATACGACTCACTATAGGATTATCGAGCCAGTCCCTTAAA
v1g216683	GTAAAGTCAATTCTCGCGCTTT	GCGTAATACGACTCACTATAGGTTACAGACGCAATGAGAAGCTG
v1g124772	ACGTGCATCAATGGTCAGAAC	GCGTAATACGACTCACTATAGGGCATTGATTACCTGTGACTTGC

>v1g165261_foxA

```
GCCCTTATGCCCTGaATACAGAAAGCGACTGTGCGAAAAGCCAAATATTAAGTGCACGCTCTTTTCATCTAATGGTCAGCTTCAACAGAAACAAGTTT
GCGAAACAATATATATGTGTATATAGGAATCTTTACAACATTTTTCTTAGCTAATAATCACTTGTGTCTGTTAATTCAAATTTCAAATATTTTGATTTGCGGA
AGATTTGCTTTCTTCAAATACTCTAATCATTTTTGTTTTTCTAGCCTAACATTTTCATCTTCTATTTTACATTTTGGATTGTTTCATAAAATTTAATTTT
CTAGCTCTTTACGAAATCTAGTTGTGCTTGTAGTAAAAAATGATCAATATAATCATACGAAAGCTCAAGCACATTGTACATAGACGAGATATTCGTCG
GATAGCTCAATTTTCAAGTGCCTCAATGTTTATATACAGGTTTATCATAATATTCAGCTTTTCTAGAAATAATTTCTACGAGTTCTAGTGTCTAAGTG
CTGCGCTAAGTTCACGAAAGATTTGATATACCACAACCTCGACGGCGTGAAGACGCAACCCTGGTAGTACGGAAGTTCGATCCGTCGTGATAGGTTGC
GATTCGATAGGCGGAATTAGGGAGTCTAGGGAACCCATGGATTGGAGTGATGGATGATAAGGACTGAAGTGCATGGGGTCGTAGCCTCGAAGCTCA
GCTTCGTGATCTTGCAGGATGATATTCTTGATAGCGAATGGGTGATTAACGACTTATTCATGGGCATACCTGCCATACCCATGGCGGTGATGCC
CCAACGtGTCCAtGCCATAgCCGtCCcgnacngAnACgGTGCatGGAAGcTAgGGGTTCCATTGAgCgAggnTgnGCCGcAtGCTCtTCgCCAtGCTT
tnnctGtGAccGGGTTGtGt
```

>v1g230810_zfC2H2_fez

GCCCTTGTTGTGTGTGTGACATGTGGAATGTGAGATTGTACACTTGATGGAACGCCTTCCGCAGATATGACACTTGAAAGGTTTCTCGCCtgTATG
 CGTGAGACGATGGTTTTATAAATCCCTTTTTGGTAAAACCTTCCCACAAATGTACACTGGAACGGCTTATAGCCTGCGTGAATTCGAATGTGA
 GTGTTTCAGAGTCGAGGAACGGTTGAACGCCTCCCACAAATGTTGCATTTATGTGGCTTCTCTTTGGTGTGAATGATCTTGTGCCGCATAGTGTG
 CTAGCTTGCCGGAATCCCTCCGCATATCTTGCCTTAAATGGGCGGGCTCTGTATGTACAGGCATGTGACGCGTCAGGTTATAGTGTGCGTTAA
 ACACCTTTCCGCATACTTCGCATGTGAAAGTTTTCGGCGTCCGTTGCGTCTTCTGCCAGTGTGTTGTTGAGCTAACTTGTGGAGTGCTCTCCGCT
 TTAGTACTGTCCAAATCTAGGAACGCCTTCCCAGGATGAAGATGCTTCTCGGAAAGCCGTCGTCACCTTTTGTGGCGATGGCACTAGCGAT
 TCGTCAAGGATTACAGGCGTACTCGATCTAGATGTCGGGAACATGGGCGAAGTGTAGTCGCGATCGCTGCTGTTGTAGCCATGATGATCTGGACTT
 AACCGGAAAGGATTGAAGGTAAAAGGGTCAAAGAAGTGCATTTCCGGGCTGAACTGTCTTTCCAGTTAGAGTCGCTGGGTGCGAGCGAGCTGAG
 GTTAAGCTTATCGGCTTGAAGATCTGTGCGGGTTCAATCTCCACCCGAATCGGCTCCCAAGCTCGGCGGAGTTGTCAACTCTGCTACTCCATCT
 TCACAAAAGACGAACAGCACGAATTTTTGGACTTGACAAGAAATGATCGTGGCATGTTGAGTTCTGTGCGAGCGGGTAGCTATCTGATAATCAC
 TGCCCCGGCTCAGGagaGCTTGTTCAGTTAGTTTAtaGGTGATGATGGTCCCTCCGTCcAGGATCAACTTgTGACAAGGGCCGaCGGTCTA
 CCAACTcCCCAgttaCtcTgcGTTg

>NVE9740_NEP19

ATTACGGCCGGGATTACACCATAAATAACAAGCCCTTACCTTTAATGAACTATACCGTCAATATTCTATTCACTTTGAAAATGTCCTCTATCC
 AGTTTTATTGTTGAAGAATCTTTGAAATGATTGAGCTGGTGCAGTTTGTCTGGTGGTGAATTCGCCTGTGGTAGCAATGACGTGGGCTCCAAAA
 TGGCCGAGGATTCGCCGATATCGAGACATCAGGGATTGATAAAGAGTTTGGGAGCGCCGGGAAAAGGAAGTTTCGACAGAAAAGCGGTGAAGG
 GAGCGACAACGAGGTACATCAAAGAAGTAGCTCCATGACCTCCATGAACCCTCTCATTGAGGATTACCGCACTCCCGATCCTTGTGTCCCAAACC
 CTTGCCTCAACGAAGGACGCTGTCTGAAGTCCCGGATGGAGAATACGACATCTATAGATGCGAGTGCCCAAAGGATATAACGGAAAAATATGT
 CAAGATGGCTTACCACCTGCTTCGAGTGAGAAAGTACGCGACAAGGGTCTGTGAGCCTAACCCATGCAAGAATCAAGGCACCTGTAAGAGA
 CTGGTGACCGAAGAAGTCCGATGAATGCGACTGCTTGCCTGAATGAAAAGGGACAACCTGCGAATATTACGTAATGATCCCTGCTTCCGAAT
 CCATGTTTGAATGTTGAAAATGCAAGCCCGGCACAATGAAATGAGACATT

Expression of transgenic reporter constructs

We generated a KikGR fluorescent reporter by modifying the previously published pNvT-MHC::mCh vector backbone (a gift from Ulrich Technau (Addgene plasmid #67943)) (Renfer et al., 2010). The mCherry fluorophore was replaced with the KikGR fluorophore amplified from the pCAG::KikGR vector (a gift from Anna-Katerina Hadjantonakis (Addgene plasmid #32608)) (Nowotschin and Hadjantonakis, 2009) with the following primers pair (6bp linker and restriction sites for *Ascl* and *SbfI* are indicated in bold):

KikGR-Fw 5'-**TACAGAGGCGCGCC**ATGAGTGTGATTACATCAGA-3'

KikGR-Rev: 5'-**GACTGTCCTGCAGG**TACTTGGCCAGCCTTGGCA-3'

gDNA was isolated from a pool of adult polyps using the MagAttract HMW DNA Kit (QIAGEN), according to the manufacturer's instructions, after replacing the original 1/3X seawater medium with Phosphate Buffered Saline solution. gDNA concentration was quantified using the Qubit High Sensitivity dsDNA quantification kit (Thermo). Promoter regions were amplified from 50ng of genomic DNA using the Phusion High Fidelity Polymerase Master Mix (Thermo) with the following primers pairs (6bp linker and restriction sites for *PacI* and *Ascl* are indicated in bold):

pEp_v1g242253-Fw: 5'-**ATCTGATTAATTAAT**TGTAGTGATCGTGAGGTGC-3'

pEp_v1g242253-Rev: 5'-**ATCTGAGGCGCGCC**GACGGGAGAGGTGCTTATTC-3'

pOtxC_v1g213735-Fwd 5'-**ATCTGATTAATTA**ACTTGTGCTCTACTTCGGTTG-3'

pOtxC_v1g213735-Rev 5'-**ATCTGAGGCGCGCC**CCGGAGTGTCTCGTTCTT-3'

pFoxL2_v1g67043-Fw 5'-**ATCTGATTAATTA**AGCCCATACGCGATTTCTTTCAAC-3'

pFoxL2_v1g67043-Rev 5'-**ATCTGAGGCGCGCC**CGCGCTTGTGTGTGACACA-3'

pMyc-v1g38935-Fw 5'- **ATCTGATTAATTA**ACCGGTATGAAGTTCGAGTGAA-3'

pMyc-v1g38935-Rev 5'-**ATCTGAGGCGCGCC**GTACAACACAACCTGAAAGATT-3'

For *myc v1g38935*, we used the primers with an added *PacI* or *Ascl* sites as above, but, due to the presence of an *Ascl* restriction site in the promoter sequence, we digested the fragment with *PacI* and *BstBI* as restriction enzymes, and we used an annealed linker to add back the missing 29bp (annealing sequences are indicated in bold).

Fw linker-pMyc-v1g38935 5'-CGAATCTTT**CAGGTTGTGTTGTAACGG**-3',

Rev linker *BstBI* Myc-v1g38935 5'-CGCG**CCGTTACAACACAACCTGAAAGATT**-3'

The *MHC* promoter of the pNvT-MHC::mCherry was replaced with 1.6kb, 2.8kb, 4.6kb and 2.5kb upstream of the ATG of the *ep2a* (v1g242253), *myc* (v1g38935), *otxC* (v1g213735) and *foxL2* (v1g67043), respectively, to generate the four reporter constructs.

Plasmids were sequenced, mini-prepped, and cleaned with magnetic beads by adding 1.8v/v of a magnetic beads solution (100mM TrisHCl pH8, 1mM EDTA, 1M NaCl, 180mg/mL PEG-8000, 1mg/mL Sera-Mag SpeedBeads Carboxyl Magnetic Beads

(GE Healthcare, #65152105050250)) washing them twice with 1mL 80% ethanol and resuspending them in nuclease-free water. Plasmid concentration was quantified using the Qubit High Sensitivity dsDNA quantification kit.

Embryos were obtained as previously described (Fritzenwanker and Technau, 2002), and prepared and injected according to (Renfer et al., 2010) with the following modification: 50ng/ μ L final vector, 1X I-SceI Buffer (10mM TrisHCl, 10mM MgCl₂, 1mM DTT, final solution pH8.8), 0.5mM Patent Blue VF (Sigma, #198218), 0.2U/ μ L I-SceI (New England Biolabs, #R0694S) and incubated for 30' at 37°C, then fresh I-SceI (I-SceI aliquots were kept on dry ice and stored at -80°C) was added to a final total concentration 0.4U/ μ L. This injection mix was microinjected into fertilized and dejellied embryos at 16-19°C (to delay cleavage) using an Eppendorf Femtojet microinjector. Microinjection needles were pulled using a Sutter P-1000 Brown micropipette puller. Injected embryos were kept in 6-wells plates and changed daily into fresh 1/2X filtered seawater for the first two days post-injection then every 2 days after and kept at 18°C.

Staining of nematocyte capsules, tubulin and actin

Polyp fixation and staining were done as previously reported (Marlow et al., 2009). Briefly, polyps were fixed in 4% paraformaldehyde and 0.2% glutaraldehyde for 90 s, followed by 4% paraformaldehyde for one hour at 4C, and then washed five times into PBS (18.6mM NaH₂PO₄·H₂O, 84.1mM Na₂HPO₄·2H₂O, 1.75M NaCl). Polyps were blocked for 30 min at room temperature and incubated overnight at 4C in mouse anti-acetylated tubulin primary antibody (Sigma, #T6793) and then washed out three times into PBS-triton. Polyps were subsequently incubated in Alexa Fluor 488 goat anti-mouse secondary antibody (Fisher Scientific, #10256302) and post-stained with Phalloidin (Fisher Scientific, #10656163). Polyps stained for nematocyst capsules were fixed and treated as previously described (Marlow et al., 2009). Basically, if calcium chelators are present during fixation, cnidocyte capsule walls can be stained with DAPI and visualized using a green filter under fluorescent illumination while nuclei (DAPI-stained DNA) are visible using a blue filter.

QUANTIFICATION AND STATISTICAL ANALYSIS

scRNA-seq analysis

Pre-processing and modeling cell-cell similarity

Reads were mapped into *Nematostella* genome v1.0 (<https://genome.jgi.doe.gov/Nemve1/Nemve1.home.html>) using Bowtie2 (Langmead and Salzberg, 2012) with default parameters and associated with a gene intervals defined by merging two existing *Nematostella* annotation (Putnam et al., 2007; Schwaiger et al., 2014). In the merging process, we gave priority to the JGI gene identifier when redundancy existed, but always kept the longest predicted form. After merging, we extended gene intervals up to 2kb downstream or until the next gene in frame is found. This accounts for the poor 3'UTR annotation of *Nematostella* genome, which causes most of the MARS-seq (a 3' biased RNA-seq method) reads to map outside genes (Figure S1B). Mapped reads were further processed and filtered as previously described (Jaitin et al., 2014). UMI filtering include two components, one eliminating spurious UMIs resulting from synthesis and sequencing errors, and the other eliminating artifacts involving unlikely IVT product distributions that are likely a consequence of second strand synthesis or IVT errors. The minimum FDR q-value required for filtering was 0.2. We filtered cells with less than 100 UMI from downstream analysis.

We summarize all UMIs in the molecule count matrix $U = [u_{gi}]$ on genes $g \in G$ and cells $i \in I$. Each cell is associated with a batch identifier, that we represent using a vector b_i over the cells. For simplicity, we use the following conventions on matrices: Given a matrix $A = [a_{ij}]$ we denote marginalization over rows as a_i and columns as a_j . For example, u_g is the total molecule count for gene g on the raw count matrix, and u_i is the total number of molecule for a cell.

The standard practice of selecting genes for modeling cell-to-cell similarity is to identify genes with high normalized variance (variance/mean on down-sampled matrices). This is problematic for our data, since down-sampling (or any form of UMI count normalization) in a single cell cohort with very broad range of UMI depths (100-10,000 in our case) decreases statistical power considerably. Instead, we can compute for each gene g the Pearson correlation with the cell depth $r_g^{sz} = cor(u_{gi}, u_i)$. As genes increase in their expression, their correlation with the cell depth will increase, even if their expression is completely homogeneous within the population, simply due to the decreasing sampling variance of their concentrations. On the other hand, truly variable genes will show a lower correlation with the cell depth compared to genes with similar expression average. We therefore compute an empirical trend $r(u_g)$ using the median r_g^{sz} correlation in a moving window of total gene expression (using windows of 100 genes). Then, we define the normalized depth scaling as $r'_g = r_g^{sz} - r(u_g)$. We select genes with sufficiently high u_g and $r'_g < T_{gr}$ (we used $T_{gr} = -0.05$).

Given a set of selected genes, we transform the raw UMI count U on the gene features F as $M = [m_{gi}] = [\log_2(\epsilon + u_{gi})]_{g \in F}$. We then compute the raw similarity matrix using Pearson correlations on the transformed features $R = [r(m_{gi}, m_{gj})]_{ij}$. Note that cells with different UMI counts may be a-priori more likely to generate high similarity scores than cells with low UMI counts, but that normalizing UMI depth, as discussed above, can be very problematic in our data. The procedure of balancing the similarity graph, as discussed below, helps controlling this effect.

The balanced K-nn cell similarity graph

Next, based on the raw similarity matrix R we perform a non-parametric transformation and compute $S = [s_{ij}] = [rank_j(r_{ij})]$. Here $rank$ is the ranking function, and each row represents the order of similarity between all cells j and a specific cell i . The S matrix is highly non-symmetric, in particular when the similarities going from an outlier cell are linking it to members of a large and homogeneous cell

group. In such cases, the outlier cell will not be among the most similar cells to any of its own neighbors. To mitigate such effects, we symmetrize S and balance the resulting matrix through the following steps:

$$[s_{ij}^1] = [\max(\alpha K^2 - s_{ij} * s_{ji}, 0)]$$

$$[s_{ij}^2] = [\max(\beta K - \text{rank}_j(s_{ij}^1), 0)]$$

$$[a_{ij}] = [\max(K - \text{rank}_j(s_{ij}^2), 0)]$$

Where K is the number of neighbors we aim to add to each cell (depending on the size of the dataset, this is in the order of 100), α (10 by default) and β (3 by default) are expansion parameters that allow more than K nearest neighbors for each node to be initially considered by the balancing process. A weighted directed graph G is constructed using $[a_{ij}]$ as the weighted adjacency matrix. The number of outgoing edges for each node in G is limited by K and the number of incoming edges is bounded by βK . Nodes with lower degrees are however still possible, since outlier cells may become disconnected or poorly connected during the balancing operations.

Seeding and optimizing a graph cover

We next wish to cover all cells with disjoint and dense subgraphs (or *metacells*) on sizes that are on a scale of the user-defined parameter K that determined the construction of the balanced K -nn graph. The K parameter reflects our downstream analysis goals – it should allow for a sufficient accuracy for estimating the UMI distribution within each metacell, but still provide sufficient flexibility to capture multiple sources of variation in the data. Metacells differ conceptually from clusters, since there is no attempt to ensure strong separation between them but only to maximize their coherence. A detailed account on the impact of capturing the similarity structure of the graph by a metacell cover will appear elsewhere.

To derive a cover of the graph G by a set of cohesive subgraphs, we first perform a seeding phase and define an initial set of metacells $I_1 \subset I, \dots, I_m \subset I$. We denote by $N(i)$ the set of graphic outgoing neighbors of i , and start by defining an empty assignment of cells to metacells $mc(i) = -1$. During seeding iterations, we define the set of covered nodes as $C = \{i \mid mc(i) > -1\}$ and the cover-free score for each node is defined as $f(i) = |N(i) - C|$. We sample seeds as follows:

While $\max_j f(j)/K > \text{size_min}$ do:

sample a new seed j by drawing a sample from cells in $I - C$ with weights $f(i)^3$

update $mc(u) = j$ for $u \in N(j) - C$

update cover set C and scores f

When we meet the stop criterion, cells that are not associated with a seed metacell (i.e., cells for which $mc(i) = -1$) have at most $K * \text{size_min}$ uncovered neighbors. Next, define the metacell groups $M_k = \{i \mid mc(i) = k\}$. The association between a cell and a metacell subgraph is based on edges to and from the cell, in a potentially non-symmetric fashion. The outgoing weight vector for each cell is defined as $w_{o_{ik}} = \sum_{j \in N(i) \cap M_k} a_{ij}$. We define $N^{in}(i)$ as the set of incoming neighbors for cell i , and the incoming weight vector is set to $w_{i_{ik}} = \sum_{j \in N^{in}(i) \cap M_k} a_{ji}$. We score metacell association by multiplying these two weights and normalizing by module size, setting $w_{ik} = w_{i_{ik}} w_{o_{ik}} / |M_k|^2$. We can now re-assign cells to metacells iteratively until convergence:

Until convergence:

Select a cell i

Reassign $mc(i) = \text{argmax}_m w_{im}$

Update weights

Since this heuristic is not guaranteeing convergence into locally optimal metacell assignment (i.e., where all cells are assigned to their maximum weight metacell), we employ a cooling strategy: We define a cooling profile $\text{cool}(c) = 1 + \max(0, c - c_{burn}) * \delta$, where by default $c_{burn} = 10$, and $\delta = 0.05$. We record the total number of metacell changes for each cell as $c(i)$, and modify the weights of the currently assigned metacell to $w_{i,mc(i)} = (w_{i,mc(i)})^{\text{cool}(c(i))}$. Using this approach convergence is guaranteed to occur after a limited number of iterations.

In the application reported here, we used $K = 250$ and $\text{size_min} = 0.2$ for the adult and larval datasets (Figures 1 and 2). For cell subset zoom-in analysis (Figures 4 and 5), we used $K = 150$ and $\text{size_min} = 0.33$.

Metacell summary statistics

A metacell cover $M_1, \dots, M_k, \dots, M_m$ can be studied as a set of meta-transcriptional states by pooling UMI counts:

$$u_{gk} = \sum_{\{i \in M_k\}} u_{gi}$$

$$u_k = \sum_g u_{gk}$$

More precisely, given general assumptions on transcriptional states as sampling from log-normal distributions, and in order to reduce the effect of outliers, we summarize metacells using log transformed statistics:

$$\rho_{gk} = \exp \left[\frac{1}{|M_k|} \sum_{\{i \in M_k\}} \log \left(\frac{(\rho + u_{gi})}{(\rho G + u_i)} \right) \right]$$

$$u'_{gk} = \rho_{gk} * u_k$$

Where the prior parameter was set to $\rho = 1/7$ in this work. To compare metacells' gene expression we define the log fold change enrichment score:

$$lfp_{gk} = \log_2(\rho_{gk} / \text{median}_k(\rho_{gk})).$$

Removing background noise using metacells

The UMI distribution represented by the matrix U is known to be affected by several sources of noise. A background or ambient noise model can be written as:

$$u_{gi} = (1 - \epsilon_g) u_{gi}^{real} + \epsilon_g \left(\frac{u_g^b}{N^b} \right)$$

where b is the batch of cell i , u_g^b is the total number of molecules observed for the gene in the batch, u_{gi}^{real} is number of molecules of gene g that were actually sampled from cell i , and ϵ_g is a gene-specific noise parameter. This model assumes that the molecules within each cell switch with probability ϵ_g their cellular identity uniformly, but only within their batch. Such process may represent amplification and sequencing errors of the single cell libraries where barcoded oligonucleotides are priming PCR reaction for molecules that were initially labeled by other cell barcodes.

Given observations U , it is difficult to infer the noise parameters ϵ and to distinguish noise from original molecules. We aggregate information within metacells to allow more robust estimation and filtering of noisy UMIs. We compute the mean per-cell UMI count for each gene in each batch as:

$$e_{gb} = \frac{u_g^b}{n^b}$$

where u_g^b is the total UMIs for gene g in batch b , and n^b is the number of cells in batch b . Our estimate for the expected background count of each gene in each metacell is then:

$$e_{gm} = \epsilon \sum_b e_{gb} \left(\sum_i b_{bi} m_{im} \right)$$

where ϵ is some initial guess on the noise level, $B = [b_{bi}]$ is the batch association matrix, set to 1 if cell i is in batch b , and $M = [m_{im}]$ is the metacell association matrix, defined similarly. The observed number of UMIs in a metacell is:

$$o_{gm} = \sum_i u_{gi} m_{im}$$

and the deviation of the expression from the background expectation can be quantified as (using element-wise arithmetic here):

$$z_{gm} = (o_{gm} - e_{gm}) / \sqrt{e_{gm}}$$

We note that in practice we modify the raw U matrix to control the effect of outliers, so that u_{gi}^{reg} is used instead of u_{gi} when computing o_{gm} :

$$u_{gi}^{reg} = \min \left(u_{ij}, \lceil e_{gb} \rceil + 3\sqrt{\lceil e_{gb} \rceil} \right)$$

Rounding up is performed in order to consider low UMI count genes conservatively. Note that a uniformly expressed (i.e., house-keeping) gene g is not expected to be expressed at ϵu_g levels in any of the metacells provided its overall expression intensity is not very low (in which case, no noise filtering should be attempted). Specifically, we define for each gene the set of metacells that are compatible with the noise model using the Boolean matrix:

$$back_{gm} = z_{gm} < T (T = 2 \text{ by default})$$

Then we test, for each gene, if the total number of cells with $back_{gm} = TRUE$ is at least α of the total number of cells, and if the total number of UMIs for the gene in all cells within such metacells is at most $2 * \epsilon * U_g$. If this is the case, we mark the gene as a valid candidate for filtering noise, generating a set G_{tofilt} . To perform UMI filtering for such genes in practice, we set $u_{gi} = 0$ for g in G_{tofilt} and cells contained in metacells with $back_{gm} = 1$. We are however not filtering UMIs from cells with outlier behavior relative to their metacell (i.e., those with $u_{gi}^{reg} < u_{gi}$).

The above procedure is sensitive to the initial parameters used. We use a minimal background subpopulation fraction $\alpha = 1/8$ and $\epsilon = 0.03$. We note that setting a higher initial epsilon is not affecting accuracy of filtering genes with high expression, but can result in too aggressive filtering in low expression noise.

To incorporate the above noise reduction procedure into the overall metacell derivation process, we used a two-phase approach. First, we grouped cells into metacells without filtering noisy UMIs. Then we identified and filtered noisy UMIs as described above. Finally, we applied the metacell cover algorithm again on the filtered UMI matrix.

Bootstrap and QC of metacells

Given a raw metacell cover, we performed downstream analysis including filtering and merging of metacells, facilitating in-depth analysis of gene expression within and between cell types as discussed in the text. We next describe the different quality metrics used for processing raw metacells, with details on how these were applied for the adult, larva, and zoom-in analyses provided in the following section.

Graph consistency. For each metacell m we computed the graph consistency:

$$Gconsist_m = \frac{\| (i,j) \|_{\{i \in M_m, j \in M_m, (i,j) \in E\}}}{\| (i,j) \|_{\{i \in M_m, (i,j) \in E\}}}$$

as the total number of edges linking two cells within the metacell divided by the total number of edges directed from a cell within it. Metacells with low overall graph consistency may represent low specificity transcriptional states, or groups of metacells forming larger cluster with arbitrary subdivision.

Co-clustering consistency. We assessed the robustness of metacells using a bootstrap approach. To this end, performed 1,000 iterations of resampling 75% of the cells and deriving metacell covers, generating matrices $O = [o_{ij}]$ and $C = [c_{ij}]$ that specify how many times the pair of cells i, j were sampled in a subset together and how many times they were both assigned to the same metacell, respectively. Given a metacell assignment matrix $M = [m_{im}]$, we can assess the consistency of assignment for a cell i using the probability of co-clustering i with m :

$$c_{im} = \sum_j \frac{c_{ij} m_{jm}}{o_{ij} m_{jm}}$$

here c_i is the total number of co-cluster observations for cell i . The overall co-clustering consistency score for a metacell m is then defined as:

$$Cconsist_m = \frac{1}{|M_m|} \sum_{\{i \in M_m\}} c_{im}$$

Given two metacells, we assess their potential merge (merge score) by counting the number of times cells within them were co-clustered together (using $C = [c_{ij}]$), normalized by the total number of times cells from these metacells were sampled together.

Expression consistency. For each metacell k we defined a signature gene module G_k by identifying genes with $lfp_{gk} > 1$ (using at most 50 genes with the highest enrichment). We compute the background distribution of total number of UMIs in this gene module across all cells (excluding the cells from the metacell k and its two most similar neighboring metacells). We then define the expression consistency score $Econsist_k$ as the fraction of cells within M_k with total G_k UMIs count higher than the top percentile of the background distribution.

Post-processing metacells

For the adult analysis discussed in [Figure 1](#), we eliminated 15 metacells (and 975 cells) for which with $Gconsist_m < 0.2$. We also filtered from our atlas 9 metacells (and 453 cells) for which no gene showed high enrichment (for all genes $lfp_{gk} < \log_2(9)$). This relatively aggressive approach ensured reported metacells represent highly specific transcriptional states. Downstream grouping and annotation of metacells into classes was done based on bootstrap values and analysis of gene annotation as described in the text.

In the analysis of the larval dataset ([Figure 2](#)), we iteratively merged metacells with merge score $> 10\%$. We then filtered 14 metacells (and 1,135 cells) with $Econsist_k < 0.2$ and a $Cconsist_m < 20\%$. Supervised downstream analysis and grouping of cells was performed using gene annotation, comparison to the adult data and the bootstrap co-clustering matrix.

In the analysis of specific cell types ([Figures 4 and 5](#)) we iteratively merged metacells with merge score $> 20\%$. We then filtered metacells with a $Econsist_k < 0.3$ and a $Cconsist_m < 30\%$, corresponding to 1 metacell in neurons ([Figure 4](#)), 0 metacells in cnidocytes, 9 metacells in gastrodermis/muscle, and 0 metacells in the gland/secretory cells ([Figure 5](#)) analyses.

Metacell regularized force-directed 2D projection

To derive a metacell 2D projection, we use the balanced similarity graph G and summarize the total number of (unweighted) edges linking metacells:

$$B = b_{ml} = \frac{K^2}{|M_m| * |M_l|} \sum_{\{i \in M_m, j \in M_l\}} [a_{ij} / N_{med}]$$

$N_{med} = \text{median}_i(|M_i|)$ is a scaling constant. To further simplify the linkage between metacells, we restrict the original graph G (built given some K -nn parameter K) to contain only the top K_{2D} edges for each cell (typically setting K_{2D} to be smaller than K)

We normalize B rows and columns:

$$b_m = \sum_l b_{ml} \quad b_{ml}^{out} = \frac{b_{ml}}{b_m}, \quad b_l = \sum_m b_{ml} \quad b_{ml}^{in} = \frac{b_{ml}}{b_l}$$

And define the score for connecting two metacells as $b_{ml}^i = b_{ml}^{in} + b_{ml}^{out}$. We retain as candidate edges only pairs for which $b_{ml}^i > T_{edge}$. We then construct a graph $G^M = (M, E^M)$ on metacells $M = (1, \dots, m)$, by adding the D_{2D} highest scoring candidate edges (if such edges exists) for each metacell. Note that any metacell in the graph can be completely disconnected if its cells are highly connected to themselves but not to any other metacell. We embed the metacell graph in 2D using a standard force-directed layout algorithm (using the graphviz implementation), deriving coordinates for each metacell (x_k, y_k) . We also position each cell i using average coordinates of the metacells containing its neighboring cells:

$$x_i = \frac{1}{Z} \sum_{\{j | a_{ij} > T_{2dcells}, \{mc(i), mc(j)\} \in E^M\}} x_{\{mc(j)\}}, \quad y_i = \frac{1}{Z} \sum_{\{j | a_{ij} > T_{2dcells}, \{mc(i), mc(j)\} \in E^M\}} y_{\{mc(j)\}}$$

Where W^{layout} is a parameter determining how many edges we will use to position a cell (0 will imply using all edges, $1 - K'/K$ will imply using the top K' neighbors). In practice, we add some Gaussian noise to both coordinates to reduce cell overlap.

Here, we used $K_{2D} = 80$, $D_{2D} = 8$, $T_{edge} = 0.02$, $T_{2dcells} = 1 - 30/250$ for the adult dataset (Figure 1) and $K_{2D} = 50$, $D_{2D} = 8$, $T_{edge} = 0.08$, $T_{2dcells} = 1 - 30/150$ for the larval dataset (Figure 2).

Implementation

All the above procedures are implemented in R scripts that are available with this paper. The Metacell package is also available from our group, implementing many of the algorithms and concepts used here, as well as scaling for larger datasets and further optimizations.

Gene functional annotation

We used blastp (with parameters $-evalue$ 1e-5 and $-max_target_seqs$ 1) to find for each protein of the merged *Nematostella* predicted proteome (JGI plus Vienna annotation) the most similar, if any, human, *Drosophila* and yeast homologs (retrieved from Uniprot). Additionally, we predicted for each protein the Pfam domain composition using Pfamscan with default curated gathering threshold. *Nematostella* TFs were identified using univocal Pfam domains for each structural TF family (de Mendoza et al., 2013). In the case of multi-TF families (Homeobox, Fox, bHLH, bZIP, DM, Smad, Myb, NR, RFX, RHD, SRF, Ets, T-box and Sox), we used phylogenetic analyses for each family in order to classify them into specific subfamilies (together with the complete TF sets of additional 10 animal species, including *Homo sapiens* and *Drosophila melanogaster* for reference annotation).

Phylogenetic distribution estimation

We used the complete predicted proteomes of 31 species at key phylogenetic positions in order to compute orthogroups, including an extensive set of 10 cnidarian species (*Acropora digitifera*, *Aiptasia pallida*, *Anthopleura elegantissima*, *Edwardsiella lineata*, *Fungia scutaria*, *Nephyrogorgia* sp., *Alatina alata*, *Atolla vanhoeffeni*, *Hydra magnipapillata*, *Podocoryne carnea*), 6 other planulozoans (*Homo sapiens*, *Branchiostoma floridae*, *Drosophila melanogaster*, *Tribolium castaneum*, *Capitella teleta*, *Lottia gigantea*), 6 other metazoans (*Trichoplax adhaerens*, *Amphimedon queenslandica*, *Oscarella carmela*, *Sycon ciliatum*, *Mnemiopsis leidyi*, *Pleurobrachia bachei*) and 8 non-metazoan eukaryotes (*Salpingoeca rosetta*, *Capsaspora owczarzaki*, *Creolimax fragrantissima*, *Saccharomyces cerevisiae*, *Spizellomyces punctatus*, *Dictyostelium discoideum*, *Arabidopsis thaliana*, *Naegleria gruberi*). We computed reciprocal blast results between all complete proteomes, with fixed database size and e-value threshold of 1e-04. Based on these reciprocal blast results, orthogroups were computed using orthoMCL algorithm (Li et al., 2003) with an inflation value (l parameter) of 1.3. We parsed these orthogroups using a parsimony criterion in order to generate an age estimation for each *Nematostella* gene.

Cross-species transcriptome comparison

In order to obtain *Nematostella*, *C.elegans* and mouse orthologous pairs, we repeated the same orthoclustering strategy as described above but removing non-metazoan eukaryotes from the dataset, and instead adding two nematodes (*C.elegans* and *Trichinella spiralis*) and mouse (*Mus musculus*). For *Nematostella*-*C.elegans* comparisons, we used the cell cluster normalized expression

values from the whole-organism scRNA-seq dataset generated by Cao et al. (Cao et al., 2017). *Nematostella-C.elegans* orthologs (see above) were used to merge the cell cluster expression profiles in both species. Both matrices were regularized before merging and then quantile normalized. We then used the expression profile of 686 highly variable genes across samples (with a fold-change > 2 in at least 1 *C.elegans* cluster and a fold-change > 2 in at least 1 *Nematostella* cluster) to compute the Pearson correlation between samples.

For *Nematostella*-vertebrate comparisons, we used the normalized expression values for amniote organ-specific transcriptomes generated by Brawand et al. (Brawand et al., 2011). *Nematostella*-human orthologs (see above) were used to merge this matrix with *Nematostella* cell cluster gene expression (represented by fraction of total molecules in the meta-cluster). Both matrices were regularized before merging and then quantile normalized. We then used the normalized expression profile of 906 highly variable genes across samples (with a fold-change > 2 in at least 4 organ samples and a fold-change > 2 in at least 1 *Nematostella* cluster) to compute the Pearson correlation between samples.

For neuronal types comparisons (Figure 4) we used the same *C.elegans* dataset and the mouse neuronal scRNA-seq generated by Tasic et al. (Tasic et al., 2016). Pairwise cross-species ortholog normalized expression tables were constructed as described above. In this case, this gene ortholog expression across neuronal cell clusters was used to compute gene-gene Pearson correlation. This correlation value was then employed to compute gene modules and to identify pairs of highly correlated genes.

Motif analysis

Inferred DNA-binding motif preferences for *Nematostella* TFs were obtained from CisBP database (<http://cisbp.ccb.utoronto.ca/>) (Weirauch et al., 2014). Briefly, experimentally determined Position Weight Matrices (PWM) from different species are transferred to other species TF given a certain threshold of primary protein sequence identity. In the case of *Nematostella*, CisBP contains 1,347 PWMs (excluding Transfac) corresponding to 255 *Nematostella* TFs. In order to reduce redundancy, we computed the binding energy of these 1,347 PWMs in all *Nematostella* promoters (TSS -200/+50 bp) and calculated the motif-motif correlation based on these energies. A correlation threshold of 0.7 was used to cluster motifs and a single representative of each cluster was selected for downstream analyses, resulting in a reduced set of 384 PWMs (Figure S6A).

We extracted promoters sequences using -200:+50 bp from annotated TSSs and associated sequences with metacells whenever their gene was at least two fold overexpressed in the module compared to the background. Enhancers sequences were defined based on Schwaiger et al., 2014 (Schwaiger et al., 2014) and grouped into metacells if their closest TSS was induced in the module at least two fold. For a short sequence element $s[1..k] = s_1, \dots, s_k$, and a 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 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 250bp (for promoters) or 150bp (for enhancers) were determined. These quantiles values were then used as thresholds to determine 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's promoters/enhancers 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 resulted 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 (ie. at least five genes in the metacell gene set with the motif in their promoters/enhancers) and a background count of 100 in this module were considered.

ATAC-seq analysis

ATAC-seq reads were trimmed to 40nt and then mapped into *Nematostella* genome v1.0 (<https://genome.jgi.doe.gov/Nemve1/Nemve1.home.html>) using Bowtie2 (Langmead and Salzberg, 2012) with the following parameters: `-D 200 -R 3 -N 1 -L 20 -i S,1,0.50-gbar 1-no-unal`. Duplicates reads were removed and then reads starting positions were shifted +5bp (for reads mapping in the forward strand) or -4 bps (for reads mapping in the reverse strand) in order to account for the Tn5 9bp insertion (Buenrostro et al., 2013). Mapped reads were extended to 50bp in total and 1bp-resolution coverage statistics over each of the genomes were computed.

To control for ATAC-seq coverage and variable ATAC-seq efficiencies, we transformed raw coverage values to quantile values. ATAC peaks were defined as regions with coverage quantiles over 0.985, merging peaks located at < 20bp and asking for a minimum peak size of 40bps. We did that for the whole-adult wild-type ATAC-seq and, independently, for the neuronal ELAV+ ATAC-seq experiments, in order to detect neuronal peaks that might not have enough coverage in whole-animal ATAC experiments. We then merged these two resulting peak intervals sets. In order to normalize our comparisons, we normalized peak size to 140bp, by taking the peak maximum ATAC signal position and extending it ± 70 bps. In downstream analysis, ATAC signal is indicated as Reads per Million (RPMs). ATAC signal in peaks/regulatory sites is computed by summing all ATAC reads within the peak (and normalizing by sequencing depth, RPMs).

Neuronal (ELAV+) enriched peaks (Figure 7E, in blue) were defined as peaks with 2-fold ELAV+ over ELAV- coverage, a minimum coverage of 1 RPM, and a p value < 0.05 (BH correction) computed with DESeq2 (Love et al., 2014) using the two ELAV+ versus the two ELAV- ATAC-seq experimental replicates. The same strategy identified ELAV- enriched peaks (Figure 7E, in red).

Motif enrichment in neuronal (ELAV+) enriched peaks was performed as described above. Briefly, for each motif PWM, the 0.98 quantiles of genome-wide binding energies in windows of 140bps (the fixed size of our ATAC peaks) were determined. These quantiles values were then used as thresholds to determine motif occurrence for each PWM at each element. The frequency of occurrence of each motif was computed for the neuronal peaks and all peaks. Enrichments were assessed statistically using a hypergeometric test.

DATA AND SOFTWARE AVAILABILITY

All data was deposited in GEO under accession number GSE95723. Processed data, annotation tables, and code for reproducing the analysis is available from our website (http://compgenomics.weizmann.ac.il/tanay/?page_id=724).

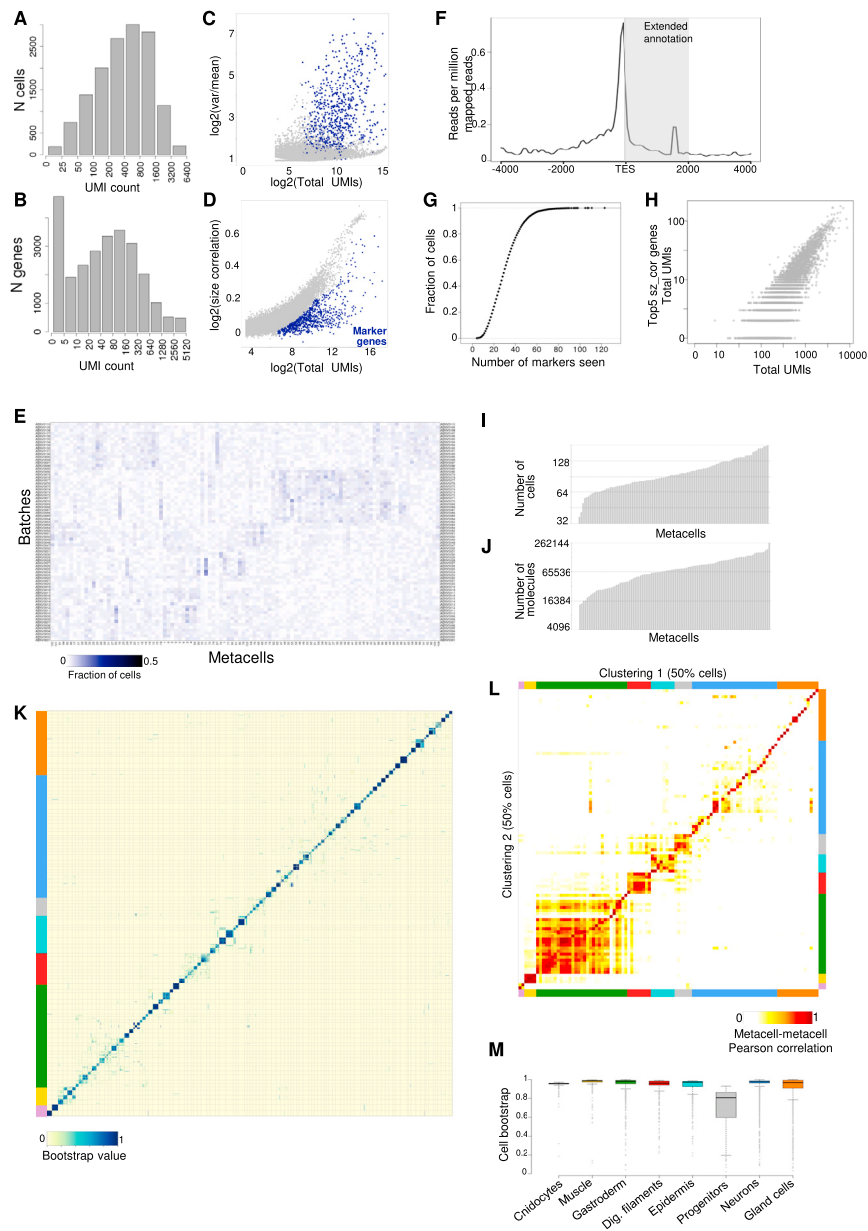


Figure S1. *Nematostella* Adult scRNA-Seq UMI Statistics and Metacell Analysis, Related to Figure 1

- (A) Distribution of total RNA molecules per cell.
- (B) Distribution of total RNA molecules per gene.
- (C) Relationship between gene expression var/mean across cells (y axis) and gene total expression (x axis). Marker genes selected for cell clustering are shown in blue.
- (D) 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.
- (E) Fraction of cells in each metacell originating from each of the batches (each batch representing half a 384-wells MARS-seq plate, see STAR Methods).
- (F) Distribution of mapped reads around Transcription End Sites (TES) of *Nematostella* originally annotated genes, we expanded the 3' end of each gene 2Kb (or until the next gene in the same strand) to capture reads mapping downstream of the original TES.
- (G) Cumulative distribution of number of marker genes detected per single cell.
- (H) Correlation of total UMIs per cell for all genes (x axis) versus total UMIs per cell for the top-5 size correlated genes.
- (I) Total number of cells per metacell.
- (J) Total number of molecules per metacell.
- (K) Bootstrap analysis. Heatmap representing the frequency of cell-to-cell association in 1,000 bootstrap subsamplings.

(legend continued on next page)

(L) Clustering robustness analysis. The heatmap shows the Pearson correlation (based on the expression profiles of the 700 marker genes) between two clustering solutions based on random selection of half of the cells. Color bars identify the new metacells based on similarity to the original clustering broad clusters (see [Figure 1](#)).

(M) Distribution of broad cell cluster association frequencies in 1,000 bootstrap subsamplings.

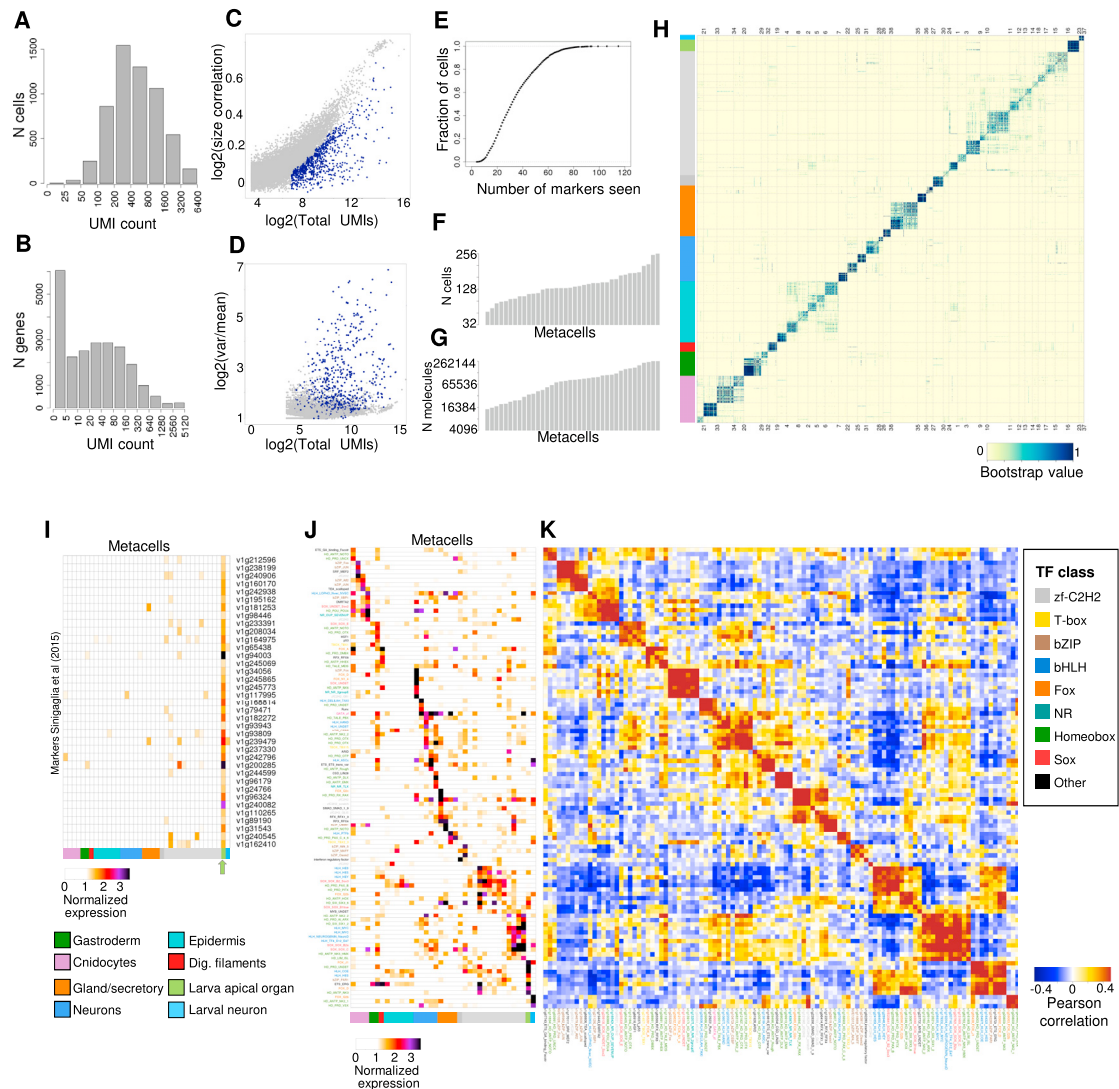


Figure S2. *Nematostella* Larval scRNA-Seq UMI Statistics and Metacell Analysis, Related to Figure 2

- (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) Relationship between gene expression var/mean across cells (y axis) and gene total expression (x axis). Marker genes selected for cell clustering are shown in blue.
- (E) Cumulative distribution of number of marker genes detected per single cell.
- (F) Total number of cells per metacell.
- (G) Total number of molecules per metacell.
- (H) Bootstrap analysis. Heatmap representing the frequency of cell-to-cell association in 1,000 bootstrap subsamplings.
- (I) Heatmap showing the expression of apical organ markers defined by (Sinigaglia et al., 2015) across larval metacells. Notice the consistent co-expression of most of them in the specific cell cluster we defined as apical organ cells (arrow).
- (J) TF expression profile across larval metacells. TF names and IDs are indicated and color-coded according to TF structural class.
- (K) TF-TF Pearson correlation based on expression profile across larval metacells.

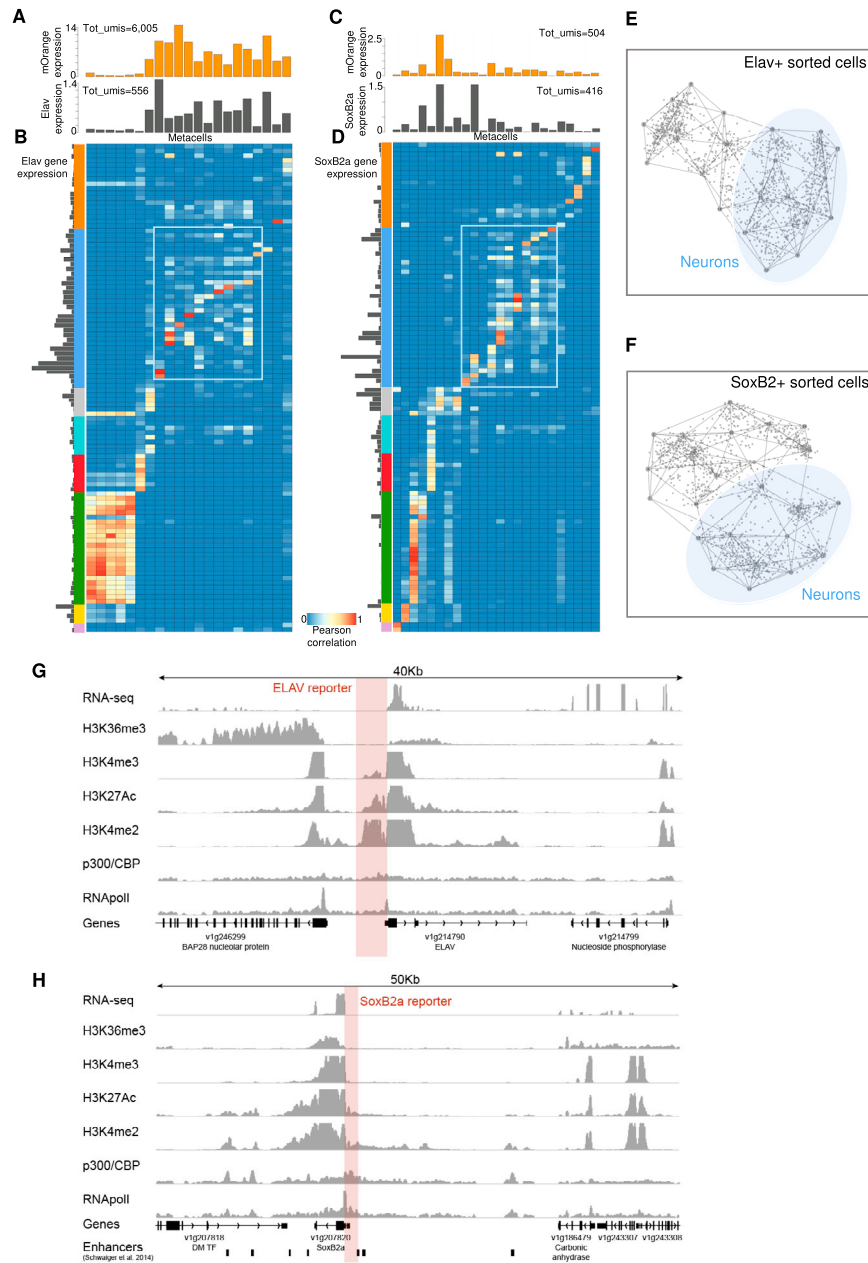


Figure S3. *Nematostella* ELAV and SoxB2a Reporter Lines scRNA-Seq Analysis, Related to Figure 4

(A) Expression of mOrange reporter and ELAV genes across the 17 metacells defined from 1,500 sorted Elav+ cells.

(B) Heatmap showing the Pearson correlation (based on the expression profiles of the 413 marker genes used to cluster Elav+ cells) between Elav+ metacells (columns) and the general metacells (rows, see Figure 1). The expression of Elav in the general metacells is shown as an horizontal barplot. Elav metacells mapping to neuronal metacells are highlighted with a blue rectangle.

(C and D), Same as (A) and (B) for 1,500 SoxB2a+ cells.

(E) 2d projection of the Elav+ cells, neuronal cells are highlighted in blue.

(F) same as (E) for SoxB2a+ cells.

(G) Chromatin features around the ELAV gene. The promoter sequence cloned for the Elav+ reporter line is highlighted in red (Nakanishi et al., 2012).

(H) Same as (G) for SoxB2a (Richards and Rentzsch, 2014). Notice that in the case of SoxB2a there are multiple regulatory elements both upstream and downstream of the cloned promoter and whose regulatory inputs are not captured by the SoxB2a reporter line. This may explain the discrepancies observed between SoxB2 and reporter mOrange expression.

All expression values are shown as molecules per 1,000 UMIs. Chromatin and bulk RNA-seq data from Schwaiger et al. (2014).

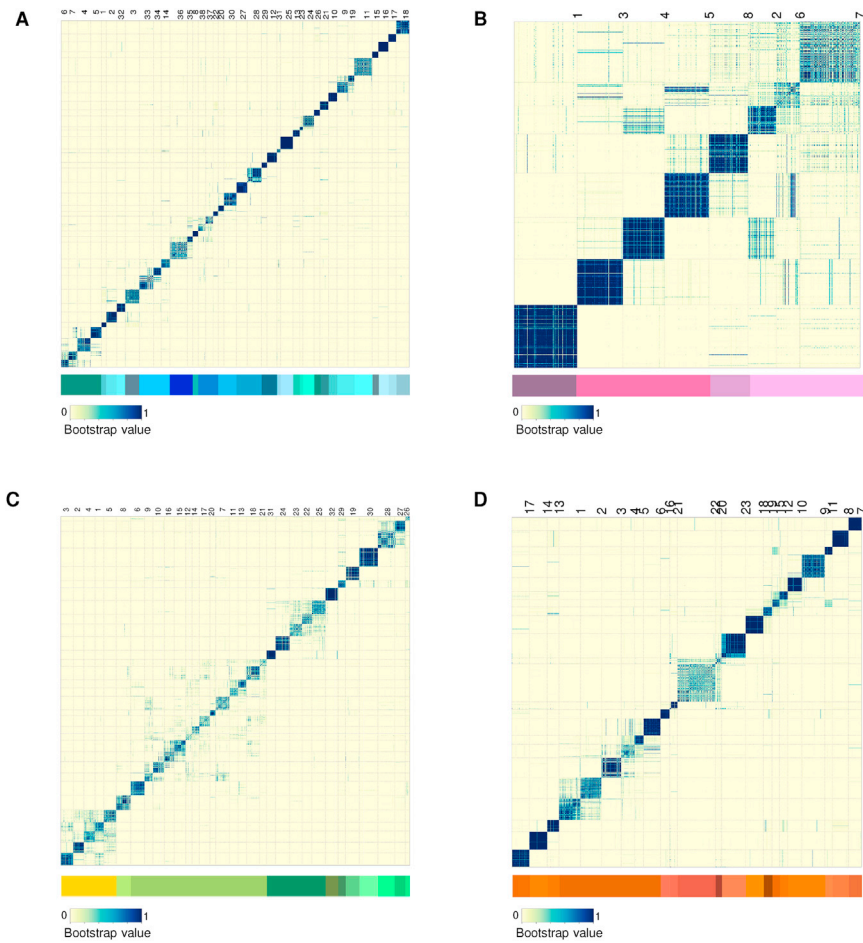


Figure S4. *Nematostella* Neuronal, Cnidocyte, Muscle/Gastrodermis, and Gland/Secretory Cell Types Characterization Supplementary Analyses, Related to Figures 4 and 5

(A) Bootstrap analysis of neuronal cells. The heatmap represents the frequency of cell-to-cell association in 1,000 bootstrap subsamplings. The cluster numbers and color-coding used in Figure 4 are shown.

(B–D) Same as (A) for cnidocyte (B), muscle/gastrodermis (C), and gland/secretory(D) cell types (as shown in Figure 5).

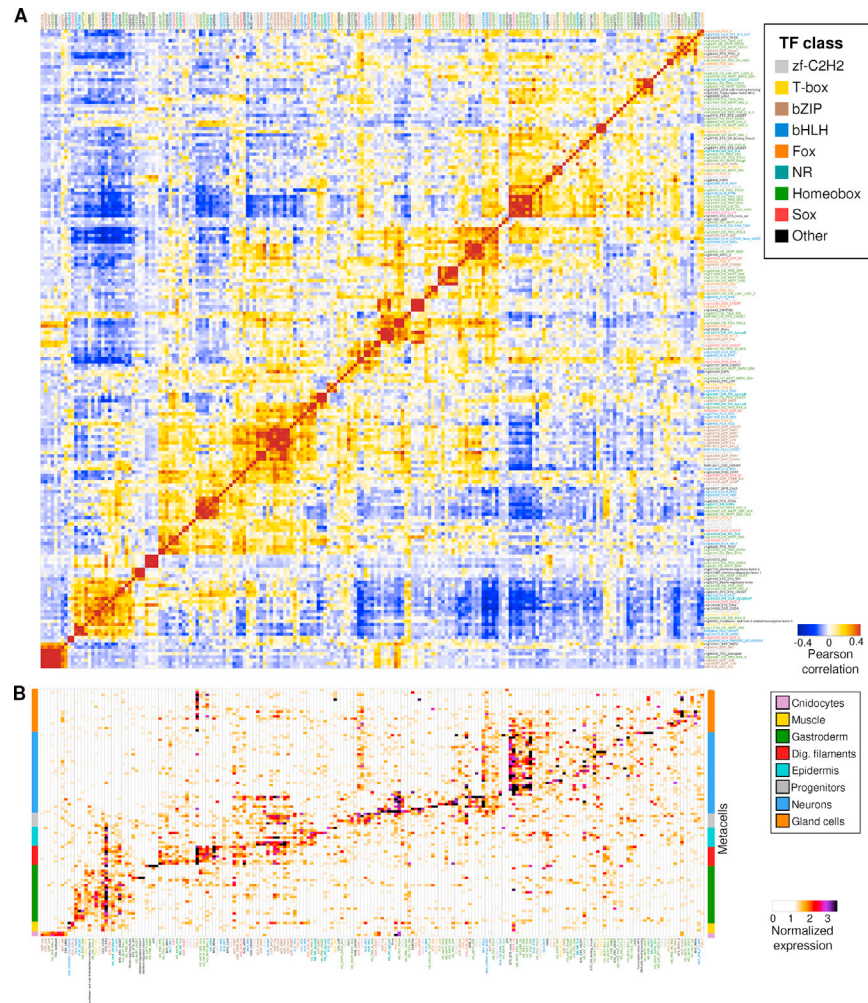


Figure S5. *Nematostella* General TF Map with Names and Gene IDs, Related to Figure 6

(A) TF-TF Pearson correlation based on expression profile across cell clusters.

(B) TF expression profile across cell clusters. TF names and IDs are indicated and color-coded according to TF structural class.

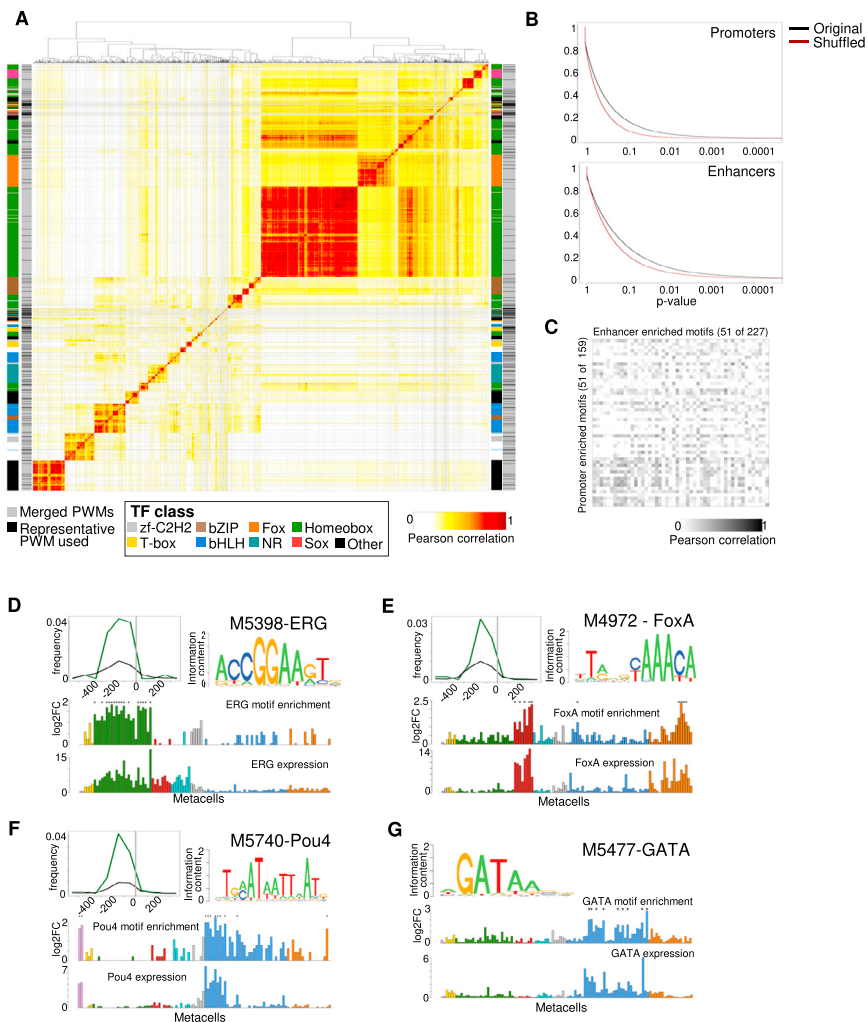


Figure S6. *Nematostella* Motif Enrichment Supplementary Analyses, Related to Figure 7

(A) Motif-motif Pearson correlation based on occurrence of each motif in *Nematostella* promoters. A single representative PWM from each metacell (indicated in black in the flanking colorbars) was used for downstream analyses. PWMs were obtained from CisBP database (Weirauch et al., 2014).

(B) Cumulative distribution of p values for the hypergeometric motif enrichment tests in promoters and enhancers in the original dataset (black) and shuffling metacell-gene associations (red).

(C) Motif-motif Pearson correlation between shared promoter (columns) and enhancer (rows) enriched motifs (presented in the same order as promoter motifs in Figure 4), showing that there is no correspondence of the same motif enrichments at enhancers and promoters.

(D) Ets-ERG TF motif promoter fold-change enrichment in 100bp windows around the TSS for a particular gene set with a significant motif enrichment (green line, FDR < 0.01) and the TSS of background genes (black line) (top-left). Motif logo derived from promoters with significant Ets-ERG motif (top-right). Promoter motif enrichment (log₂ Fold change) across metacells (middle). Ets-ERG expression (molecules per 10,000 UMIs) profile across metacells (bottom).

(E-F) same as (D) for FoxA (E) and Pou4 (F) TFs, respectively.

(G) Same as (D) for GATA TFs in enhancers. Asterisks indicate significant enrichments (FDR < 0.001). IDs represent CisBP entries (Weirauch et al., 2014).

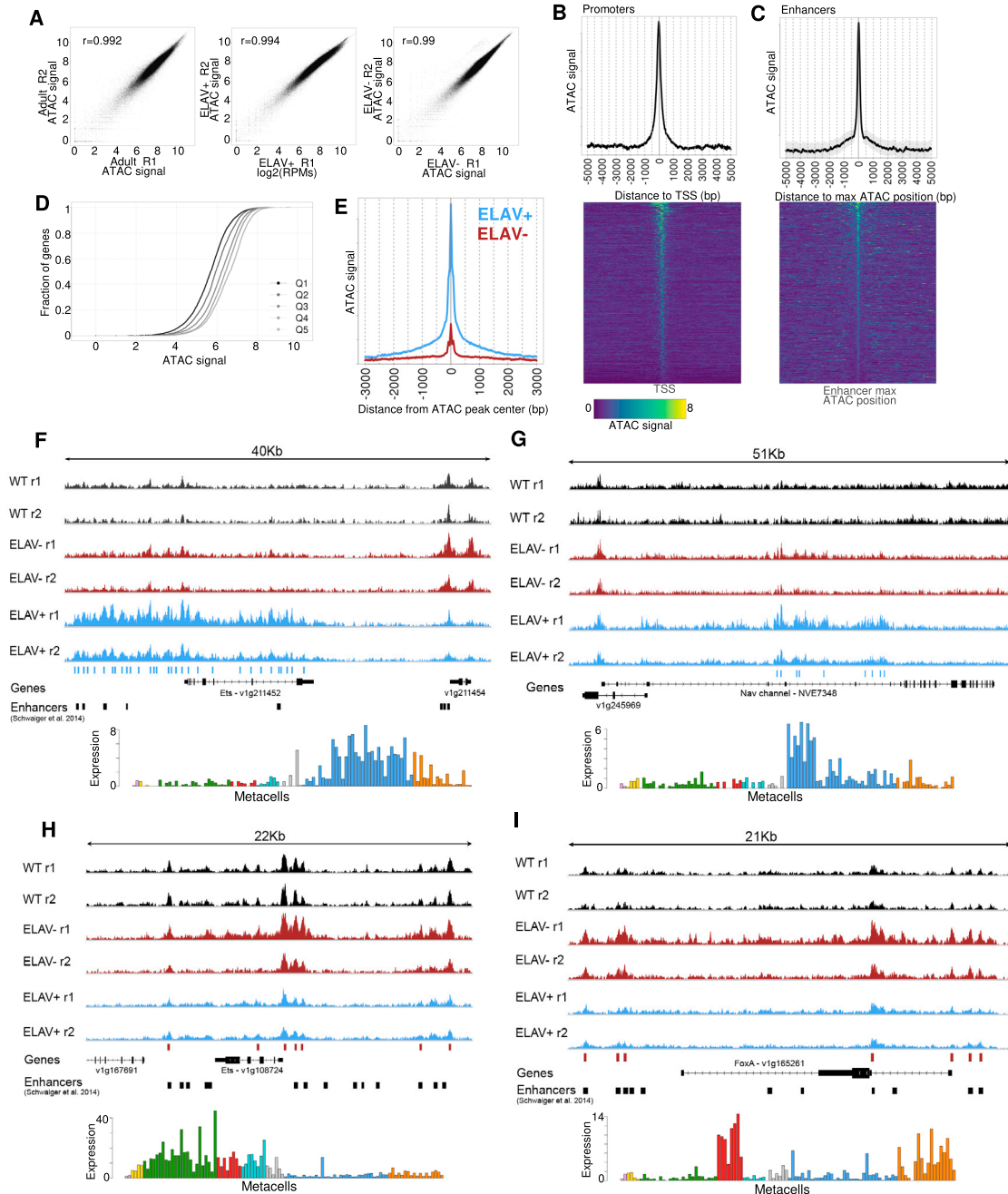


Figure S7. *Nematostella* ATAC-Seq Supplementary Analyses, Related to Figure 7

(A) ATAC-seq replicates reproducibility. Scatterplots compare the normalized coverage (reads per million) between replicates of adult ATAC-seq experiments (left), ELAV+ sorted cells (middle), and ELAV- sorted cells (right).

(B) Normalized ATAC signal density in *Nematostella* promoters, centered in the TSS. The heatmap show the normalized ATAC signal in individual sites, sorted by signal intensity.

(C) Same as (B) for enhancers (as defined by (Schwaiger et al., 2014)), centered in the maximum ATAC signal position.

(D) Relationship between fraction of cells expressing a gene (in our adult scRNA-seq dataset) and whole-adult ATAC signal in the gene promoter. Genes are grouped into five quantiles, from smaller (Q1) to higher (Q5) cell fraction.

(E) ELAV+ ATAC signal (blue) and ELAV- (red) ATAC signal around neuronal regulatory sites, centered in the maximum ELAV+ ATAC signal position.

(F) Chromatin accessibility landscape associated to the neuronal TF Ets. The tracks show the normalized ATAC signal for each wild-type whole-adult ATAC replicate (black), the two ELAV- (non-neurons) replicates (red), and the two ELAV+ (neurons) replicates (blue). Peaks that are significantly enriched in ELAV+ samples are highlighted by blue bars. Enhancer intervals as defined by (Schwaiger et al., 2014) are indicated. The barplot below shows the expression profile of Ets TF across adult cell clusters.

(legend continued on next page)

(G) same as (F) for a neuronal sodium voltage-gated ion channel.

(H and I) same as (F) for two non-neuronal TFs, Ets-ERG (H) and FoxA (I). In this case, the regulatory sites enriched in ELAV- (non-neuronal cell population) ATAC samples are highlighted in red.

All expression values are shown as molecules per 1,000 UMIs.