nature ecology & evolution

Article

https://doi.org/10.1038/s41559-025-02906-1

Decoding cnidarian cell type gene regulation

Received: 30 April 2025

Accepted: 23 October 2025

Published online: 22 December 2025

Check for updates

Anamaria Elek[1,8,9], Marta Iglesias[1,9] ✉, Lukas Mahieu[2], Grygoriy Zolotarov[1], Xavier Grau-Bové[1], Stein Aerts[2,3,4] & Arnau Sebé-Pedrós[1,5,6,7] ✉

Animal cell types are defined by differential access to genomic information—a process orchestrated by the combinatorial activity of transcription factors that bind to *cis*-regulatory elements (CREs) to control gene expression. Changes in these gene regulatory networks (GRNs) underlie the origin and diversification of cell types, yet the regulatory logic and specific GRNs that define cell identities remain poorly resolved across the animal tree of life. Cnidarians, as early-branching metazoans, provide a critical window into the early evolution of cell type-specific genome regulation. Here we profiled chromatin accessibility in 60,000 cells from whole adults and gastrula-stage embryos of the sea anemone *Nematostella vectensis*. We identified 112,728 putative CREs and quantified their activity across cell types, revealing pervasive combinatorial enhancer usage and distinct promoter architectures. To decode the underlying regulatory grammar, we trained sequence-based models predicting CRE accessibility and used these models to infer cell type similarities that reflect known ontogenetic relationships. By integrating sequence motifs, transcription factor expression and CRE accessibility, we reconstructed the GRNs that define cnidarian cell types. Our results show the regulatory complexity underlying cell differentiation in a morphologically simple animal and highlight conserved principles in animal gene regulation. This work provides a foundation for comparative regulatory genomics to understand the evolutionary emergence of animal cell type diversity.

In multicellular animals, cell type-specific gene expression is orchestrated by transcription factors (TFs), which recognize specific sequence motifs located within CREs such as gene promoters and enhancers. These TF–CRE networks ultimately interpret genomic information in each cell, determining the transcriptional state of individual genes and collectively shaping specific gene regulatory networks (GRNs). By measuring the transcriptional output of these gene programs, single-cell transcriptomics provides unprecedented insights into the molecular diversity of cell types across animal lineages[1–9]. However, our understanding of the structure and logic of the regulatory programs that define cell types remains limited for most species except for fruit fly[10–12] and vertebrates[13–17].

The development of single-cell chromatin accessibility sequencing (scATAC-seq) assays[18], together with the generation of high-quality genomes and gene expression data, has created new avenues to study whole-organism, cell type-specific gene programs in non-model species. In the context of evolutionary studies, dissecting cell type regulatory identity can offer new opportunities for cross-species comparisons

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [2]Laboratory of Computational Biology, VIB Center for AI and Computational Biology, Leuven, Belgium. [3]VIB-KU Leuven Center for Brain and Disease Research, Leuven, Belgium. [4]Department of Human Genetics, KU Leuven, Leuven, Belgium. [5]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [6]ICREA, Barcelona, Spain. [7]Tree of Life, Wellcome Sanger Institute, Hinxton, UK. [8]Present address: Center for Molecular Biology of Heidelberg University (ZMBH), Heidelberg, Germany. [9]These authors contributed equally: Anamaria Elek, Marta Iglesias. ✉e-mail: marta.iglesias@crg.eu; arnau.sebe@crg.eu

that go beyond similarities in gene expression, instead focusing on *cis*-regulatory sequence grammars[19]—the content and arrangement of TF motifs—or on gene modules[20]—co-regulated sets of genes that can be co-opted modularly between cell types during development and evolution. These approaches ultimately bring us closer to bridging the gap between genome evolution and cell type diversification[9].

In this context, given their key phylogenetic position as sister group to all bilaterian animals, the study of cnidarians (anemones, corals and jellyfish) can offer insights into the evolution of animal regulatory complexity and cell types. Notably, cnidarian genomes show hallmarks of bilaterian gene regulation such as distal enhancer elements[21] that contact gene promoters through chromatin loops[22]. Furthermore, although historically considered simple animals with relatively few cell types, single-cell transcriptomics studies have revealed that cnidarians encode a diverse repertoire of cell types[2,7,23–26], including several neuronal and secretory cell types[2,27], distinct muscle cells[28,29] and stem cell populations[30,31]. In addition, cnidarians are defined by the presence of cnidocytes—specialized stinging cells that offer the opportunity to study the regulatory mechanisms underlying the emergence of new cell types[32–34].

To begin to understand the genomic basis of this cell diversity, we systematically dissected cell type *cis*-regulatory programs in the sea anemone *Nematostella vectensis* (Fig. 1a), including cell-specific open chromatin regions representing putative CREs (hereafter referred to simply as CREs), regulatory motif grammars defined by sequence models and GRNs defined from the integration of TF expression and target CRE accessibility. Comparative analyses uncovered shared and stage-specific CRE landscapes, with early accessibility often preceding gene activation, and cell type ontogenetic relationships reflected in regulatory similarities. Detailed dissection of developmentally convergent retractor muscle programs showed that shared effector genes can be governed by highly distinct regulatory states, highlighting how divergent regulatory architectures underlie similar cell phenotypes.

Results

Nematostella cell type-specific chromatin landscapes

To define CRE usage across *Nematostella* cell types, we profiled chromatin accessibility in 51,866 adult and 6,882 gastrula-stage single cells using 10x Genomics scATAC-seq (Fig. 1b and Extended Data Fig. 1a–c). We sequenced libraries to an average of 19,575 reads per cell and obtained a median of 2,788 fragments per cell (2,458 in adult; 1,869 in gastrula). Cells were grouped on the basis of their accessibility profiles into metacells[35], which served as the basic units for downstream analyses (Fig. 1b and Extended Data Fig. 1d,e). This resulted in 693 metacells for the adult stage and 95 for the gastrula stage, with each metacell containing a median of 67 and 53 single cells, respectively. We then clustered metacells based on their accessibility profiles using neighbour joining (NJ) (Extended Data Fig. 1f–h). We identified chromatin accessibility peaks using cluster-level aggregated pseudobulk ATAC-seq signal and iteratively merged overlapping peaks[36], generating a catalogue of 112,728 CREs across the 269 Mb *Nematostella* genome (Fig. 1c and Extended Data Fig. 1i). We assigned peaks to genes based on their distance to transcription start sites (TSSs) and covariation across cell types (Fig. 1d) and we annotated scATAC-seq cell clusters using previously defined scRNA-seq cell types[2,27] (Extended Data Fig. 1j–l; Methods). To achieve this, we calculated an ensemble gene accessibility score as a weighted sum of peak accessibility for each gene (Fig. 1d) and correlated these scores with gene expression to match scATAC-seq clusters to cell types defined by scRNA-seq (Extended Data Fig. 1j–l). This analysis resulted in 32 annotated cell clusters (22 in the adult, 10 in the gastrula), each with both specific and combinatorial gene accessibility patterns (Fig. 1b) and with cluster-specific accessible CREs ranging from 2,000 to 30,000 (median 21,156 CREs). To validate these cell type-specific CREs, we generated transgenic reporter lines for two predicted alternative promoters (APs) of the *Gabrb4* gene. The two promoters drove expression in either the tentacle retractor (TR) muscle or tentacle neurons, recapitulating their respective accessibility profiles (Fig. 1e).

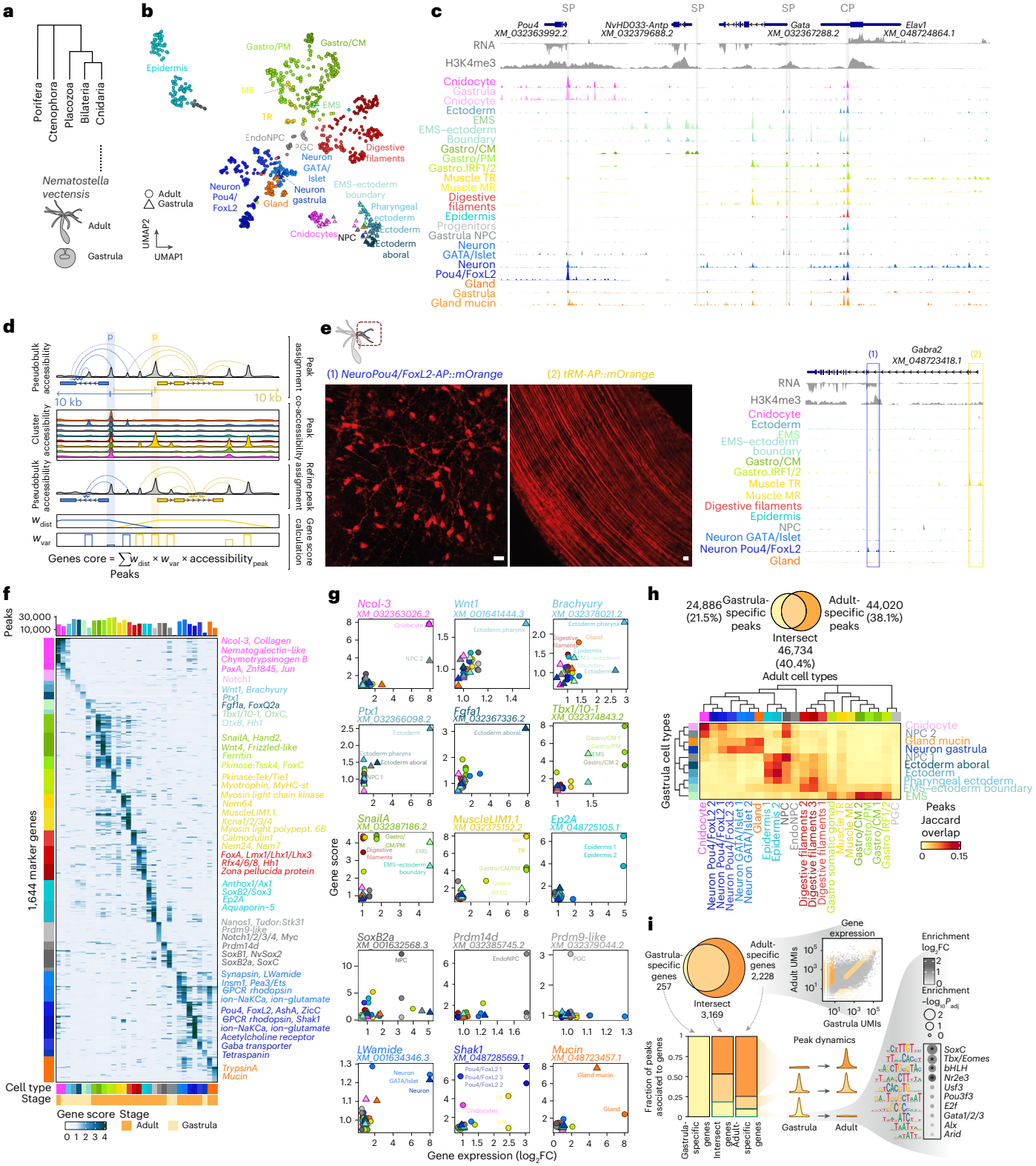
Adult cell clusters included eight previously described broad adult cell types[2], characterized by high CRE accessibility around known markers (Fig. 1f,g and Extended Data Fig. 2) such as *Ncol-3* (cnidocytes), *MuscleLIM* protein (retractor muscle), *EP2A* (epidermis) and *Shak3* ion channel (Pou4/FoxL2 neurons). Consistent with previous single-cell RNA sequencing (scRNA-seq) reports[37], we also identified three distinct clusters of adult progenitor cells. One represents adult neurosecretory progenitor cells (NPCs) characterized by differential accessibility near TF genes such as *SoxC*, *SoxB2a* and *Ath-like*[27,38]. Another, that we termed endodermal NPCs (endo-NPCs), exhibited accessibility near *Prdm14d*, a marker for endodermal neurogenesis[39]. The third precursor cluster probably represented primordial germ cells (PGCs), characterized by the differential accessibility near *Prdm9* (ref. 37). Gastrula-stage cell clusters included both differentiated cell types, such as gland cells, cnidocytes and neurons, as well as progenitor cells such as NPCs. We also identified the main germ layers and spatial territories within the gastrula[37,40]: ectoderm and aboral ectoderm, showing *Ptx1* (ref. 27) and *Ffg1a*[41] accessibility, respectively; endomesoderm (EMS, sometimes classified as mesoderm[40,42]), showing *Tbx1/10-1* (ref. 40) and *SnailA*[40] accessibility; and pharyngeal ectoderm (sometimes classified as endoderm[40,42]), showing *Brachyury*[40], *FoxA*[40] and *Wnt1* (ref. 43) differential accessibility (Fig. 1g and Extended Data Fig. 2).

We then compared CRE usage between adult and gastrula cell types, identifying 46,734 shared CRE (40.4%) between adult and gastrula (Fig. 1h). Comparisons of CRE accessibility revealed strong similarities between neurons, cnidocytes and gland cells, as well as between NPCs at both stages (Fig. 1h). Furthermore, the CRE landscapes of gastrula germ layers showed resemblances to some adult cell types, including similarities between EMS and gastrodermal/

Fig. 1 | Cell type-specific chromatin landscapes in *N. vectensis*. **a**, *Nematostella* phylogenetic position. **b**, UMAP two-dimensional projection of scATAC metacells, coloured by cell type, with broad cell type labels. Gastro/PM, Gastrodermis/parietal muscle; Gastro/CM, gastrodermis/circular muscle. **c**, Example regulatory landscapes for selected genes. Forward and reverse RNA signals are shown above and below baseline, respectively. Promoter peaks are highlighted with vertical grey bars. **d**, Peak assignment and gene accessibility score calculation strategy. Peaks up to 10 kb are assigned to genes unless they are downstream of another gene's promoter (p). When a peak is assigned to more than one gene, peak–peak co-accessibility is used to refine peak assignment. Gene score is then calculated as the sum of the accessibility of peaks assigned to a gene, weighted by distance from the TSS (w_{dist}) and peak variability across clusters (w_{var}). **e**, Transgenic reporter validation of *Gabrb4* (XM_048723418.1) APs. Images correspond to the tentacle region showing reporter expression in neurons and longitudinal muscle fibres (left) corresponding to the regulatory regions highlighted in the genome browser (right). Three animals were imaged independently with similar results; a representative image is shown. Scale bars: 10 µm. **f**, Heatmap of gene scores for marker genes across cell types. Colour code for genes indicates the cell type where gene has the highest score. Selected known markers are highlighted on the right. **g**, Comparison between gene accessibility scores and gene expression levels for selected marker genes. **h**, Euler diagram showing the total number of overlapping peaks (accessibility FC > 1.5) between the two life stages (top) and heatmap representing peak overlap between adult and gastrula cell types (bottom). Rows and columns are clustered based on peak overlap between cell types within each life stage. **i**, CRE dynamics across development. We first select genes expressed in adult and/or gastrula (Euler diagram, top) and then analyse the accessibility dynamics of CREs associated to these three gene groups across development (bottom). For CREs associated to genes expressed in adult and/or gastrula, which are accessible only in gastrula but not in adult, top enriched motifs ($\log_2$FC > 1 and adjusted *P* value (P_{adj}) < 0.05) are shown in the dotmap on the right. Asterisks indicate $P_{adj} < 0.01$ (hypergeometric test, FDR adjustment for multiple testing).

https://doi.org/10.1038/s41559-025-02906-1

mesenteric retractor (MR) muscles, between pharyngeal ectoderm and digestive filaments, and between ectoderm and epidermis. Although the precise developmental trajectories of these tissues remain to be fully characterized, these patterns align with the hypothesis of three germ layers in *Nematostella* as proposed by Steinmetz et al.[40]. Finally, we found that approximately 25% of the genes expressed exclusively in adult *Nematostella* had CREs accessible already at the gastrula stage, including 10% of CREs accessible only in the gastrula stage (Fig. 1i). This is consistent with findings in other species, where chromatin accessibility often precedes transcriptional activation[44], and could reflect the early activity of putative pioneer TFs[45,46], as revealed by enriched motifs such as Sox, Pou or GATA, in these gastrula CREs (Fig. 1i). Overall, our *Nematostella* single-cell accessibility atlas represents a comprehensive inventory of cell type-specific *cis*-regulatory landscapes in a non-bilaterian animal. This atlas is available for exploration through an interactive database and genome browser: https://sebelab.crg.eu/nematostella-cis-regulatory-atlas/ and https://sebelab.crg.eu/nematostella-cis-reg-jb2.



https://doi.org/10.1038/s41559-025-02906-1

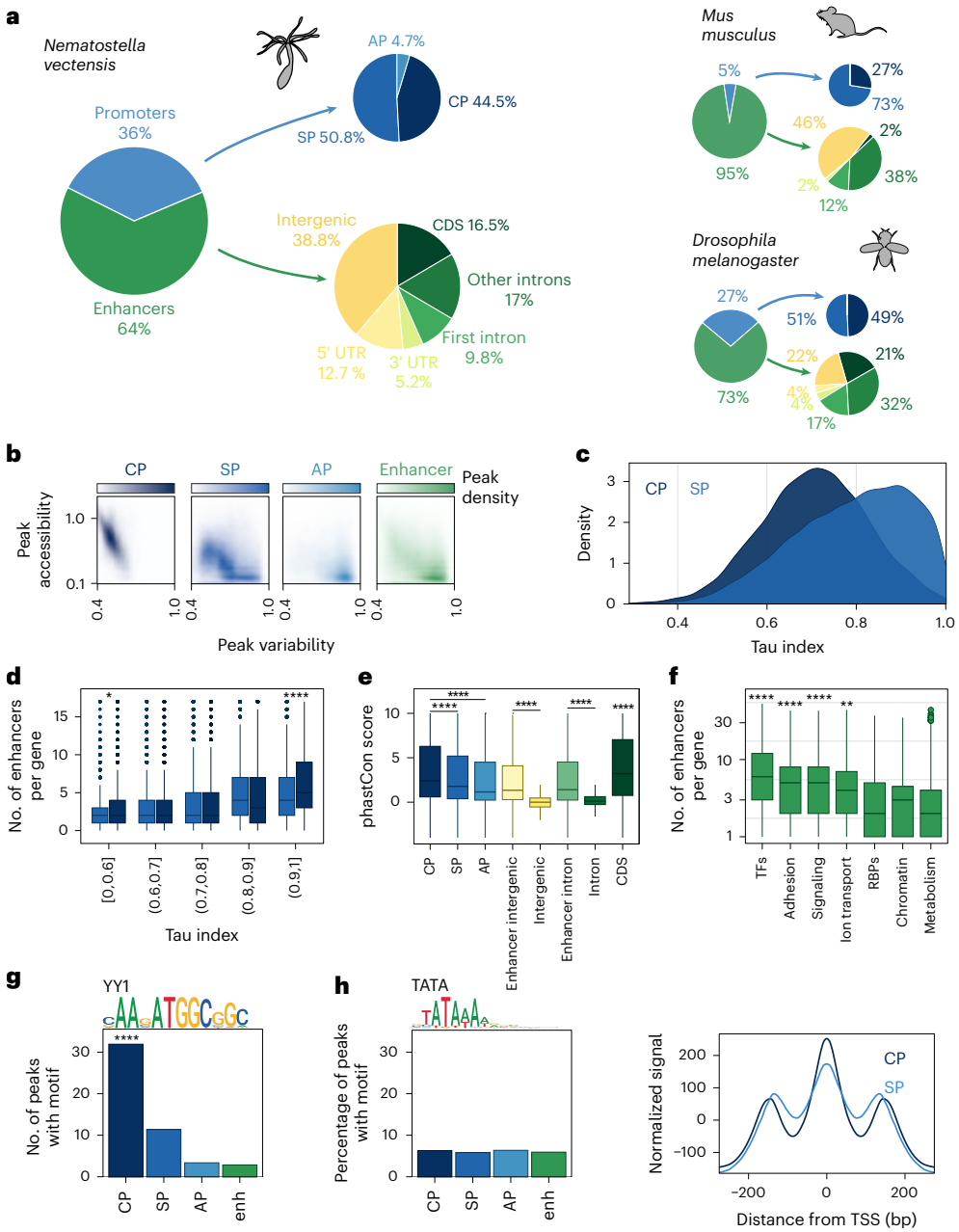


Fig. 2 | Prebilaterian gene regulatory architecture. a, Fraction of CREs classified as promoters and enhancers. Promoters are further classified as CP, SP and AP. Enhancers are classified based on their overlap with different genomic regions. The same is shown for *Nematostella* (top), mouse (bottom left) and *Drosophila melanogaster* (bottom right). CDS, coding sequence; UTR, untranslated region. **b**, Comparison of accessibility versus variability across cell clusters for different CRE classes. **c**, Cell type expression specificity (as measured by the Tau index[47]) for genes with CP and SP. **d**, Number of enhancers for group of genes with CP or SP and different levels of expression specificity (Tau index bins). **e**, Sequence conservation (phastCon score) of different promoter classes and enhancers overlapping intergenic and intronic regions. Conservation of intergenic and intronic regions not overlapping predicted enhancers is shown for comparison. **f**, Number of enhancers for different functional gene sets. Significance for each set compared to all (that is, basemean) is indicated. TF, transcription factor; RBP, RNA-binding protein. **g**, Fraction of peaks with YY1 motif in different CRE classes. **h**, Fraction of peaks with TATA motif in different CRE classes (left) and aggregated ATAC signal around promoters (right). $^{*}P < 0.05$, $^{****}P < 0.0001$ (one-sided Wilcoxon test).

Cnidarian gene regulatory architecture

We next investigated the different CRE configurations associated with *Nematostella* genes. First, we classified CREs into promoters and non-promoters (which we termed *enhancers*) using a combination of distance to TSS, histone post-translational modifications (H3K4me3)[21], and newly generated 5′ scRNA-seq data (Extended Data Fig. 3a). Among the 58,954 CREs identified in adult cell types, we classified 21,344 (36%) as promoters and 37,610 (64%) as enhancers (Fig. 2a and Extended Data Fig. 3b). These proportions are similar to those observed in *Drosophila*, which has 27% promoters and 73% enhancers, whereas in mice, the fraction of promoters among scATAC-defined CREs is smaller (5% versus 95% enhancers). In *Nematostella*, enhancers are predominantly located in intergenic regions (38.8%), followed by intronic regions (26.8%). Enhancers in mouse also tend to be found in intergenic regions (46%), whereas in *Drosophila* 49% are intronic and 22% are intergenic (Fig. 2a).

Focusing on promoters, we identified approximately half (44.5%) as constitutively accessible across all cell types (constitutive promoters

(CP)), while roughly another half (50.8%) were cell type-specific (specific promoters (SP)). A smaller fraction (4.7%) represented alternative promoters (AP) of the same gene accessible in different cell types (Figs. 1e and 2a and Extended Data Fig. 3a). These proportions are similar in *Drosophila* (51% SP versus 49% CP), whereas in mouse SP promoters are more frequent (73% SP versus 27% CP). *Nematostella* CPs showed higher and less variable accessibility compared to SP, AP and enhancers (Fig. 2b)—a pattern similar to that observed in *Drosophila*[44]. Furthermore, CPs were generally associated with genes expressed across several cell types, whereas genes with SPs tended to exhibit more restricted, cell type-specific expression (Fig. 2c), as measured by the Tau index[47]. Regardless of promoter type, genes with cell type-specific expression were linked to an increased number of associated enhancer elements (Fig. 2d), as observed in other species[48]. Comparing CRE sequence conservation across cnidarian genomes, we found that CPs are significantly more evolutionarily conserved than SPs, APs or enhancers (Fig. 2e). Furthermore, TFs represent the gene class with the highest number of associated enhancers[21] (Fig. 2f).

We also examined sequence motifs enriched in different promoter types and found that YY1 motif was strongly enriched in CP (Fig. 2g). *YY1* is a metazoan-specific TF that has been involved in enhancer–promoter contacts in different cell types[49], suggesting that *Nematostella* CPs may rely on this factor for integrating regulatory signals from their associated enhancers. In bilaterian animals, adult cell type-specific promoters—often called Type I promoters[50,51]—are characterized by the presence of TATA motifs and fuzzy nucleosomes. In contrast, *Nematostella* SP have well-positioned flanking nucleosomes and lack TATA motifs (Fig. 2h and Extended Data Fig. 3e), suggesting that this class of promoters may be a bilaterian-specific feature. These findings offer a comprehensive perspective on the landscape of cell type-specific gene regulation in a non-bilaterian animal. Our results highlight similarities to bilaterians that have relatively compact genomes like *Drosophila*, such as CRE-type proportions and their genomic distributions, while also revealing key differences, including the absence of TATA-containing Type I promoters.

Nematostella cis-regulatory programs

Having defined the CREs accessible in different cell types, we next sought to identify the key TFs and *cis*-regulatory sequences in each cell type. To identify sequence motifs, which represent putative TF binding sites that are important for CRE accessibility, we employed two complementary approaches: (1) calculating motif enrichments in accessible CREs using both de novo discovered and known motif collections (Extended Data Fig. 4) and (2) training sequence-to-function machine learning models that explain the relationship between sequence features and accessibility[10,52–54], to then extract important model features and discovering motifs[55,56] (Extended Data Fig. 5). To reduce redundancy in motif annotations, we grouped similar motifs into broader archetypes[57], and then systematically compared the motif collections obtained with each method (Extended Data Fig. 6a–i). Motif enrichment analyses uncovered a larger number of motifs (1,292) compared to sequence models (637), with different sequence models only recovering up to 15% (n = 193) of the enriched motifs (Extended Data Fig. 6j). This discrepancy is probably related to the fact that sequence models prioritize motifs that are predictive of accessibility patterns rather than capturing an exhaustive set of all enriched motifs. However, it is worth noting that up to 30% of motifs (n = 216) identified by sequence models were absent from enrichment analyses (Extended Data Fig. 6h–i, new motifs), suggesting that sequence models offer higher sensitivity and can detect important motifs with fewer genome-wide binding sites.

Beyond motif discovery, we also leveraged sequence models to investigate cell type-specific CRE codes, considering both motif composition (lexicons) and the combinatorial rules governing motif arrangement, orientation and spacing (syntax). For example, in adult cnidocytes, the most common motif grammar contained Pou4 in combination with E-box bHLH, Fox and zf-C2H2 motifs (Fig. 3a and Extended Data Fig. 6k). Across all cell types, we identified 15–36 key motifs per cell type, and each CRE contained a median of thee to four motif instances (Fig. 3b). The co-occurrence of TF binding motifs within CREs ranged from 10% to 75% depending on the cell type (Fig. 3c). When analysing motif combinations, we found that most motif pairs and triplets exhibited flexible order and orientation (Fig. 3d), with only a few exceptions involving YY1 and zf-C2H2 binding sites. This pattern observed in *Nematostella* is compatible with a billboard-like model of TF binding sites[58], which emphasizes the importance of TF combinations for CRE function while allowing flexibility in the arrangement, order and spacing of these motifs. Similarly, TF motif footprinting analyses in human tissues suggest that CRE accessibility is shaped by synergistic, yet largely independent, binding of the cognate TFs[57].

To further link CRE sequences to TF function, we assigned motifs to specific *Nematostella* TFs using a combination of orthology, sequence similarity-based motif transfer[59], and correlations between TF expression and motif accessibility (Extended Data Fig. 6l–n; Methods). This analysis enabled us to predict candidate binding motifs for 96% (571 of 590) of expressed/accessible TFs in *Nematostella*. Then, we compared TF expression to the aggregated accessibility of the assigned motif in each cell type[60] (TF motif activity), observing good agreement between TF expression and TF activity (Fig. 3e). For example, we found that *PaxA* was expressed and active specifically in cnidocytes, *FoxA* and *Rfx4/6/8* in digestive filaments, *Hes2* in ectodermal cells and *FoxQ2d* in epidermis. *Pou4* is expressed and active in cnidocytes and one broad type of neurons; while *Gata*, *Islet* and *OtxC* are active in the other broad neuronal type (Fig. 3f and Extended Data Fig. 7).

We integrated CRE accessibility with TF motif binding scores and gene expression to infer cell type-specific GRNs. Specifically, we used in silico chromatin immunoprecipitation[61] (ChIP), which links a TF to candidate target CREs if (1) the CRE contains a high-scoring motif for the TF and (2) the CRE accessibility is correlated with the expression of the TF across metacells. This allowed us to reconstruct a global TF–CRE network, which we then partitioned per cell type based on TF motif activity, TF expression and target CRE accessibility in each cell type. In the global GRN model, TFs lacking self-regulation were predicted to target a median of 66 genes, whereas self-regulating TFs targeted a median of 196 genes (Fig. 3g). This suggests that self-regulating TFs may control larger networks of effector genes, contributing to long-term maintenance of cell functions. From a complementary perspective, each effector gene is predicted to be regulated by a median of three TFs (Fig. 3h). Within a cell type, TFs were found to regulate very different sets of genes (median overlap fraction between predicted targets, 0.03) (Extended Data Fig. 8a). Across cell types, TFs tend to regulate distinct sets of genes as a function of the number of cell types in which these TFs are active (Extended Data Fig. 8b), and most TFs are predicted to bind only one CRE per gene (91%, for genes with more than one associated CRE) (Extended Data Fig. 8c). The analysis of predicted GRN structure also highlights important TF for cell type identity (Fig. 3i,j, Extended Data Fig. 8d–i). For example, the reconstructed GRN for adult cnidocytes (Fig. 3j) indicates that *FoxL2*[2] is the TF with most regulatory connections (highest degree of centrality) and *Pou4*[62] is the TF bridging most submodules in the network (highest betweenness centrality) and also the global cnidocyte TF regulator with connections spread most evenly across different submodules (highest participation). This GRN model also highlights other TFs known to be important in cnidocyte differentiation, such ase *PaxA*[63], *Sox2* (ref. 34), *Znf845* (ref. 33) and an unclassified Fox TF[64] (Fig. 3i). In each cell type, we also identified a subset of TFs with predicted self-regulation, for example *FoxL2*, *Pou4* and *Sox2* in the case of cnidocytes (Fig. 3i,j).

Cell type relationships defined by regulatory characters

We explored the relationships between the identified *Nematostella* cell clusters by comparing different regulatory features. We first grouped

https://doi.org/10.1038/s41559-025-02906-1

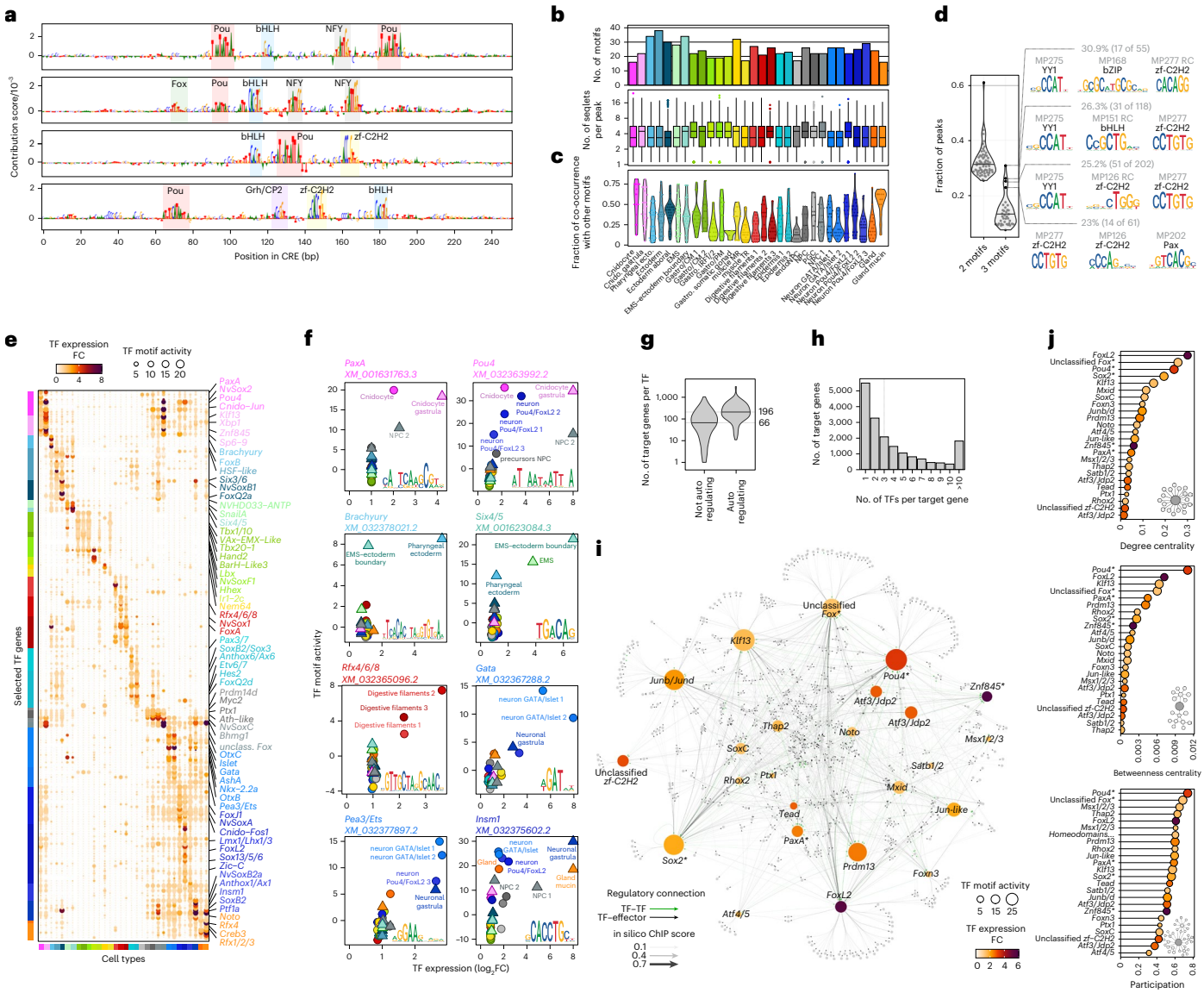


Fig. 3 | *Nematostella* cell type regulatory programs. a, Nucleotide importance scores for four representative cnidocyte CREs, highlighting detected TF motifs. **b**, Top: total number of motifs learned from sequence models per cell type. Bottom: total number of motif instances (seqlets) in each peak per cell type. **c**, Fraction of TF motif co-occurrences per cell type. We define two motifs as co-occurring when they appear nonoverlapping in the same CRE, for example, co-occurrence fraction of 0.5 would mean that 50% of CREs where TF motif is identified also have at least one more motif of another TF. **d**, Frequency distribution of motif pairs (doublets) and triplets that co-occur in more than 50 CREs. The violin plot (left) shows the fraction of peaks (out of all peaks with given motif combination) in which motifs appear in specific order and orientation. The examples of motif triplets that deviate from flexible ordering and orientation are highlighted (right). **e**, Dotmap showing TF motif activity (dot size) and expression (colorscale) for selected variable TFs across cell types. Cell types are colour coded as in Fig. 1. Colour code for TFs indicates the cell type where a TF has the highest motif activity. **f**, Examples of correlated TF expression and TF motif activity. **g**, Number of target genes per TF, shown separately for (predicted) self-regulating TFs (median = 196) and not-self-regulating TFs (median = 66). **h**, Number of TF regulators per target gene (median = 3). **i**, Inferred GRN for cnidocytes in adult *Nematostella*. TF nodes are coloured by expression, scaled by TF motif activity and labelled; target genes are indicated by small grey dots. Width and transparency of connections represents interaction strength (in silico ChIP binding score). Asterisks highlight TFs known to be involved in cnidocyte development. **j**, Network centrality metrics for the top TFs in the inferred cnidocyte GRN.

cell types based on Euclidean distances between gene accessibility profiles (gene scores; Fig. 4a), which we expected to largely reflect shared effector gene usage, similar to gene expression. This analysis revealed that functionally related cell types tended to cluster together, for example, adult muscle cell types (fast-contracting retractor muscles, and slow-contracting parietal and circular muscles), as well as a group composed of neurosecretory cells (cnidocytes, neurons and gland/secretory cells) alongside epidermal cells and NPCs.

In contrast, clustering based on the overlap of accessible CREs (Fig. 4b) resulted in a different grouping: TR muscle cells clustered with epidermal cells and adult NPCs, whereas the remaining muscle cell types grouped together with gastrula EMS cells. A similar pattern, consistent with known ontogenetic relationships in *Nematostella*[40,65], was observed when we compared cells based on *cis*-regulatory sequence similarity, using area under the curve (AUC) values derived from gkm-SVM classifiers performance across cell types (Fig. 4c). This analysis revealed the strongest cross-stage associations. For instance, gastrula ectodermally derived adult cell types such as epidermis, NPCs, cnidocytes[27] and TR muscle[65], along with Pou4/FoxL2-expressing neurons. Separately,

https://doi.org/10.1038/s41559-025-02906-1

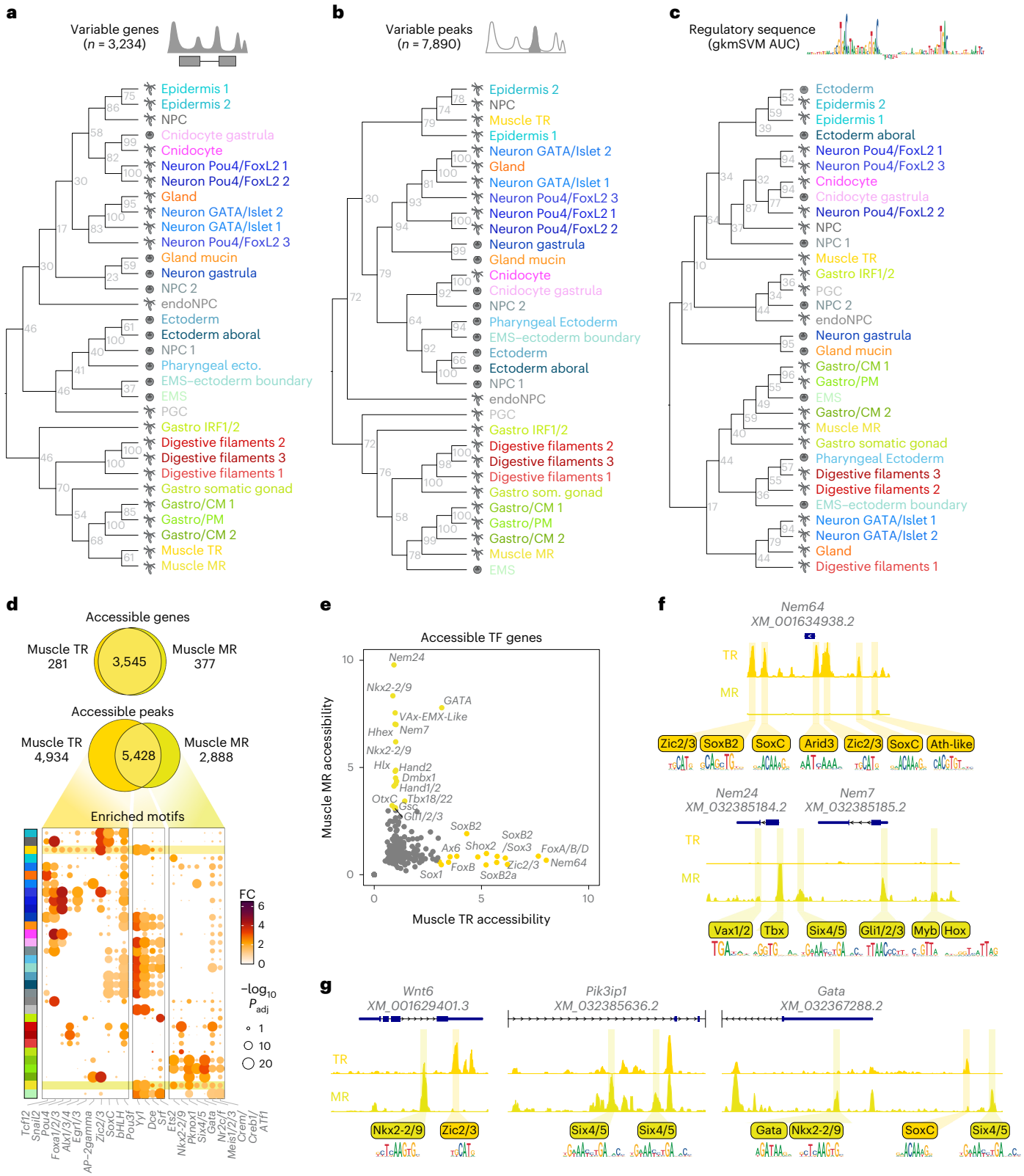


Fig. 4 | Comparing *Nematostella* cell type regulatory identities. a, NJ cell type tree based on accessibility scores for 3,234 variable genes. Node labels indicate bootstrap support values calculated by resampling genes and recalculating pairwise distances across 100 bootstrap iterations. **b**, NJ cell type tree based on shared accessibility of 7,980 variable peaks. Node labels indicate bootstrap support values calculated by resampling peaks and recalculating peak overlaps across 100 bootstrap iterations. **c**, NJ cell type tree based on regulatory sequence similarity, based on AUC values obtained applying cell type gkm-SVM classifiers between cell types. Node labels indicate bootstrap support values calculated by resampling test-set peaks per model and recalculating AUC across 100 bootstrap iterations. **d**, Euler diagram showing the overlap of genes (top, based on gene scores) and peaks (bottom) accessible in two retractor muscles. The dotmap below shows the top enriched motifs in each group of peaks (hypergeometric test, FDR adjustment for multiple testing). **e**, TF accessibility in TR and MR muscle cells (in all cell type peaks, not a subset shown in **d**). **f**, CRE accessibility around key TFs in TR and MR muscle cells. **g**, Examples of genes with shared accessibility but different set of accessible peaks in two retractor muscles. Specific instances of TF binding sites are shown below the coverage tracks.

gastrula EMS cells clustered with endomesodermally derived adult muscle types, including MR muscle, circular and parietal muscles[65]. Another cross-stage association reflecting known ontogenetic relationships included gastrula pharyngeal ectoderm cells clustering with digestive filaments[40]. Adult gland/secretory cells and GATA/Islet positive neurons formed a distinct cluster that was more similar to the group of EMS and pharyngeal derivatives than to Pou4/FoxL2-positive neurons and cnidocytes. This may suggest the existence of developmentally distinct populations of enteric and ectodermal/epidermal neurons in *Nematostella*[39,66].

The distinct affinities of TR and MR muscle cell types have implications for the evolution of their expression programs. TR muscle has been proposed to arise from ectodermal progenitors through co-option of the MR muscle program—a process thought to be mediated by the emergence of the *Nem64* paralog[29]. Cole et al.[29] identified TR and MR muscles as transcriptionally similar cell types that differ in their regulation by bHLH TFs: TR muscle by *Nem64*, and MR muscle by *Nem7* and *Nem24*. To further investigate this co-option process, we compared TR–MR similarity at three levels: genes, CREs and regulatory sequences. As expected from their high transcriptional similarity, the overall gene accessibility scores for TR and MR muscles were largely overlapping (Fig. 4d), yet the two muscles have different accessible TF (Fig. 4e), including the *Nem* bHLH TFs (Fig. 4f). Likewise, there was far less overlap for TR and MR muscles at the level of individual CREs, and these non-overlapping CREs harbour distinct TF motif signatures (Fig. 4d,g), overall suggesting highly distinct regulatory states. These motif differences help explain the divergence in *cis*-regulatory sequences between TR and MR muscles and their different germ layer origin[65]. For example, among the motifs enriched in TR-specific CREs we find Tcf12, SoxC and Pou4—motifs also associated with NPCs and/or its ectodermal derivatives—whereas MR-specific CREs are enriched for motifs such as Six4/5, Nkx2 and GATA, shared with EMS and/or its derivatives (Fig. 4d). Many of these motifs are present in the CREs linked to *Nem64*, *Nem7* and *Nem24* (Fig. 4f).

Together, these results indicate that, although TR and MR muscles rely on a largely shared repertoire of effector genes, they are governed by distinct CRE landscapes that interpret different upstream regulatory inputs during differentiation. Even genes with shared accessibility between the two retractor muscles, show different accessible CREs and TF binding motif occurrences between TR and MR muscles (Fig. 4g). This suggests that the co-option of the MR muscle program into a ectodermal lineage involved more than the recruitment of a paralogous terminal selector (*Nem64*). It also required thousands of TR-specific CREs capable of activating shared muscle genes in ontogenetically distinct progenitors—ectodermal for TR muscle and EMS for MR muscle—potentially by establishing a permissive chromatin landscape compatible with their respective developmental origins and/or redundantly reinforcing the activation of muscle genes by *Nem* bHLH TFs.

Discussion

Here we present a whole-organism single-cell chromatin accessibility atlas for the cnidarian *N. vectensis*. This atlas allowed us to dissect the regulatory logic underlying cell type-specific gene expression in cnidarians. We identified 112,728 CREs across the 269 Mb *Nematostella* genome, including 91,362 putative enhancers (that is, non-promoter CRE). This number substantially exceeds previous estimates and approaches the number of CREs reported in *Drosophila*, which has a similar genome size (180 Mb).

We identified key TFs associated with each cell identity by analysing their expression, aggregated motif accessibility and regulatory influence. In parallel, we defined the *cis*-regulatory motif grammars that characterize cell type-specific CREs. By integrating TF activity with CRE accessibility and motif composition, we inferred GRN models for main *Nematostella* cell types, enabling systematic analysis of GRN structure and composition. These analyses reveal the intricate regulatory logic that governs cell type-specific gene programs in a morphologically simple, non-bilaterian animal and provide a framework to dissect the conserved and lineage-specific regulatory networks that enabled the emergence of new cell types (for example, cnidocytes) in future comparative studies. A key limitation for both GRN reconstruction and the interpretation of CRE sequence models is our incomplete knowledge of TF binding preferences in *Nematostella*. Although protein sequence conservation can, in some cases, be used to transfer experimentally defined motifs from other species[59,67], a substantial fraction of *Nematostella* TFs lack predictable binding motifs and will therefore require direct experimental characterization, for example, by SELEX or related assays[68–71].

Our findings further show that although effector gene usage groups functionally similar cell types, regulatory features can reveal ontogenetic relationships between cell types[72]. For instance, GATA/Islet-expressing neurons exhibit regulatory sequence similarities with EMS and pharyngeal derivatives, clearly distinguishing them from the ectodermally associated Pou4/FoxL2 neurons. This suggests a possible enteric origin for this broad class of neurons in *Nematostella*. This analysis also sheds light into the regulatory mechanisms underlying the convergent differentiation of fast-contracting muscles from distinct germ layers. Here the activation of a similar set of fast muscle effector genes occurs through largely distinct CREs and regulatory sequence information, even for the same target genes. This suggests that developmental homoplasy may not result merely from the duplication and redeployment of a terminal selector TF in a different germ layer. Instead, such convergent activation of effector programs also requires access to distinct regulatory states, such as those mediated by pioneer TFs that establish CRE accessibility in distinct progenitor populations. Mapping single-cell chromatin dynamics through development will be essential for resolving the TF hierarchies and sequential CRE activation events that underpin the deployment of these convergent cell type programs. Furthermore, comparative analyses across closely related anthozoan species could reveal the evolutionary flipside of this developmental co-option, shedding light on how these distinct muscle CRE landscapes evolved.

Our *cis*-regulatory atlas moves beyond conventional transcriptome-based cell type characterization by analysing regulatory traits that define cell type identities in *Nematostella*, such as CREs sequence motif composition, active TFs and GRN architecture. We anticipate that applying similar approaches in other organisms will further advance our understanding of animal genome regulation and serve as a powerful tool for resolving cell type ontogenetic and evolutionary relationships.

Methods

N. vectensis culture

The *N. vectensis* culture is derived from CH2 males and CH6 females[73]. Adult polyps were maintained at 18 °C in filtered seawater diluted 1:3 (*Nematostella* medium (NM)), and spawned by a temperature and light shock[74]. Fertilized egg packages were treated with a 3% L-cysteine in NM solution to remove the egg jelly. Embryos were raised at 21 °C until midgastrula stage (26 hours post-fertilization (hpf)) and collected based on their morphology.

Sample preparation for single-cell experiments

Depending on the sample input and single-cell omics protocol, different approaches were used to obtain single-cell suspensions as described below.

Whole-gastrula scATAC-seq. Embryos were washed twice in cold PBS before nuclei isolation and permeabilization. Nuclei from 300 pooled gastrula were isolated in 300 µl 1× OmniATAC lysis buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA, 0.1% NP-40, 0.1% Tween-20, 0.01% digitonin)[75]. If nuclei were processed fresh, OmniATAC lysis buffer was supplemented with Pitstop2 (70 µM, Abcam, catalogue

number 120687) to increase nucleus permeability to Tn5 (ref. 76). Nuclei were isolated gently and permeabilized by Dounce homogenization and mechanical pipetting for a maximum of 3 min in cold conditions; 1.7 ml of cold ATAC wash buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 2% BSA, 0.1% Tween-20) was added and the nuclei filtered through a 40-µm strainer into a new 2-ml LoBind tube. Nuclei were pelleted at 500*g* in a swinging bucked rotor for 7 min at 4 °C. The resulting pellet was washed twice in cold PBS-1%BSA, gently resuspended in 1× diluted buffer (10x Genomics) and filtered through a 40-µm cell strainer (Flowmi). A step-by-step version of this protocol can be found at protocols.io[77].

If nuclei suspension was purified from debris and aggregates using fluorescence-activated cell sorting (FACS) (see below), nuclei were Dounce-homogenized in OmniATAC lysis buffer without digitonin and fixed mildly in 0.1% methanol-free paraformaldehyde (PFA) (ThermoFisher, catalogue number 28906) to mitigate nuclei damage during sorting. Briefly, after washing in ATAC wash buffer, nuclei were incubated in PBS-1% BSA for 5 min on ice, gently resuspended and fixed for 5 min at room temperature by adding 1% PFA in PBS to reach a final concentration of 0.1% PFA. The reaction was quenched by adding glycine (0.125 M final concentration), Tris-HCl pH 8 (50 mM final concentration) and BSA (1.7% final concentration) and left for 5 min at 4 °C. Nuclei were pelleted at 500*g*, 5 min at 4 °C, and washed once with cold PBS-1% BSA. The resulting pellet was resuspended gently and stained in PBS-1% BSA with 4′;,6-diamidino-2-phenylindole (DAPI; 10 µg ml^{-1} final concentration) before FACS. Single nuclei (1 million; 2*n* and 4*n* DNA content) were sorted using FACS Influx (100 µm nozzle, 12 psi, cold conditions) into PBS-1% BSA. Nuclei were pelleted and permeabilized in 0.1× OmniATAC lysis buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA, 0.01% NP-40, 0.01% Tween-20, 0.001% digitonin) supplemented with 70 µM Pitstop2 for 2 min on ice while gently pipetting. After washing, nuclei were processed as described above for fresh nuclei (sample name: 2_Gastrula_fix).

Before each Chromium scATAC-seq run (10x Genomics), an aliquot of nuclei suspension was taken to assess their quality and concentration. For this, nuclei were stained with DAPI and loaded on a Neubauer chamber for counting under a fluorescence microscope. Nuclei concentration was adjusted to encapsulate ~10,000 nuclei from each sample with the 10x Chromium platform after tagmentation in bulk. scATAC-seq libraries from gastrula stage were prepared using the Chromium scATAC v.2 (Next GEM) kit from 10x Genomics, following the manufacturer's instructions.

Whole-adult scATAC-seq. *Nematostella* polyps (two- to three-months old) were obtained from non-sexed wild-type polyps, starved for at least 3 days, and spawned 1 day before dissociation to avoid any possible contamination with gametes. Two to four adult polyps were washed in PBS before plunging them into ice-cold TST lysis buffer (10 mM Tris-HCl pH 7.5, 146 mM NaCl, 1 mM $CaCl_2$, 21 mM $MgCl_2$, 0.03% Tween-20, 1× complete protease inhibitor)[78]. Polyps were transferred on a clean slide on ice and minced with a pre-chilled knife into small chunks. Chopped tissue was then crushed gently in an ice-cold Dounce homogenizer until homogenous suspension was achieved, and further dissociated by pipetting (Gilson Pipetman, p1000 strokes). Sample was maximum 12 min in TST lysis buffer, then diluted with 1 volume of cold 2% BSA in ST buffer (without Tween-20). Resulting cell/nuclei suspension was filtered through a 70-µm strainer into LoBind protein tube and pelleted at 800*g* for 5 min at 4 °C. To purify single nuclei from debris and aggregates, sample was fixed mildly in 0.1% PFA before FACS as described above. Between 700,000 and 1 million single nuclei were sorted into PBS-2% BSA, pelleted and permeabilized for 2 min in cold 0.1× OmniATAC lysis buffer with Pitstop2. A step-by-step version of this protocol can be found at protocols.io[79].

When nuclei from adult samples were processed fresh (without PFA fixation), NP-40 was added to TST lysis buffer (0.01% NP-40 final concentration) after Dounce homogenization and further dissociated for 5 min by pipetting. In this case, nuclei were purified from debris using an OptiPrep continuous density gradient. Fresh purified nuclei were permeabilized in ice-cold 1× OmniATAC lysis buffer with Pitstop2 for 4 min while pipetting gently. Finally, fresh or fixed and permeabilized nuclei were washed in ATAC wash buffer, resuspended in 1× diluted buffer and filtered through a 40-µm strainer (Flowmi) before counting.

A total of 16 scATAC-seq libraries were generated from adult fixed samples (sample name: 3–15 and 17–19 Adult_Fix) and 1 scATAC-seq library from fresh sample (16_Adult_Fresh). All of them using the Chromium scATAC v.1.1 (Next GEM) kit from 10× Genomics and following the manufacturer's instructions.

scATAC-seq of *NvElav1::mOrange*-positive cells. To enrich our adult scATAC-seq dataset with neural cells, *NvElav1::mOrange*-positive cells were purified by FACS as described previously[80], with minor modifications. Briefly, 1-month-old *NvElav1::mOrange*-positive polyps were dissociated at 25 °C in calcium- and magnesium-free NM (CMF/NM) containing 5 mM EDTA and 0.25% α-chymotrypsin (Sigma, catalogue number C4129). Single-cell suspensions were then stained with Hoechst 33342 (1 µg ml^{-1}, ThermoFisher, catalogue number 62249) and TO-PRO-3 (50 nM, Invitrogen, catalogue T3605) to remove debris and nonviable cells by FACS (FACS Aria II, 100-µm nozzle, cold conditions). Nuclei from 150,000 sorted mOrange-positive cells were isolated, fixed mildly in 0.1% PFA and permeabilized (samples: 20 and 21 Elav_fix) or permeabilized directly in OmniATAC lysis buffer with Pitstop2 (samples: 22–24 Elav_fresh). Finally, permeabilized nuclei were washed in ATAC wash buffer, resuspended in 1× diluted buffer and encapsulated using the 10x Chromium platform. Five scATAC-seq libraries were generated using the Chromium scATAC v.1.1 (Next GEM) kit from 10x Genomics, following the manufacturer's instructions.

Whole-adult scMultiome (ATAC+RNA). Two adult wild-type polyps were dissociated and stained for FACSorting as described above for *NvElav1::mOrange* samples. Nuclei from 250,000 single viable cells were isolated and permeabilized for 3 min in ice-cold 0.1× OmniATAC lysis buffer supplemented with Pitstop (70 µM), RNAse inhibitor (1 U $µl^{-1}$) and dithiothreitol (1 mM). Nuclei were then washed in ATAC wash buffer and resuspended in 1× diluted nuclei buffer, both supplemented with RNAse inhibitor (1 U $µl^{-1}$) and dithiothreitol (1 mM). Nuclei were counted after filtering through a 40-µm strainer with the help of DAPI, and encapsulated using the 10x Chromium platform.

One run of Chromium Next GEM single-cell multiome kit from 10x Genomics was performed, following the manufacturer's instructions and performing eight PCR cycles in Step 5.1 for scATAC library construction (24_Adult_Fresh_MultiomeATAC), or nine PCR cycles of cDNA amplification in Step 6.1 for scRNA library construction (24_Adult_Fresh_MultiomeGE).

Whole-adult 5′ scRNA-seq. Single-cell suspensions were obtained after ACMEsorbitol (0.4 M) fixation and dissociation as described previously[8,81]. One single-cell 5′ GE library was generated using the Chromium Next GEM 5′ GEX scRNA-seq v.2 kit from 10x Genomics, following the manufacturer's instructions with 14 PCR cycles of cDNA amplification.

scATAC-seq libraries (Supplementary Table 1) were sequenced using a 50/8/16/50 sequencing format to reach ~5,000 reads per cell (average: 19,575), with median 2,788 unique fragments per cell on average. The scATAC-seq library derived from scMultiome kit (07564AAD) was sequenced using a 50/8/24/49 sequencing format at 3,929 reads per cell and 1,556 unique fragments per cell, whereas the scRNA-seq library (07563AAD) was sequenced using a 28/10/10/90 sequencing format at 8,080 reads per cell and median 789 unique molecular identifiers (UMIs) per cell. The 5′ scRNA-seq library (07575AAD) was sequenced using a 26/10/10/90 sequencing format at 24,297 reads per cell and

median 863 UMIs per cell. All libraries were sequenced using Illumina NextSeq500 platform.

scATAC-seq processing and cluster annotation

We processed scATAC sequencing data using a modified scATAC-pro workflow[82]. Briefly, we mapped sequencing reads to *Nematostella* Darwin Tree of Life genome[83] using bwa[84], and filtered nucleosome free reads for downstream analysis. Initial cell calling with was done using EmptyDrops[85], with false discovery rate 0.05. scATAC downstream analysis was done using ArchR[36]. Cells called with EmptyDrops that had TSS enrichment below four and fewer than 200 fragments were filtered out. We also added doublet scores using ArchR's in silico doublets method, and removed cells predicted to be doublets using filterRatio = 1 (4% of input cells). We then performed dimensionality reduction using iterative latent semantic indexing (four iterations) and clustering using top 10,000 variable features, with resolution set at 0.3. We identified and removed clusters of low-quality cells with TSS enrichment < 8. We then repeated dimensionality reduction and clustering iteratively until all resulting clusters were of good quality. Next we used SEACells[35] for grouping cells into metacells, with target of ~75 single cells per metacell. Metacells obtained from SEACells approach were grouped in clusters and annotated broadly by label transfer from scRNA-seq data using AUCell[86]. Briefly, AUCell calculates enrichment score for a given reference gene set (for example, scRNA-seq-derived cell type marker signatures) within ranked genes profile of a query cell or group of cells (for example, scATAC metacell). For each scATAC metacell, we computed AUC scores for all reference broad cell type signatures in scRNA, and annotated the metacell as the highest-scoring broad cell type. In cases where broad AUCell-based annotations were not sufficient to resolve more specific subtypes (for example, muscle, gastrodermis or progenitor subpopulations), we assigned more specific cell type annotations by inspecting the accessibility (gene scores, described below) of known marker genes. To validate our annotations, we compared correlations between gene scores and expression in matched 10x multiome cells (scATAC + scRNA-seq) and in unmatched RNA/ATAC data linked by annotation transfer. Correlations were similar in both cases (Extended Data Fig. 1l), supporting the accuracy of our strategy.

We then aggregated metacells into pseudobulk cell types and generated final consensus set of peaks using MACS2 (ref. 87) and iterative reduction approach implemented in ArchR. Differential peaks per cell type were determined as those with Log_2fold change (FC) ≥ 1 and false discovery rate (FDR) ≤ 0.1 when compared to peaks in other cell types using Wilcoxon test and FDR *P* value adjustment, and accounting for TSS enrichment and log_{10}(number of fragments) bias. Up to this point, the adult and gastrula datasets were analysed independently. Next, to integrate the two datasets, we overlapped gastrula and adult peaks to construct a reference peak set (union of all peaks). We then constructed a combined peak-by-cell count matrix for both stages. Counts were quantile-normalized and aggregated at both the metacell and cell type level. Aggregated accessibility profiles were used for hierarchical clustering, NJ tree construction, and dimensionality reduction with uniform manifold approximation and projection (UMAP) to visualize metacells and types (Extended Data Fig. 1d–h).

Peaks to gene assignment and gene score calculation

To each gene we assigned peaks that are within the gene's body or <10 kb away from the gene's TSS, unless they were coming after (upstream or downstream) a TSS of another gene (implemented in mta_match_peaks_to_genes() function). Initially, 52,526 (63%) peaks were assigned to a single gene, and 31,098 (37%) peaks were assigned to more than one gene. For the latter, we refined the assignment by taking into account peaks co-accessibility (calculated by Cicero) and the correlation of accessibility to gene expression. Briefly, for all co-accessible peaks groups (co-accessibility > 0.5) assigned to more than one gene, we looked for a sharp drop in ranked peak-to-gene correlation (Δcorrelation < −0.1) for all peaks in the group, and removed those assignments that followed the drop (this procedure is implemented in mta_refine_peaks_to_genes_by_coaccessibility() function). As a result, we refined the assignment of 2,142 peaks. Next, we calculated gene scores as a weighted sum of the accessibility of all peaks assigned to gene. Each peak is weighted by distance from the gene (peaks inside the gene body get maximum weight of 1) and by peak specificity, measured by Gini index (Fig. 1d). This procedure is implemented in mta_gene_scores() function. Using 5′ scRNA-seq and H3K4me3 data together with scATAC peaks, we devised a decision tree approach (Extended Data Fig. 3a) to assign promoters to genes, and further classify them as CPs, which are accessible in all cell types, SPs accessible in one or several cell types, but not all, and potential APs, with different promoters being used in different cell types (this is implemented in mta_class_promoters() function).

Motif archetypes

We aimed to collect a comprehensive catalogue of all possible TF binding motifs in *Nematostella* genome. To this end, we combined motifs for *Nematostella* TFs that were either determined experimentally or inferred from other species based on TFs' DNA-binding domain (DBD) sequence similarity, with motifs we found to be significantly enriched or depleted in either all accessible or specifically accessible peaks in cell types, or enriched in different promoter classes (AP, SP, CP). To reduce redundancy of this comprehensive catalog of motifs, we calculated pairwise similarities between position weight matrices (PWMs) using compare_motifs() function from universalmotif R package (Pearson correlation coefficient (PCC) with normalize.score option to favour alignments that leave fewer unaligned positions, as well as alignments between motifs of similar length), and then we applied complete hierarchical clustering, choosing the number of clusters that maximizes the ratio of within- and between-cluster median pairwise similarities. These initial clusters of similar motifs were further split into smaller clusters that contain only motifs above a desired similarity threshold (0.8). For all the motifs in each cluster we applied information content (IC) block filtering[88], retaining only motifs with a block of at least four consecutive bases with IC ≥ 0.5 (ungapped motif), or at least two blocks of at least three consecutive bases with IC ≥ 0.5 (gapped motif). Then we generated a consensus PWM by averaging aligned PWMs at each position. Finally, we trimmed off the leading and trailing positions with IC < 0.5 in the consensus archetype motif. This entire procedure is implemented in the mta_merge_archetype() function. By doing this, we reduced the filtered input set of 2,951 motifs to 1,292 archetypes (Extended Data Fig. 6). We show that minimum–maximum normalized motif scores in accessible peaks are comparable for archetypes and highest-scoring motifs in each archetyping cluster, as well as the motif enrichments in cell type-specific peaks (Extended Data Fig. 6g). Motifs scores in peaks were computed by first calculating genome-wide motif alignment scores and retaining only those above 98th percentile of the genome-wide score distribution. This procedure was implemented in mta_gw_motif_score_monalisa() function, using findMotifHits() function from monaLisa R package. Motif scores were used to calculate motif enrichment in the set of cell type-specific peaks, with other peaks as the background; *p* value of enrichment was calculated using hypergeometric test, followed by FDR correction. This is implemented in mta_motif_enrichment_test() function.

Assigning motifs to TFs

Binding motifs have been determined experimentally only for a subset of *Nematostella* TFs[59]. We devised a computational approach to assign a motif from our comprehensive set of motif archetypes (hereafter, motifs) to each TF gene without experimentally determined motif (Extended Data Fig. 6l). We first calculated motif activity scores for all motifs using chromVAR[60]. Next, we calculated correlations between each motif's activity score and both expression and accessibility (gene score) of each TF. We ranked motifs based on gene score correlation

and for each gene we selected the best-correlated motif of the same structural class, if either expression or gene score correlation was greater than 0.3. To improve the accuracy of assignment, particularly for large structural classes such as Homeodomains, we also considered closest human, mouse, rat and zebrafish orthologs of each TF, and, if the motif activity of ortholog gene's motif correlated better than that of previously selected archetype, we assigned that ortholog motif to a given *Nematostella* TF (Extended Data Fig. 6m,n).

Gene regulatory network inference

We used an in silico ChIP method[61] to link TFs to target scATAC peaks. Briefly, in silico ChIP links TFs to a peak, if the peak contains a motif hit for the TF and if the accessibility of the peak correlates with the RNA expression of the TF. Correlation between peak accessibility and RNA expression at metacell level was calculated after mapping each scATAC metacell to the best-correlated scRNA metacell of the same broad cell type. Motif hits were determined using findMotifHits() function from monaLisa R package[89] with 95th quantile of genome-wide motif scores distribution for each motif used as a minimum score for counting a hit. In silico ChIP outputs a matrix of TF binding scores for each peak, ranging from 0 to 1, and it is necessary to select a threshold value for '+'. For each motif, we calculated its cell type activity as a *Z*-score of accessibility deviation of the target peaks set (selected with different in silico ChIP binding score thresholds) in a given cell type, compared to assumption of equal chromatin accessibility across cell types, and normalized by a set of background peaks matched for GC and average accessibility. From this, we selected 0.1 as a binding score cut-off because this was the value that maximized the correlation of TF expression and TF activity for most TFs.

TFs and target peaks for which binding score is greater than 0.1 constitute a global GRN. We further partitioned this into cell type-specific GRNs by filtering TFs based on expression and TF activity, and filtering target genes based on expression and accessibility. We used a per cell type 0.4 quantile threshold of expression FC to filter genes (both TFs and target genes) by expression. To filter peaks, we used a per cell type 0.4 quantile threshold of normalized peak accessibility. To filter TFs based on activity, we used a *Z*-score threshold of 4. For plotting GRNs (Fig. 3i and Extended Data Fig. 8), we also filtered out genes with expression FC < 1.2.

Sequence models

To learn the sequence determinants of CREs, we first used gapped kmer support vector machine (gkm-SVM) classifiers[90]. gkm-SVM represents each DNA sequence by short words (k-mers) that can contain gaps, thereby capturing both exact and flexible sequence patterns. We trained per cell type classifiers on the set of accessible peaks in each cell type ($\log_2$FC > 1 and *P* value < 0.1), with a 70–30 train–test split and fivefold cross-validation. Of note, we used a relaxed threshold for selecting specific peaks, because we wanted the models to learn more general CRE sequence features that might be shared across similar cell types. We applied each model to all left-out sets of peaks and calculated test-set AUC statistics. We used gkmexplain[55] on the top 1,000 scored peaks per model to identify important sequence features for each cell type classifier. Reasoning that the deep learning models may be better suited for identifying complex sequence grammar than the classic machine learning kmer classifiers, we next trained two types of deep learning model on *Nematostella* chromatin accessibility data: ChromBPNet and CREsted. ChromBPNet[53] (v.1.5) is a fully convolutional neural network that predicts accessibility at base pair resolution from underlying CRE sequences, after directly removing the known Tn5 bias. For this, we first train a bias models to learn Tn5 sequence biases that will be regressed out during the subsequent models training. Then we trained individual regression models to predict total counts and accessibility profile signals for each cell type using cell type accessible peak and GC-matched nonpeak sequences as inputs. The models were trained using default architecture with four dilation layers and 512 filters, on 500-bp sequences as input, with prediction on 250 bp. In addition, we also trained the peak regression CREsted model (https://github.com/aertslab/CREsted), which adapts the original ChromBPNet architecture but jointly learns accessibility prediction across all cell types. We reasoned that this information sharing may benefit the model and lead to more efficient motif discovery approach then training individual models for each cell type, as is the case for ChromBPNet. CREsted model was trained on peak logcounts, normalized by subtracting mean value across class. The model was trained using the default architecture, with learning rate of 5×10^{-1} and Huber loss function, on 500-bp sequences as input, with prediction also on 500 bp. During training of both CREsted and chromBPNet models, we left out CREs on one chromosome for validation (NC_064034.1) and on another for testing (NC_064035.1). We used SHAP DeepExplainer[91] to estimate the predictive importance of each base in CRE sequence, and TF MoDISco-lite[91] to identify sequence patterns (motifs) that are relevant for accessibility prediction. As input for TF MoDISco, we selected the 5,000 most specific regions per class and, of those, used the 1,000 regions with the highest predictions scores for that class. We then used the same archetyping procedure as described above to reduce redundant patterns from different models, with the only difference being that here we used Jannson–Shannon divergence (JSD) as a metric of motif similarity. To compare pattern archetypes to known motif archetypes, we calculated JSD for every motif archetype-pattern archetype pair in two ways: along the entire length of motifs alignments ($JSD_{complete}$) and along only the overlapping fraction of alignment (JSD_{min})—in this way we could better distinguish new motifs from similar motifs in different contexts (Extended Data Fig. 6h).

Generation of *Nematostella* transgenic lines

NvGabrb4::mOrange transgenic reporter lines driven by differentially accessible APs in TR muscle (TR-AP) or neuron Pou4/FoxL2 (NeuroPou4/FoxL2-AP) cells were generated by meganuclease-mediated transgenesis as described by Renfer and Technau[92].

The genomic coordinates for the roughly 2.8-kb regulatory region of tRM-AP are 11660621–11657766 on minus strand chr. 2. The genomic coordinates for the roughly 2-kb regulatory region of NeuroPou4/FoxL2-AP are 11644315–11642257 on the same minus strand of chr. 2 (ref. 83). These regulatory regions were cloned in frame with mOrange reporter gene into the meganuclease (I-Sce1)-mediated transgenesis vector kindly provided by the Technau laboratory[92]. Wild-type fertilized eggs were injected with a mix containing: plasmid DNA (20 ng μl^{-1}), I-Sce1 (1 U μl^{-1}, NEB, catalogue number R0694), Dextran Alexa Fluor 488 (50 ng μl^{-1}, Life Technologies, catalogue number D22910) and CutSmart buffer (1×). The mix was incubated at 37 °C for at least 20 min, then injection was performed at 18 °C with a FemtoJet 4i microinjector (Eppendorf). Constructs and/or transgenic lines are available from the authors upon request.

Immunofluorescence

One-month-old F1 polyps derived from *NeuroPou4/FoxL2-AP::mOrange* transgenic line were relaxed in 0.34% $MgCl_2$/NM solution to prevent tentacle contraction before cutting with a sharp knife at the level of the pharynx. The resulting heads were fixed in 3.7% formaldehyde in PBS-0.1% Tween-20 (PBTw) overnight at 4 °C, washed several times in PBTw the day after and left in PBS overnight at 4 °C.

For immunostaining against mOrange, samples were washed several times in PBS-0.3%TritonX (PBTx) for 1 h at room temperature, blocked in blocking solution (1% BSA/5% normal goat serum/PBTx) for 1 h at room temperature, and incubated with rabbit anti-DsRed primary antibody (1:100, Clontech, catalogue number 632496) in blocking solution overnight at 4 °C. Samples were then washed several times in PBTx-BSA for 2 h at room temperature, blocked in blocking solution for 30 min at room temperature and incubated with goat anti-rabbit

Alexa568 secondary antibody (1:250, Life Technologies, catalogue number A11011) in blocking solution overnight at 4 °C. Samples were then washed five times in PBTx-BSA for 2 h at room temperature, 5 min in PBS, and left in 70% glycerol in PBS at 4 °C for at least overnight. Samples were mounted in ProLong Glass antifade mountant (ThermoFisher Scientific, catalogue number P36982) and imaged on a Leica SP8 confocal microscope. Images were extracted from *Z*-stacks with Fiji and adjusted for brightness/contrast applied to the whole image.

Live imaging

Adult F1 polyps derived from *TR-AP::mOrange* transgenic line were relaxed in 0.34% $MgCl_2$/NM solution before cutting with a sharp knife at the level of the pharynx. The resulting heads were then mounted in a slide with 2.43% $MgCl_2$/NM solution for live imaging on a Leica SP8 confocal microscope and images extracted as described above.

Reporting summary

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

Data availability

Raw and processed files will be available in GEO repository under accession number GEO: GSE294388. In addition, the atlas can be explored in an interactive database: https://sebelab.crg.eu/nematostella-cis-regulatory-atlas/ and also in an interactive genome browser: https://sebelab.crg.eu/nematostella-cis-reg-jb2/.

Code availability

Scripts to reproduce the data processing and downstream analysis are available via Zenodo at https://doi.org/10.5281/zenodo.17425383 (ref. 93). Unless otherwise specified, scripts are based on R v.4.2.2 and Python v.3.8.10, and the language-specific libraries specified in Methods.

References

1. Musser, J. M. et al. Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. *Science* **374**, 717–723 (2021).
2. Sebé-Pedrós, A. et al. Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-seq. *Cell* **173**, 1520–1534 (2018).
3. Sebé-Pedrós, A. et al. Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat. Ecol. Evol.* **2**, 1176–1188 (2018).
4. Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M. & Reddien, P. W. Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science* **360**, eaaq1736 (2018).
5. Plass, M. et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **1723**, eaaq1723 (2018).
6. Cao, J. et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
7. Levy, S. et al. A stony coral cell atlas illuminates the molecular and cellular basis of coral symbiosis, calcification, and immunity. *Cell* **184**, 2973–2987.e18 (2021).
8. Najle, S. R. et al. Stepwise emergence of the neuronal gene expression program in early animal evolution. *Cell* **186**, 4676–4693 (2023).
9. Tanay, A. & Sebé-Pedrós, A. Evolutionary cell type mapping with single-cell genomics. *Trends Genet.* **37**, 919–932 (2021).
10. Janssens, J. et al. Decoding gene regulation in the fly brain. *Nature* **601**, 630–636 (2022).
11. Cusanovich, D. A. et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
12. Calderon, D. et al. The continuum of *Drosophila* embryonic development at single-cell resolution. *Science* **377**, eabn5800 (2022).
13. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
14. Sarropoulos, I. et al. Developmental and evolutionary dynamics of *cis*-regulatory elements in mouse cerebellar cells. *Science* **373**, eabg4696 (2021).
15. Li, Y. E. et al. A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* **382**, eadf7044 (2023).
16. Zhang, K. et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001 (2021).
17. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324 (2018).
18. Minnoye, L. et al. Chromatin accessibility profiling methods. *Nat. Rev. Methods Prim.* **1**, 10 (2021).
19. Hecker, N. et al. Enhancer-driven cell type comparison reveals similarities between the mammalian and bird pallium. *Science* **387**, eadp3957 (2025).
20. Parker, J. & Pennell, M. The cellular substrate of evolutionary novelty. *Curr. Biol.* **35**, R626–R637 (2025).
21. Schwaiger, M. et al. Evolutionary conservation of the eumetazoan gene regulatory landscape. *Genome Res.* **24**, 639–650 (2014).
22. Kim, I. V. et al. Chromatin loops are an ancestral hallmark of the animal regulatory genome. *Nature* **642**, 1097–1105 (2025).
23. Chari, T., et al. Whole-animal multiplexed single-cell RNA-seq reveals transcriptional shifts across *Clytia medusa* cell types. *Sci. Adv.* **7**, eabh1683 (2021).
24. Siebert, S. et al. Stem cell differentiation trajectories in Hydra resolved at single-cell resolution. *Science* **365**, eaav9314 (2019).
25. Li, Y. et al. Single-cell transcriptomic analyses reveal the cellular and genetic basis of aquatic locomotion in scyphozoan jellyfish. Preprint at *bioRxiv* https://doi.org/10.1101/2023.02.06.527379 (2023).
26. Hu, M., Zheng, X., Fan, C.-M. & Zheng, Y. Lineage dynamics of the endosymbiotic cell type in the soft coral Xenia. *Nature* **582**, 534–538 (2020).
27. Steger, J. et al. Single-cell transcriptomics identifies conserved regulators of neuroglandular lineages. *Cell Rep.* **40**, 111370 (2022).
28. Steinmetz, P. R. H. et al. Independent evolution of striated muscles in cnidarians and bilaterians. *Nature* **487**, 231–234 (2012).
29. Cole, A. G. et al. Muscle cell-type diversification is driven by bHLH transcription factor expansion and extensive effector gene duplications. *Nat. Commun.* **14**, 1747 (2023).
30. Denner, A. et al. *Nanos2* marks precursors of somatic lineages and is required for germline formation in the sea anemone *Nematostella vectensis*. *Sci. Adv.* **10**, eado0424 (2024).
31. Miramón-Puértolas, P., Pascual-Carreras, E. & Steinmetz, P. R. H. A population of Vasa2 and Piwi1 expressing cells generates germ cells and neurons in a sea anemone. *Nat. Commun.* **15**, 8765 (2024).
32. Babonis, L. S. & Martindale, M. Q. Old cell, new trick? Cnidocytes as a model for the evolution of novelty. *Integr. Comp. Biol.* **54**, 714–722 (2014).
33. Babonis, L. S., Enjolras, C., Ryan, J. F. & Martindale, M. Q. A novel regulatory gene promotes novel cell fate by suppressing ancestral fate in the sea anemone Nematostella vectensis. *Proc. Natl Acad. Sci. USA* **119**, e2113701119 (2022).
34. Babonis, L. S. et al. Single-cell atavism reveals an ancient mechanism of cell type diversification in a sea anemone. *Nat. Commun.* **14**, 885 (2023).
35. Persad, S., et al. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.* **41**, 1746–1757 (2023).

36. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 935 (2021).
37. Cole, A. G. et al. Updated single cell reference atlas for the starlet anemone *Nematostella vectensis*. *Front Zool.* **21**, 8 (2024).
38. Richards, G. S. & Rentzsch, F. Regulation of *Nematostella* neural progenitors by *SoxB*, *Notch* and *bHLH* genes. *Development* **142**, 3332–3342 (2015).
39. Lemaître, Q. I. B. et al. NvPrdm14d-expressing neural progenitor cells contribute to non-ectodermal neurogenesis in *Nematostella vectensis*. *Nat. Commun.* **14**, 4854 (2023).
40. Steinmetz, P. R. H., Aman, A., Kraus, J. E. M. & Technau, U. Gut-like ectodermal tissue in a sea anemone challenges germ layer homology. *Nat. Ecol. Evol.* **1**, 1535–1542 (2017).
41. Rentzsch, F., Fritzenwanker, J. H., Scholz, C. B. & Technau, U. FGF signalling controls formation of the apical sensory organ in the cnidarian *Nematostella vectensis*. *Development* **135**, 1761–1769 (2008).
42. Haillot, E. et al. Segregation of endoderm and mesoderm germ layer identities in the diploblast Nematostella vectensis. *Nat. Commun.* **16**, 7979 (2025).
43. Lebedeva, T. et al. Cnidarian-bilaterian comparison reveals the ancestral regulatory logic of the β-catenin dependent axial patterning. *Nat. Commun.* **12**, 4032 (2021).
44. Reddington, J. P. et al. Lineage-resolved enhancer and promoter usage during a time course of embryogenesis. *Dev. Cell* **55**, 648–664 (2020).
45. Zhu, F. et al. The interaction landscape between transcription factors and the nucleosome. *Nature* **562**, 76–81 (2018).
46. Bulyk, M. L., Drouin, J., Harrison, M. M., Taipale, J. & Zaret, K. S. Pioneer factors—key regulators of chromatin and gene expression. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-023-00648-z (2023).
47. Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
48. Marlétaz, F. et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64–70 (2018).
49. Weintraub, A. S. et al. YY1 is a structural regulator of enhancer-promoter loops. *Cell* **171**, 1573–1588 (2017).
50. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* **13**, 233–245 (2012).
51. Haberle, V. & Lenhard, B. Promoter architectures and developmental gene regulation. *Semin. Cell Dev. Biol.* **57**, 11–23 (2016).
52. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput. Biol.* **10**, e1003711 (2014).
53. Pampari, A. et al. ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. Preprint at *bioRxiv* https://doi.org/10.1101/2024.12.25.630221 (2025).
54. De Winter, S., Konstantakos, V. & Aerts, S. Modelling and design of transcriptional enhancers. *Nat. Rev. Bioeng.* https://doi.org/10.1038/s44222-025-00280-y (2025).
55. Shrikumar, A., Prakash, E. & Kundaje, A. GkmExplain: fast and accurate interpretation of nonlinear gapped k-mer SVMs. *Bioinformatics* **35**, i173–i182 (2019).
56. Shrikumar, A. et al. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. Preprint at https://arxiv.org/abs/1811.00416v5 (2018).
57. Vierstra, J. et al. Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
58. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
59. Lambert, S. A. et al. Similarity regression predicts evolution of transcription factor sequence specificity. *Nat. Genet.* **51**, 981–989 (2019).
60. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).
61. Argelaguet, R. et al. Decoding gene regulation in the mouse embryo using single-cell multi-omics. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.15.496239 (2022).
62. Tournière, O. et al. NvPOU4/Brain3 functions as a terminal selector gene in the nervous system of the cnidarian *Nematostella vectensis*. *Cell Rep.* **30**, 4473–4489 (2020).
63. Babonis, L. S. & Martindale, M. Q. PaxA, but not PaxC, is required for cnidocyte development in the sea anemone *Nematostella vectensis*. *Evodevo* **8**, 14 (2017).
64. Danladi, B. et al. Conserved and lineage-restricted gene regulatory programs modulate developmental cnidocyte specification in *Nematostella vectensis*. Preprint at *bioRxiv* https://doi.org/10.1101/2025.05.08.652877 (2025)
65. Jahnel, S. M., Walzl, M. & Technau, U. Development and epithelial organisation of muscle cells in the sea anemone *Nematostella vectensis*. *Front. Zool.* **11**, 44 (2014).
66. Nakanishi, N., Renfer, E., Technau, U. & Rentzsch, F. Nervous systems of the sea anemone Nematostella vectensis are generated by ectoderm and endoderm and shaped by distinct mechanisms. *Development* **139**, 347–357 (2012).
67. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
68. Jolma, A. et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* **20**, 861–873 (2010).
69. Jolma, A. et al. DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339 (2013).
70. Jolma, A. et al. Perspectives on Codebook: sequence specificity of uncharacterized human transcription factors. Preprint at *bioRxiv* https://doi.org/10.1101/2024.11.11.622097 (2024).
71. Jolma, A. et al. GHT-SELEX demonstrates unexpectedly high intrinsic sequence specificity and complex DNA binding of many human transcription factors. Preprint at *bioRxiv* https://doi.org/10.1101/2024.11.11.618478 (2024).
72. Wang, M. et al. Distinct gene regulatory dynamics drive skeletogenic cell fate convergence during vertebrate embryogenesis. *Nat. Commun.* **16**, 2187 (2025).
73. Hand, C. & Uhlinger, K. R. The culture, sexual and asexual reproduction, and growth of the sea anemone *Nematostella vectensis*. *Biol. Bull.* **182**, 169–176 (1992).
74. Fritzenwanker, J. H. & Technau, U. Induction of gametogenesis in the basal cnidarian *Nematostella vectensis* (Anthozoa). *Dev. Genes Evol.* **212**, 99–103 (2002).
75. Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959 (2017).
76. De Rop, F. V., et al. Hydrop enables droplet-based single-cell ATAC-seq and single-cell RNA-seq using dissolvable hydrogel beads. *eLife* **11**, e73971 (2022).
77. Iglesias, M. Gastrula_Nvectensis_scATAC-seq, v.1, https://doi.org/10.17504/protocols.io.81wgbwrpogpk/v1 (2025).
78. Drokhlyansky, E., et al. The human and mouse enteric nervous system at single-cell resolution. *Cell* **182**, 1606–1622 (2020).

79. Iglesias, M. Adult_Nvectensis_scATAC-seq, v.1, https://doi.org/10.17504/protocols.io.261gek5mwg47/v1 (2025).
80. Torres-Méndez, A. et al. A novel protein domain in an ancestral splicing factor drove the evolution of neural microexons. *Nat. Ecol. Evol.* **3**, 691–701 (2019).
81. García-Castro, H. et al. ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics. *Genome Biol.* **22**, 89 (2021).
82. Yu, W., Uzun, Y., Zhu, Q., Chen, C. & Tan, K. scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. *Genome Biol.* **21**, 94 (2020).
83. Fletcher, C. & Pereira da Conceicoa, L. The genome sequence of the starlet sea anemone, *Nematostella vectensis* (Stephenson, 1935). *Wellcome Open Res.* **8**, 79 (2023).
84. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
85. Lun, A. T. L. et al. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
86. van den Oord, J. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
87. Zhang, Y. et al. Model-based analysis of ChIP–Seq (MACS). *Genome Biol.* **9**, R137 (2008).
88. Huber, B. R. & Bulyk, M. L. Meta-analysis discovery of tissue-specific DNA sequence motifs from mammalian gene expression data. *BMC Bioinform.* **7**, 229 (2006).
89. Machlab, D., et al. monaLisa: an R/Bioconductor package for identifying regulatory motifs. *Bioinformatics* **38**, 2624–2625 (2022).
90. Lee, D. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* **32**, 2196–2198 (2016).
91. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. Preprint at https://arxiv.org/abs/1704.02685v2 (2017).
92. Renfer, E. & Technau, U. Meganuclease-assisted generation of stable transgenics in the sea anemone *Nematostella vectensis*. *Nat. Protoc.* **12**, 1844–1854 (2017).
93. Elek, A. sebepedroslab/nvec-scatac: Nematostella_scATAC_atlas. *Zenodo* https://doi.org/10.5281/zenodo.17425383 (2025).

Acknowledgements

We thank I. Kim, A. de Mendoza, S. Montgomery, M. Irimia and N. Maeso for critical comments on the paper, as well as all members of the Sebe-Pedros group for discussion and suggestions. We thank F. Rentzsch for access to *Nematostella Elav1*::*mOrange* transgenic line. We are grateful to D. Cañas-Armenteros for taking care of *Nematostella* cultures and to the CRG Flow Cytometry, Genomics and ALMU facilities for technical support. Research in A.S-P. group has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 851647) and the Spanish Ministry of Science, Innovation and Universities (PID2021-124757NB-I00 funded by MICIU /AEI /10.13039/501100011033 / FEDER, UE). We acknowledge support of the Spanish Ministry of Science and Innovation through the Centro de Excelencia Severo Ochoa (CEX2020-001049-S, MCIN/AEI /10.13039/501100011033), the Generalitat de Catalunya through the CERCA program and to the EMBL partnership. A.E. was supported by FPI PhD fellowship from the Spanish Ministry of Science and Innovation (PRE2019-087793SO funded by MCIN/AEI/10.13039/501100011033 and FSE+). M.I. has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement number 75442. X.G-B. was supported by the European Union's H2020 research and innovation program under Marie Skłodowska-Curie grant agreement 101031767.

Author contributions

A.S.-P. conceived and supervized the study. M.I. performed single-cell experiments and generated transgenic reporter lines. A.E. analysed scATAC-seq data, performed motif analyses and trained sequence models with the support of L.M. and S.A. G.Z. and X.G.-B. performed phylogenetic and comparative genomics analyses. A.E. created visualizations. A.E., M.I. and A.S.-P. interpreted the data and wrote the paper with contributions from all authors.

Competing interests

The authors declare no competing interests.

Additional information

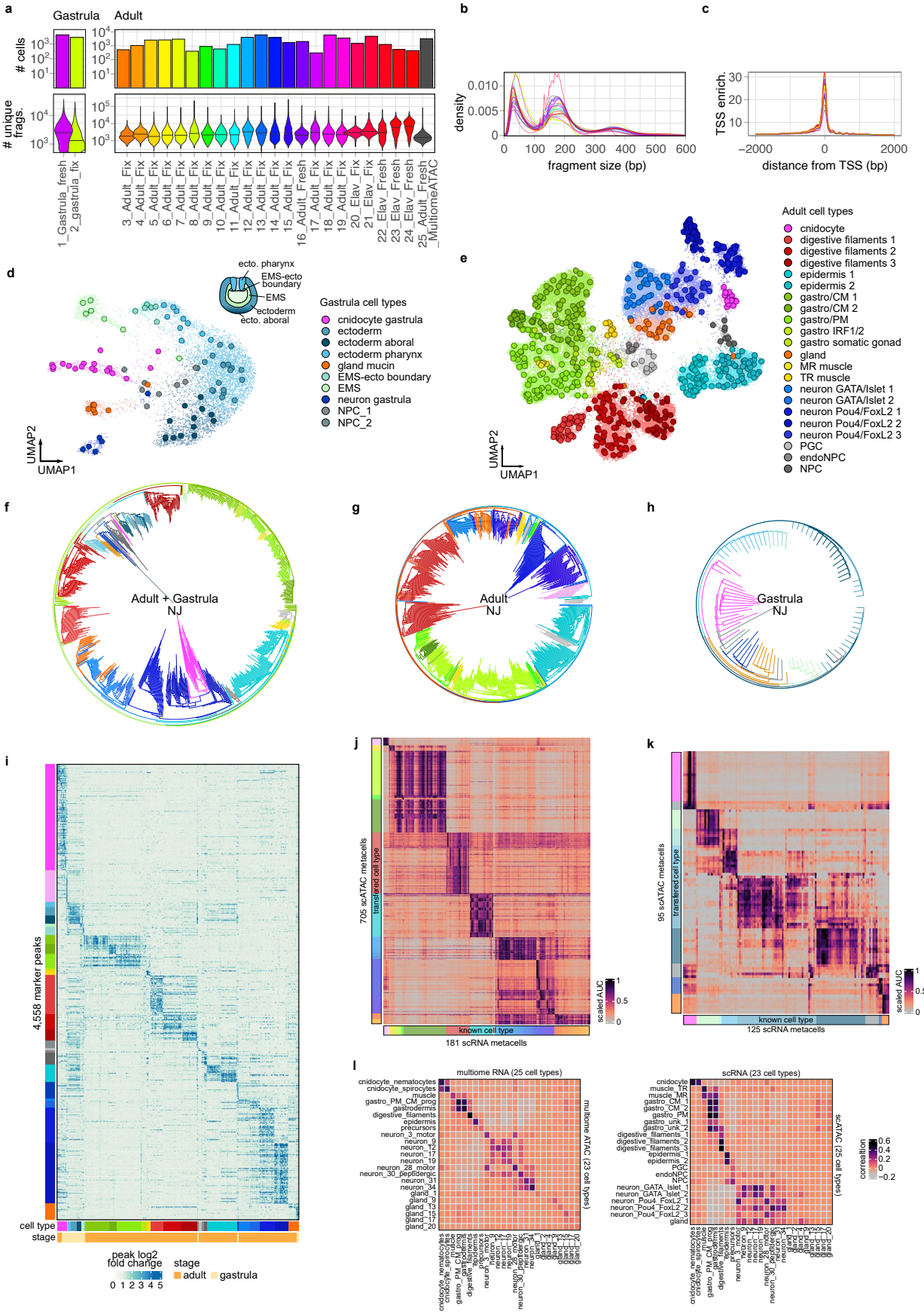
Extended data is available for this paper at https://doi.org/10.1038/s41559-025-02906-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-025-02906-1.

Correspondence and requests for materials should be addressed to Marta Iglesias or Arnau Sebé-Pedrós.

Peer review information *Nature Ecology & Evolution* thanks Maria Ina Arnone, Ferdinand Marlétaz and Juan Tena for their contribution to the peer review of this work.

https://doi.org/10.1038/s41559-025-02906-1



Extended Data Fig. 1 | See next page for caption.

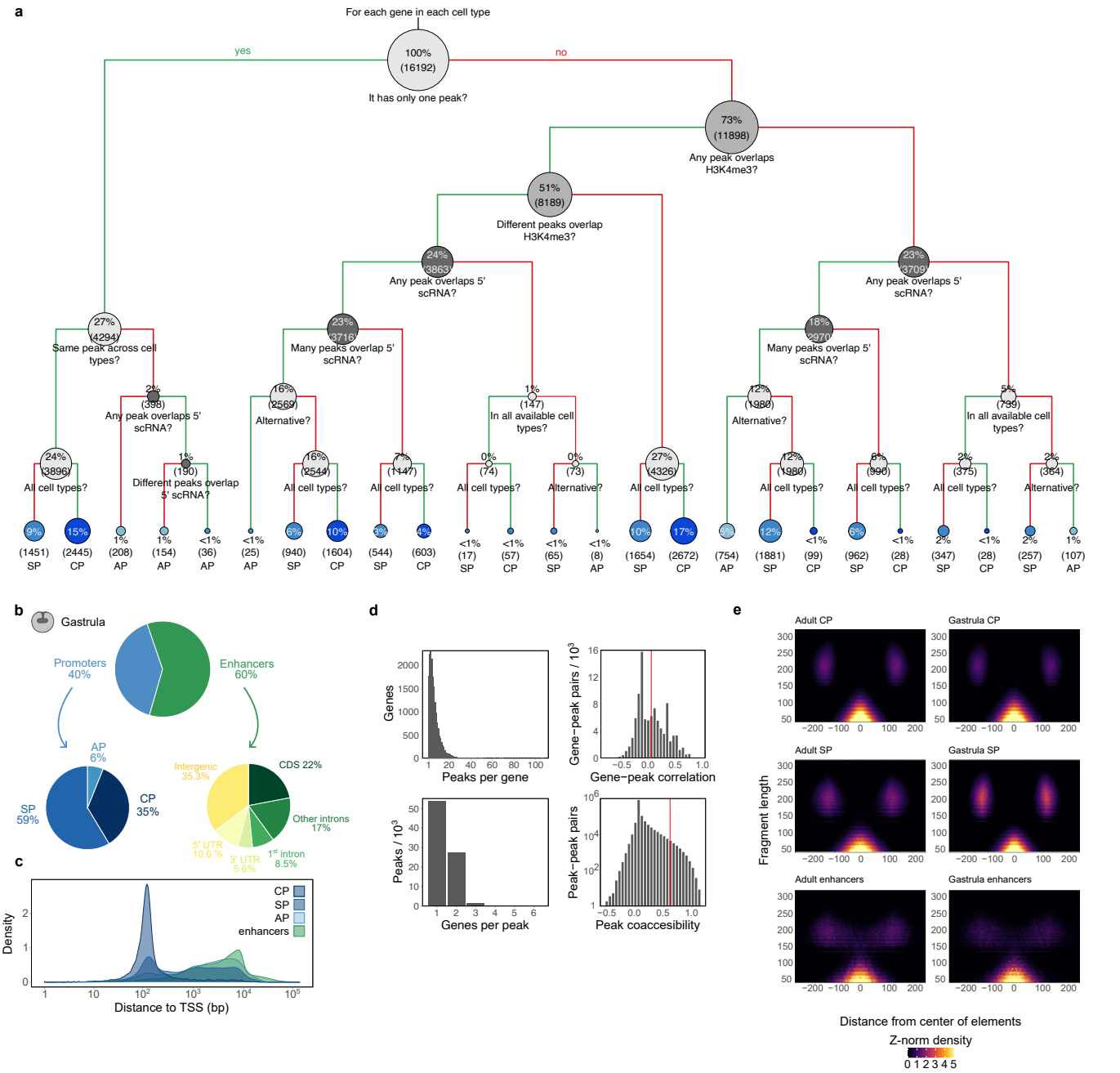
https://doi.org/10.1038/s41559-025-02906-1

Extended Data Fig. 1 | scATAC-seq dataset QC, clustering and annotation. **a**, Number of cells (top) and unique fragments per cell (bottom), **b**, scATAC-seq fragment size distribution for each sample. **c**, TSS enrichment signal for each sample. **d**, UMAP projection of single cells and metacells for gastrula dataset. **e**, UMAP projection of single cells and metacells for adult dataset. **f**, NJ clustering of metacells for adult and gastrula together, only for adult (**g**) and only for gastrula (**h**). **i**, Heatmap showing peak accessibility per cell type. **j**, Annotation transfer heatmap for adult scATAC-seq clusters. **k**, Annotation transfer heatmap for gastrula scATAC-seq clusters. **l**, Comparison of ATAC and RNA correlations for multiome (left) and separately profiled scATAC-seq and scRNA-seq data (right).

https://doi.org/10.1038/s41559-025-02906-1

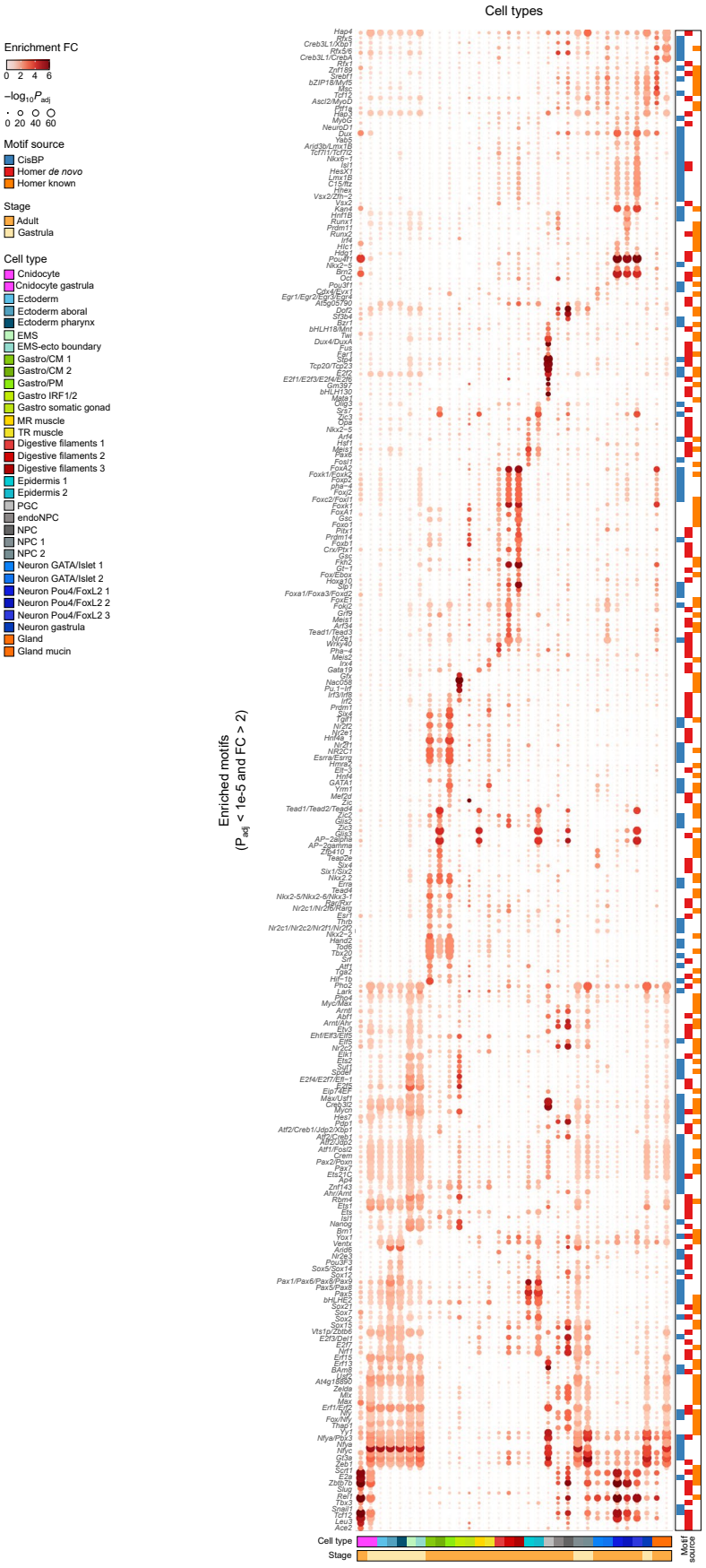
Extended Data Fig. 2 | Cell type markers. Comparison between accessibility scores and expression for selected marker genes.

https://doi.org/10.1038/s41559-025-02906-1



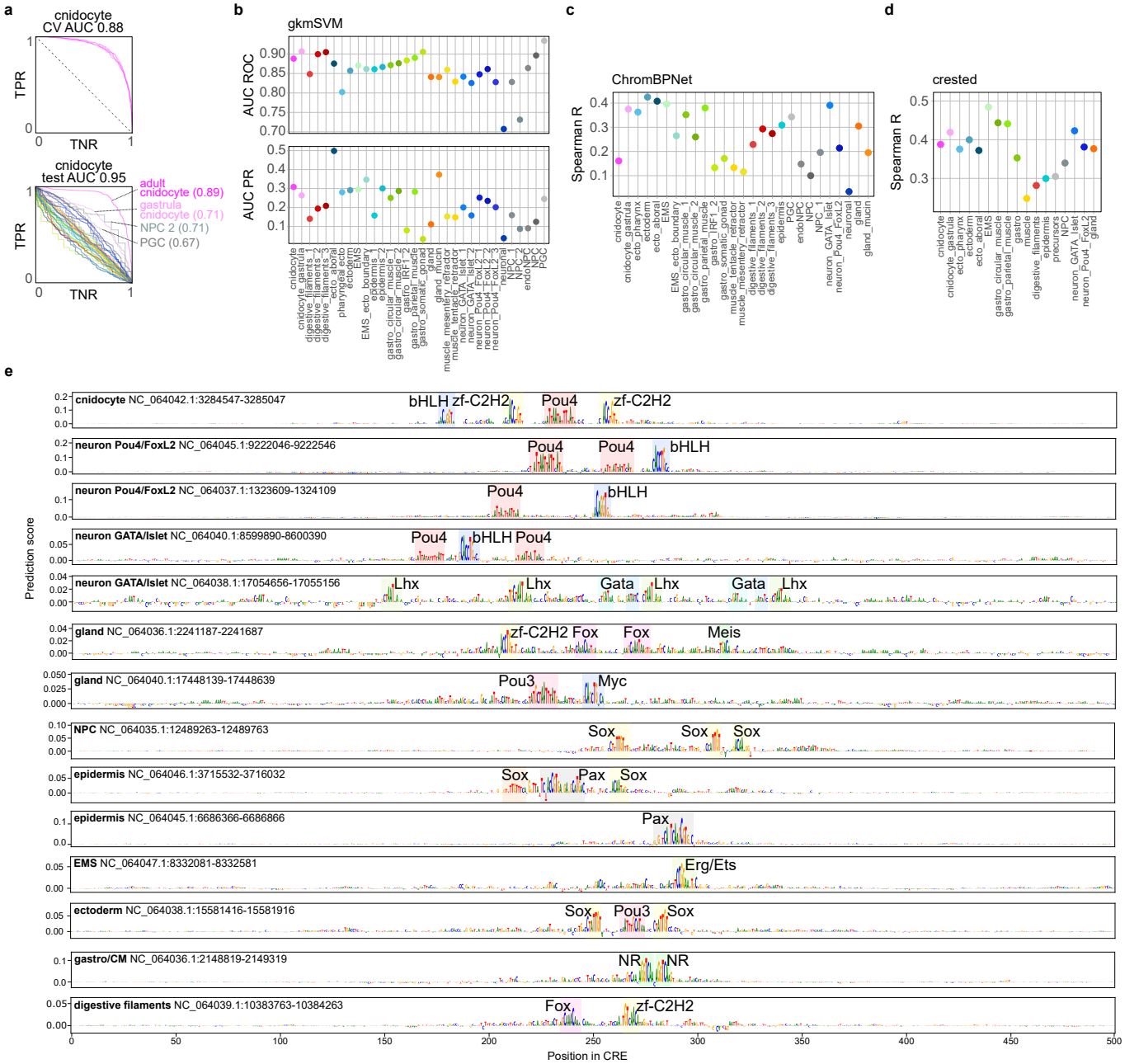
Extended Data Fig. 3 | *Cis*-regulatory element classification. a, Decision tree used to classify CRE into different promoter types. **b**, Fraction of gastrula CREs classified as promoters and enhancers. Promoters are further classified as constitutive promoters (CP), specific promoters (SP) and alternative promoters (AP). Enhancers are classified based on their overlap with different genomic regions. **c**, Distance to the nearest TSS distributions for different types of CREs. **d**, Summary peak statistics. Number of peaks per gene (top-left) and number of genes that each peak gets assigned to (bottom-left), correlation across metacells between peak accessibility and expression of the genes they are assigned to (top-right), and co-accessibility across cell clusters of all pairs of peaks (bottom-right). **e**, V-plots showing tagmentation fragment size distributions (y-axis) at different distances (x-axis) around CP, SP and distal CREs in adult and gastrula stages.

https://doi.org/10.1038/s41559-025-02906-1



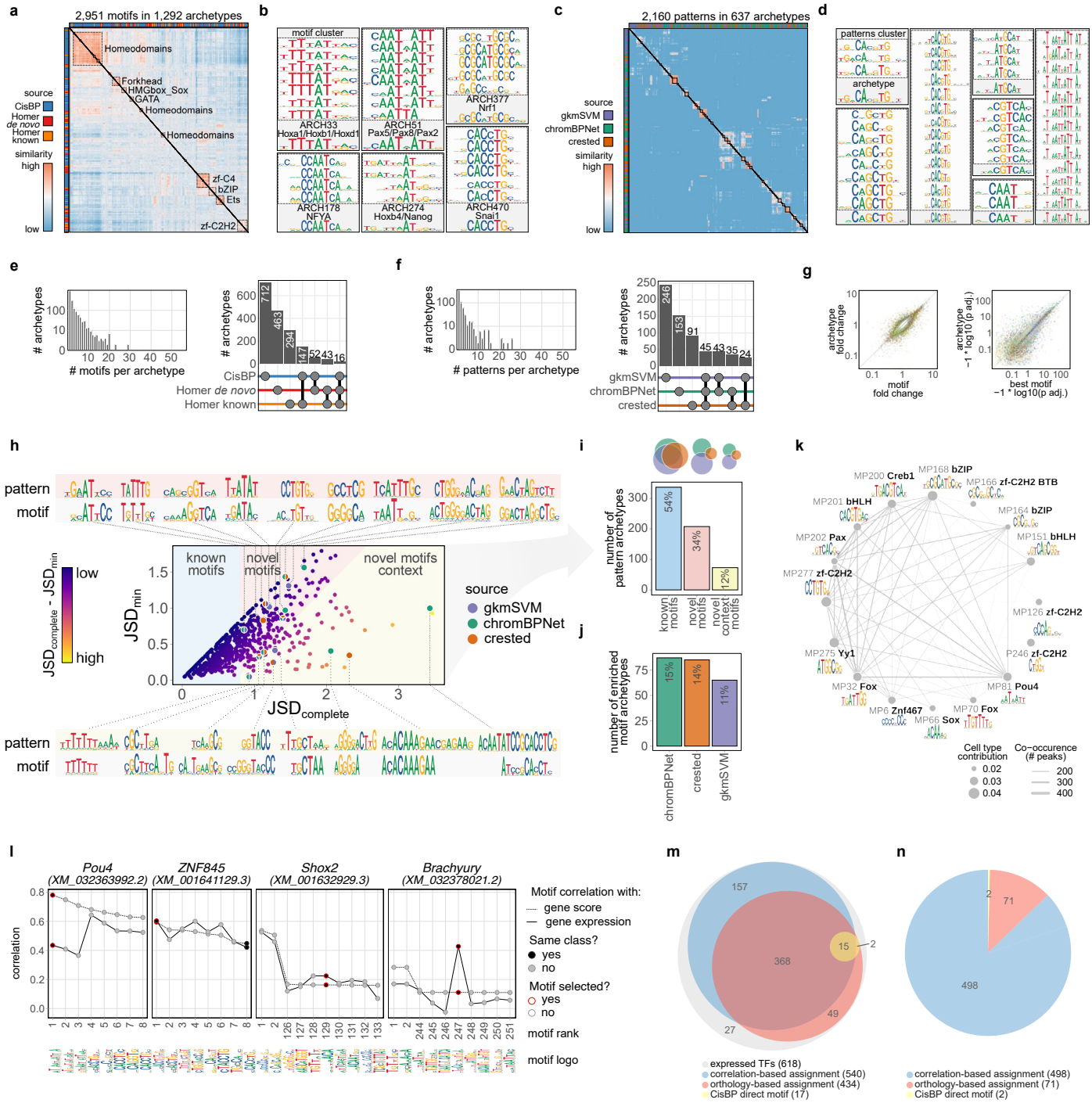
Extended Data Fig. 4 | Motif enrichment analysis. Dotmap of representative motifs enriched in different cell types. Significantly enriched motifs (fold change > 1, padj < 0.001) are grouped by motif-motif similarity and top enriched representative motifs for each cluster are indicated (hypergeometric test, FDR adjustment for multiple testing).

https://doi.org/10.1038/s41559-025-02906-1



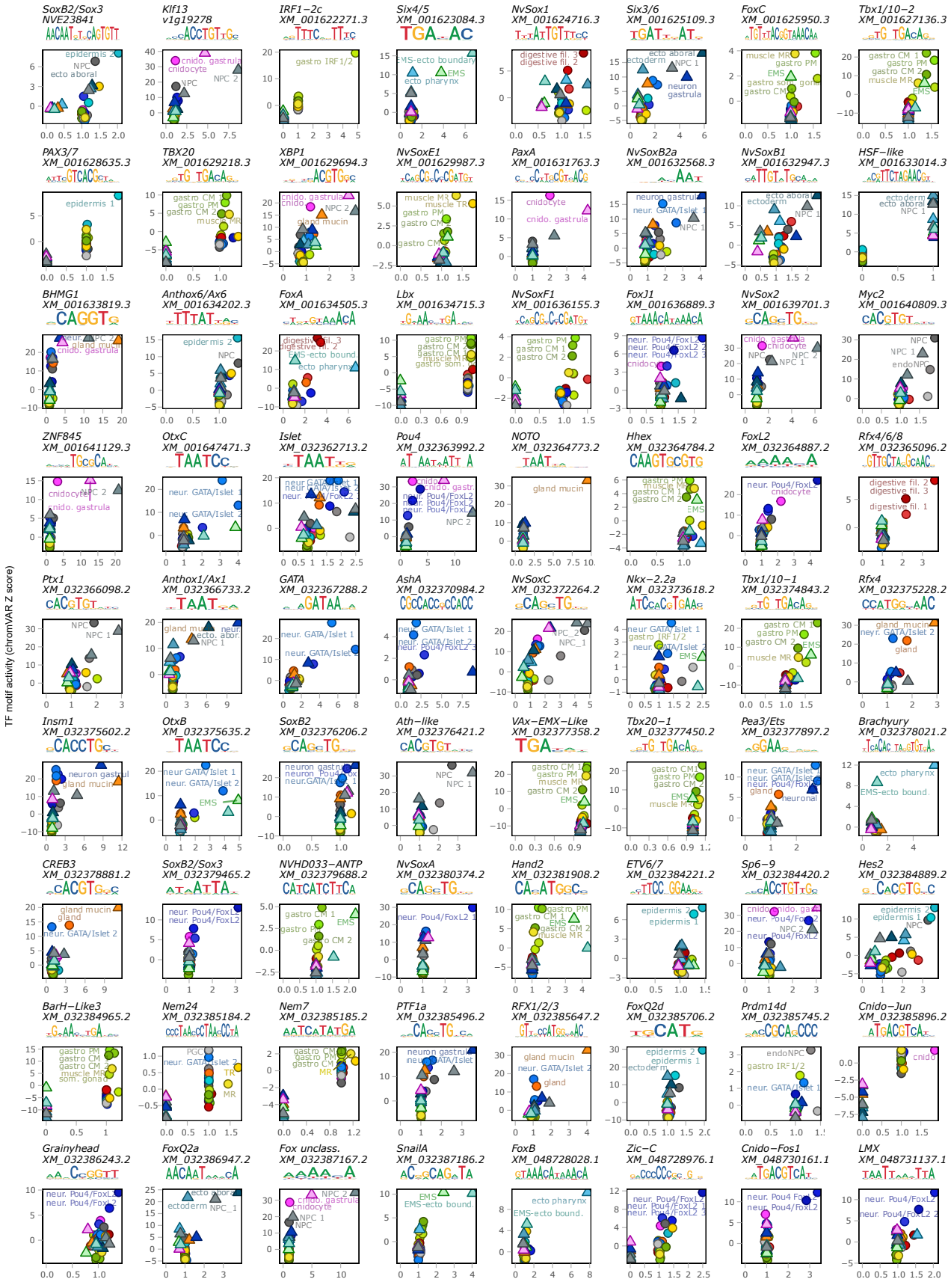
Extended Data Fig. 5 | Sequence models. a, Five-fold cross-validation (CV) area under the curve (AUC, top) and test set AUC (bottom) for gkm-SVM classifiers trained on adult cnidocytes. **b**, Area under receiver operator curve (AUC ROC, top) and area under precision recall curve (AUC PR, bottom) for all cell type gkm-SVM classifiers. **c**, Spearman correlation for ChromBPNet predicted accessibility counts in test set peaks. **d**, Spearman correlation for crested predicted accessibility counts in test set peaks. **e**, Nucleotide importance scores for top scored CREs in different cell types.

https://doi.org/10.1038/s41559-025-02906-1



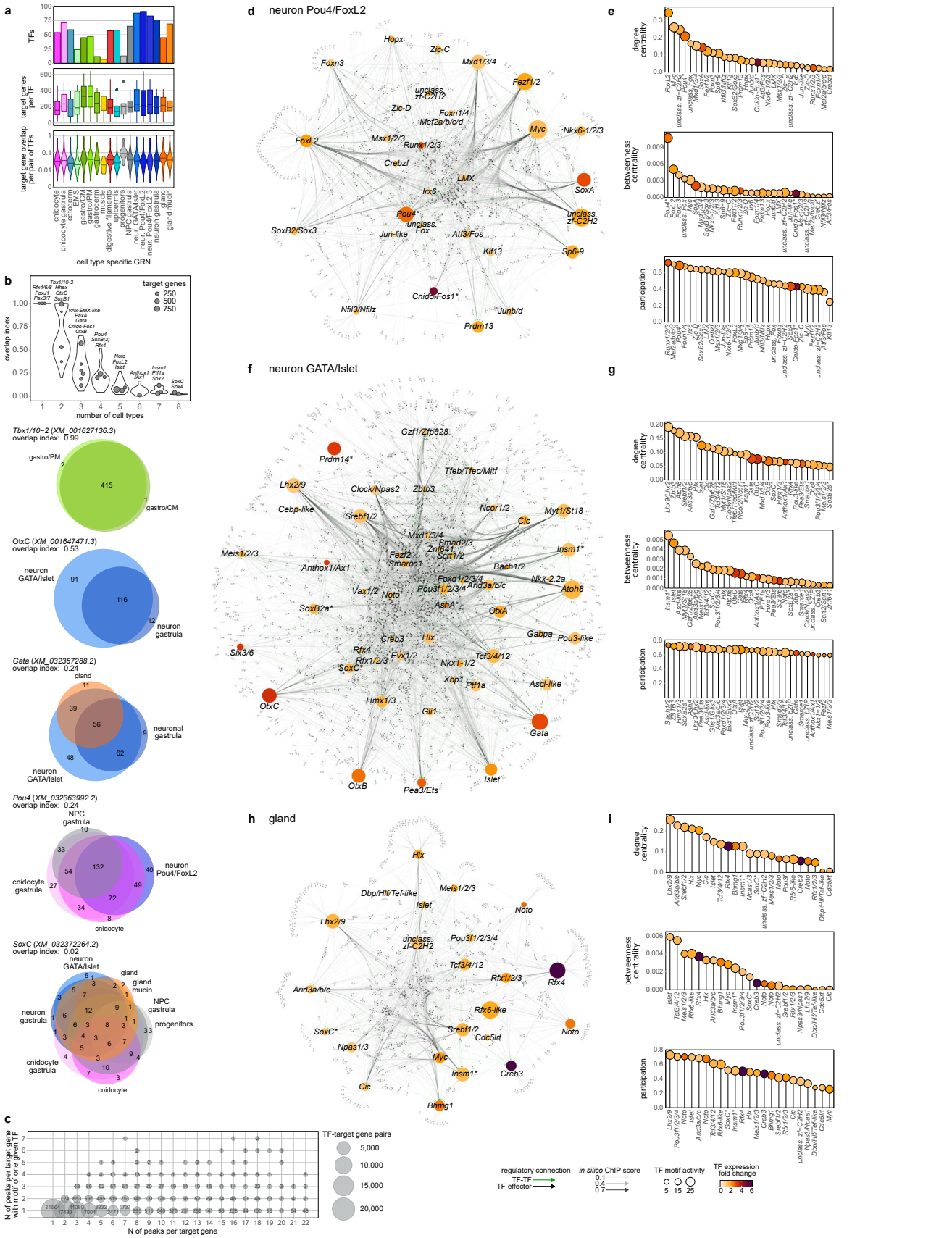
Extended Data Fig. 6 | Sequence motif discovery. a, Heatmap showing pairwise motif similarities used to generate motif archetypes from enriched motifs. **b**, Examples of motif clusters. **c**, Examples of motif archetypes. **d**, Same as (a) and (b) for patterns discovered with sequence models. **e**, Number of motifs per archetype (left) and number of archetypes composed of motifs from different sources (right). **f**, Number of patterns per archetype (left) and number of archetypes composed of patterns from different sequence models (right). **g**, Comparison of motif enrichment fold change (left) and adjusted p-value (right) for archetypes versus best scoring motif in each archetype cluster (hypergeometric test, FDR adjustment for multiple testing). **h**, For all pattern archetypes and their most similar motif archetype, Jensen-Shannon divergence (JSD) calculated across the best pairwise alignment of archetypes (x-axis, $JSD_{complete}$), and calculated across the best alignment spanning the length of shorter archetype (y-axis, JSD_{min}). Based on these two metrics, pattern archetypes are classified as being novel motifs, having novel context or resembling known motifs from motif enrichment analysis. **i**, Fraction of pattern archetype classes defined in h, Euler diagrams summarizing the source of pattern archetypes (that is sequence models) for each of these three categories are shown on top. **j**, Fraction of enriched motifs found with each of the sequence models, **k**, Co-occurrence network of pattern archetypes with contribution in cnidocytes. Size of the node reflects its cell type contribution, and width of the connection scales with the number of CREs in which two motifs co-occur. **l**, Correlation-based approach for assigning motifs to TFs. For each TF, we rank motifs based on correlation of motif activity to TF accessibility and expression, and assign it top ranking motif of the same structural class. **m**, Euler diagram showing TF coverage using different motif-to-TF assignment methods for expressed *Nematostella* TFs. **n**, Final motif assignment sources for expressed *Nematostella* TFs.

https://doi.org/10.1038/s41559-025-02906-1



Extended Data Fig. 7 | Examples of TF expression and TF motif activity correlations for selected marker genes.

https://doi.org/10.1038/s41559-025-02906-1



Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41559-025-02906-1

Extended Data Fig. 8 | Cell type gene regulatory networks. a, Number of TFs in GRNs inferred for each broad cell type (top), number of genes targeted by each TF (middle), and fraction of overlapping target genes for each pair of TFs (bottom). **b**, Overlap of target genes for the same TF across cell types, plotted for groups of TFs active in different number of cell types. Selected TFs are highlighted on the plot and overlap of their target genes is shown as Euler diagrams below. **c**, Number of CREs per target gene (x-axis) compared to number of CREs of the same gene with any single TF motif (y-axis). Most TFs have binding motif in a single CREs of their target genes**. d-g**, Additional inferred GRN and TF connectivity measurements for neuro-secretory cell types: GATA/Islet neurons (**d-e**), Pou4/FoxL2 neurons (**f-g**) and gland cells (**h-i**). Asterisks highlight TFs known to be involved in neurosecretory development.

nature portfolio

Corresponding author(s): Arnau Sebe-Pedros

Last updated by author(s): Oct 14, 2025

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection

Sofwtare used to collect/preprocess data in this study (package/version/source):
bcl2fastq 2.20 Illumina
CellRanger 7.0.0 10x genomics

Data analysis

Code https://github.com/sebepedroslab/nvec-scatac

Open source software/packages used in for data analysis in this study (package/version/source):
bwa 0.7.17 github.com/lh3/bwa
metacell 0.3.41 github.com/tanaylab/metacell
ArchR 1.0.2 github.com/GreenleafLab/ArchR
SEACells 1.0.0 github.com/dpeerlab/SEACells.git
py4cytoscape 1.11.0 github.com/cytoscape/py4cytoscape
pygenometracks 3.5 Bioconda
macs2 2.2.7.1 pypi.org
numpy 2.1.3 numpy.org
pandas 2.2.3 pandas.pydata.org
modisco-lite 2.3.2 pypi.org
crested 1.4.1 pypi.org
rocker/shiny 4.0.5 docker.com
ChromBPNet 1.0 github.com/kundajelab/chrombpnet
data.table 1.13.0 cran.r-project.org

ComplexHeatmap 2.12.1 cran.r-project.org
ggplot2 3.3.2 cran.r-project.org
ggrepel 0.9.6 cran.r-project.org
patchwork 1.3.0 cran.r-project.org
monaLisa 1.2.0 Bioconductor
rtracklayer 1.56.1 Bioconductor
SummarizedExperiment 1.26.1 Bioconductor
SingleCellExperiment 1.18.1 Bioconductor
DropletUtils 1.16.0 Bioconductor
ape 5.0 Bioconductor
phytools 0.7-47 Bioconductor
treeio 1.6.2 Bioconductor
tidytree 0.3.3 Bioconductor
stringr 1.4.0 Bioconductor
GenomicRanges 1.54.1 Bioconductor
GenomicFeatures 1.54.3 Bioconductor

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All sequencing data is deposited in GEO under accession number GSE294388. Processed data, annotation tables, and code for reproducing the analysis will be is in available in our lab Github page (: https://github.com/sebepedroslab/nvec-scatac). All generated datasets can be explored in interactive genome browsers: hhttps://sebelab.crg.eu/nematostella-cis-regulatory-atlas/ and https://sebelab.crg.eu/nematostella-cis-reg-jb2

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	NA
Population characteristics	NA
Recruitment	NA
Ethics oversight	NA

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	Sequencing depth and number of libraries were defined to allow support for the paper's main conclusions (there is not "sample size" in this paper)
Data exclusions	No data were excluded from the analysis.
Replication	Replication is not applicable to descriptive single-cell atlas studies. Robustness is instead achieved by profiling large numbers of cells across individuals and consistently identifying similar cell states across experiments, represented by dozens or hundreds of individual cell profiles.

Randomization	Randomization was not applicable in this study because the experimental design did not involve treatment groups, interventions, or subjective outcome assessments where investigator knowledge could introduce bias.
Blinding	Blinding was not applicable in this study because the experimental design did not involve treatment groups, interventions, or subjective outcome assessments where investigator knowledge could introduce bias. A single-cell atlas study is primarily descriptive and exploratory, aiming to comprehensively map cell types, states, and molecular features within a given tissue or organism.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Antibodies

Antibodies used	rabbit anti-DsRed (Clontech 632496) goat anti-rabbit Alexa568 (Life Technologies A11011)
Validation	No custom-made were used in this study. Commercial antobodies were previously validated in other Nematostella studies.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	This study used gastrula-stage embryos and non-sexed adult polyps derived from wild-type Nematostella vectensis animals (laboratory strain CH2 xCH6), as well as one-month-old NvElav1::mOrange positive polyps derived from a previously published transgenic line (PMID: 22159579)
Wild animals	No wild animals were used in this study
Reporting on sex	The sex of sampled individuals was not determined
Field-collected samples	No field-collected animals were used in this stufy
Ethics oversight	Ethical approval is not required for work on Nematostella

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