Science Advances
AAAS

Supplementary Materials for

A single-cell, time-resolved profiling of *Xenopus* mucociliary epithelium reveals nonhierarchical model of development

Julie Lee *et al.*

Corresponding author: Jakub Sedzinski, jakub.sedzinski@sund.ku.dk; Kedar Nath Natarajan, kenana@dtu.dk

Sci. Adv. **9**, eadd5745 (2023)
DOI: 10.1126/sciadv.add5745

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

Supplementary Text

Supplementary note 1: scRNA-seq processing and quality control

Here, we further describe the steps involved pre-processing scRNA-seq data, quality control and normalisation.

Mapping: The ten individual stages and corresponding scRNA-seq datasets were aligned to *Xenopus laevis* genome (XenBase RRID# SCR_003280; NCBI GCA_001663975.1) using annotation from the GitLab repository (https://gitlab.com/Xenbase/genomics/XENLA_9.2; 2018-May version) (*45*, *68*) using zUMI's alignment and feature counting (version 2.4.5b), which utilises STAR (version 2.5.4b) (*69*, *70*). The zUMIs were run with the following parameters: *BC_filter: (num_bases: 1, phred: 20)*, *UMI_filter: (num_bases: 1, phred: 20), barcodes: (automatic: yes, BarcodeBinning: 2, nReadsperCell: 100), counting_opts: (introns: yes, downsampling: '0', strand: 0, Ham_Dist: 1, velocyto: yes, primaryHit: yes, twoPass: no)*

Barcode filtering: To filter out background barcodes with few/minimal aligned reads from the ten individual stages and corresponding scRNA-seq datasets, we applied EmptyDrops (from DropletUtils version 1.2.2) using default parameters and selected good barcodes i.e., cells < 0.01 FDR threshold, for every individual stage dataset (*72*).

Doublet filtering: For removing doublets from the ten individual stages and corresponding scRNA-seq datasets, we applied scrublet (version 0.2.1) using parameters *expected_doublet_rate = 0.06*, *min counts = 2, min_cells = 3, min_gene_variability_pct = 85 and n_prin_comps = 30* to all datasets individually (*73*).

Normalisation: For each of the ten individual stages and corresponding scRNA-seq datasets, we first normalised each dataset for library size, log-transformed read counts and scaled genes to mean zero and unit variance, followed by identification of highly variable genes (scanpy version 1.4) (*74*). We integrated the count matrices from 10 ten stages and corresponding scRNA-seq datasets together, followed by normalisation of cell library size, log transformation of read counts and integration using Harmony (*78*).

Scaling: For individual analysis of ten stages and corresponding scRNA-seq datasets, we scaled gene expression to mean zero and unit variance genes using scanpy (version 1.4) (*74*).

Highly variable genes: For individual analysis of ten stages and corresponding scRNA-seq datasets, we used *scanpy.pp.highly_variable_genes with parameters min_mean=0.0125, max_mean=3, min_disp=0.5* to compute highly variable genes.

For integrated analysis of count matrices from 10 ten stages and corresponding scRNA-seq datasets together, we used *harmony.utils.hvg_genes no_genes=2000*. For heatmaps showing highly variable genes, we used *harmony.utils.hvg_genes no_genes=500* for each stage, amounting to 3906 unique genes.

Supplementary note 2: Public dataset comparison

Across our ten individual stages and corresponding scRNA-seq datasets, we first computed highly variable genes per stage (3906 genes), removing LOC, Xetrov, Xaleav and MGC gene groups from the list. We analysed the bulk RNA-seq dataset and marker genes from Quigley et al. (*38*) and identified 1,376 common genes expressed across both datasets. We computed pseudobulk cells across our scRNA-seq 10 stages and 15 developmental clusters, and calculated pairwise spearman correlation to capture cell type correlations between our data and respective public datasets.

We selected core-multiciliated and ionocyte markers from Quigley et al. (*38*) and basal cell markers (*18*) across pseudobulk cells from our scRNA-seq 10 stages and 15 developmental clusters.

For comparison with X. tropicalis atlas (*19*), we subset all the cell-types from non-neural ectoderm lineage (37, 270 cells), combining technical and biological replicates, spanning developmental stages 8 to 22, with 8 shared stages between both studies (stages 8, 10, 12, 13, 16, 18 and 20).

Supplementary note 3: Lineage inference

Here, we describe the construction and inference of lineages from scRNA-seq dataset over MCE clusters and developmental stages.

To infer developmental lineages over time-series datasets, we utilised a neighbour mapping and voting strategy. Briefly, we use the scRNA-seq annotations of 10 developmental stages and 15 developmental clusters as a reference and compare the cluster association within and across subsequent developmental stages. We compute Louvain clusters for each stage and map its association across subsequent development stages. For every cell identified within a stage, the cell assigns a vote for the cluster of its nearest neighbour in the preceding timepoint (For example, stages 27 and 24). The votes are aggregated within a cluster and a connection (link) to its likely developmental ancestor is made. We performed the neighbour search using Scipy (version 1.1.0) function scipy.spatial.cKDTree with default parameters. We assess the voting confidence between clusters based on marker genes (expression trends), refine ambiguous links based on key branching genes and prune spurious low-confidence links. For example, we refined classification of Stage 16, cluster 1 (1_st16) to Stage 13, cluster 0 (0_st13), instead of Stage 13, cluster 1 (1_st13) based on expression trends of multiciliated marker genes (*mcidas*, *cav3*, *tubb4b*). Similarly, we updated basal cell lineage classification from Stage 20, cluster 2 (2_st20) to Stage 18, cluster 2 (2_st18) and to Stage 16, cluster 0 (0_st16), based on marker gene expression (*otog*, *foxi1* and *pfn1*). We repeat the approach for all cells across pairs of successive developmental stages to form the lineage map.

Supplementary note 4: Single-cell evolutionary development comparison

To investigate the cell types, expression modules and relationships within mucosal epithelium across species, we performed a comparative analysis of 9 datasets of nasal, airway epithelium spanning 144 cell-types across 120,842 single-cells, with our dataset. We downloaded single-cell expression datasets from respective repositories for the below datasets (Table S12).

Mapping algorithm: We first converted all mouse and Xenopus genes to orthologous human gene names across all datasets, retaining only unique gene names.

The mapping algorithm is structured in 4 steps:

1. Pairwise differential expression between all cell types in all datasets.
2. Identification of dataset-wise intersecting marker genes.
3. Cell type enrichment of marker gene sets.
4. Pairwise spearman correlation of gene enrichment in cell types.

Pseudocode

```
For i in datasets:
    For j in datasets:
     DBi = DE(i)
     DBj = DE(j)
     GS =  DB_i ∩ DB_j

     Enr_i = Mean_per_type(i[GS]) / Mean_all_type(i[GS])
     Enr_j = Mean_per_type(j[GS]) / Mean_all_type(j[GS])
     Out[CT_i, CT_j ] = Spearman_correlation(Enr_i, Enr_j)
```

Where:

i and j are datasets with uniquely names cell types and ,
DBi and DBj are marker genes per cell type in datasets i and j respectively,
GS is the intersect in marker genes found in datasets i an j,
Enr_i and Enr_j are per cell type arrays of mean expression relative to all cell type mean expression in dataset i and j respectively,
Out is a matrix with all unique cell types across all datasets as rows and columns. Values represent pairwise spearman correlation between cell types, based on dataset-wise intersecting genes.

Next, we identified the differentially expressed (DE) genes for each dataset using scanpy rank_genes_groups (version 1.4) using author-annotated cell types and compared them against all cells (within the same dataset; $p < 0.05$). While we retain the author annotation for 'Clara

cells' for consistency, these cells should be referred to as Club cells. For each dataset and author annotated, cell types, a set of unique DE genes were therefore retained. For each dataset, the specific enrichment of DE genes to the author annotated cell types is quantified as:

$$DB = ct_1 \bigcup \quad ct_2 \bigcup \quad \dots ct_n$$

Finally, we calculated the mean expression of all genes, per author, annotated cell types (i.e., cell group) resulting in a cell type-by-gene matrix per dataset.

The specific gene enrichment per dataset was calculated as the mean of cell type relative to the mean in the whole dataset.

For cross-species comparison, we correlated each cell type (spearman rank correlation) to all other cell types both within and across datasets. The correlation was performed using DE genes identified and common across respective datasets, irrespective of species

$$Ct_A \in D_1$$
$$Ct_B \in D_2$$
$$GS = D_1 \cap D_2$$
$$AB = corr(ct_{A[Gs]}, ct_{B[Gs]})$$

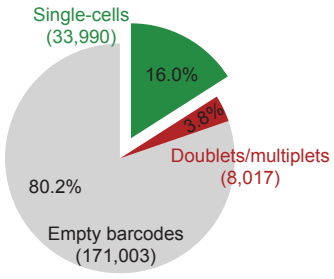
where:

$$Ct_A = Enrichment\ of\ genes\ for\ celltype\ A\ as\ an\ array$$
$$D_1 = Uniun\ of\ DE\ genes\ from\ all\ celltypes\ in\ dataset\ 1$$

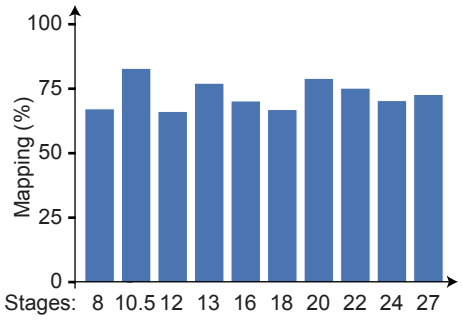
Module score: We annotated three major cell groups (Basal, Secretory, and Multiciliated), and scored each cell type (irrespective of dataset/species, 0 to 1 scale) based on differentially expressed gene sets (computed above). The scoring is performed based on the mean enrichment of genes that are differentially expressed in a minimum of 20% of cell types within the three major groups (Basal, Secretory, and Multiciliated).

Supplementary figure 1

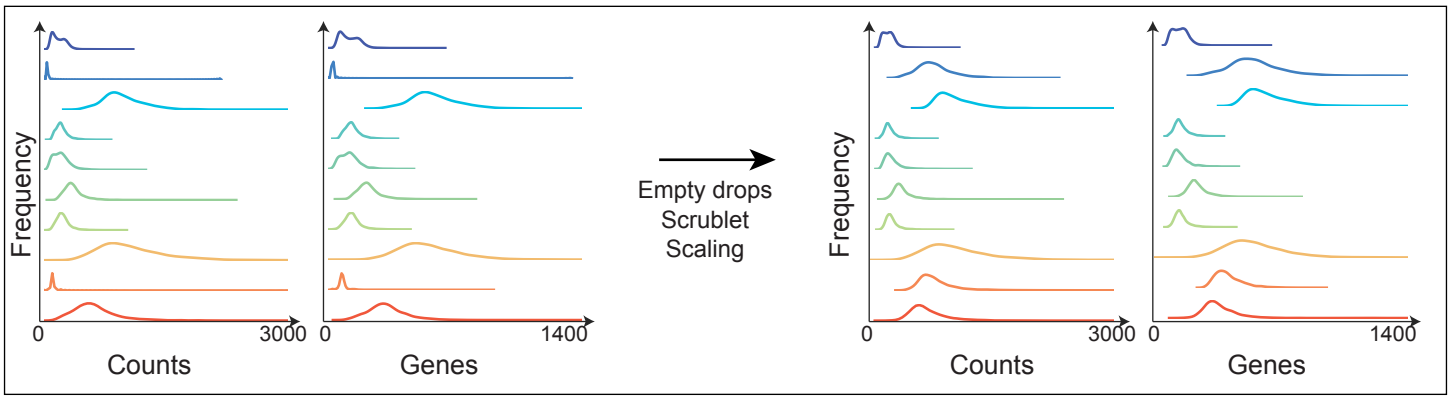
A



Stages	#cells
8	3408
10.5	1261
12	868
13	5823
16	3252
18	4228
20	3450
22	5222
24	3107
27	3371



B



C

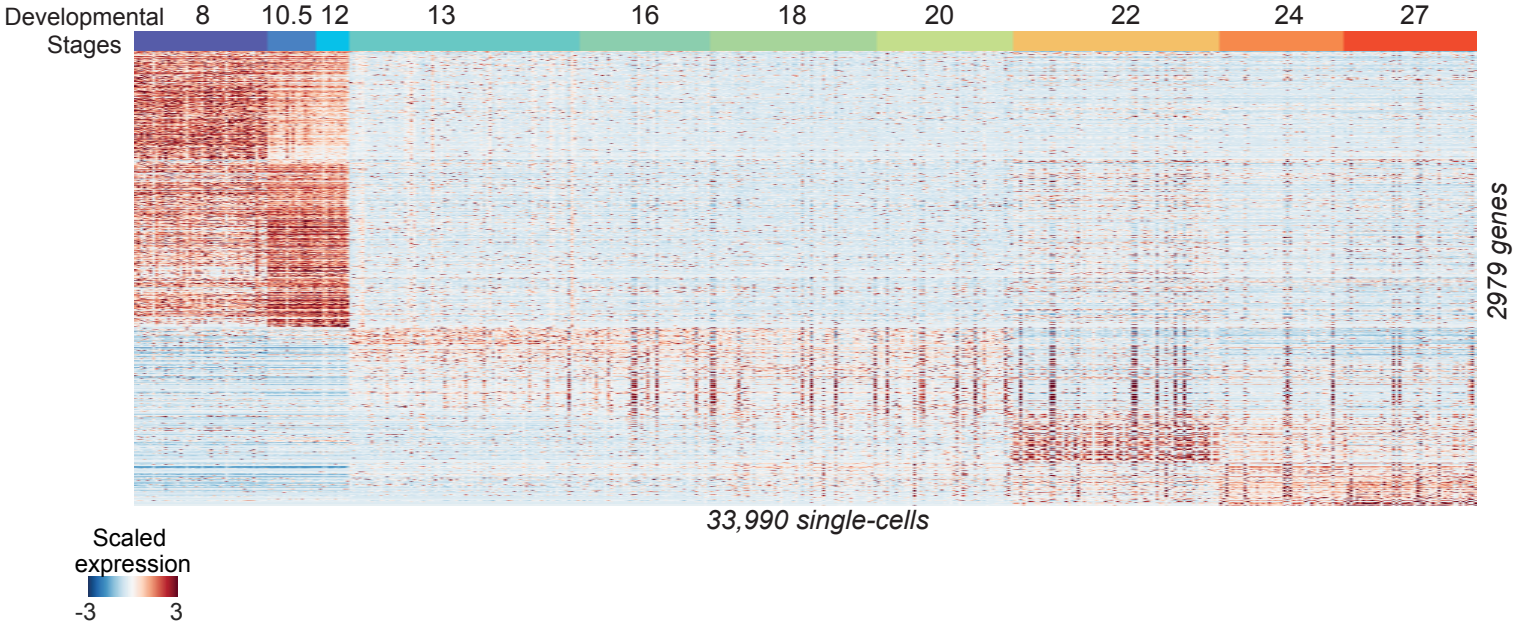


Fig. S1 scRNA-seq statistics and processing across MCE stages.

(**A**) Relative proportion of cells passing quality control after droplet scRNA-seq and mapping statistics across 10 stages of developing MCE.

(**B**) The expression distribution (counts and genes) before and after quality control for each developmental stage. The quality control analysis is performed to remove empty (residual/floating RNA), doublet/multiples and scale across different MCE stages.

(**C**) Expression pattern of 2,979 highly variable genes (HVGs) over 33,990 single-cells across 10 developmental stages. The scale bars indicate relative z-scaled expression across cells.

Supplementary figure 2

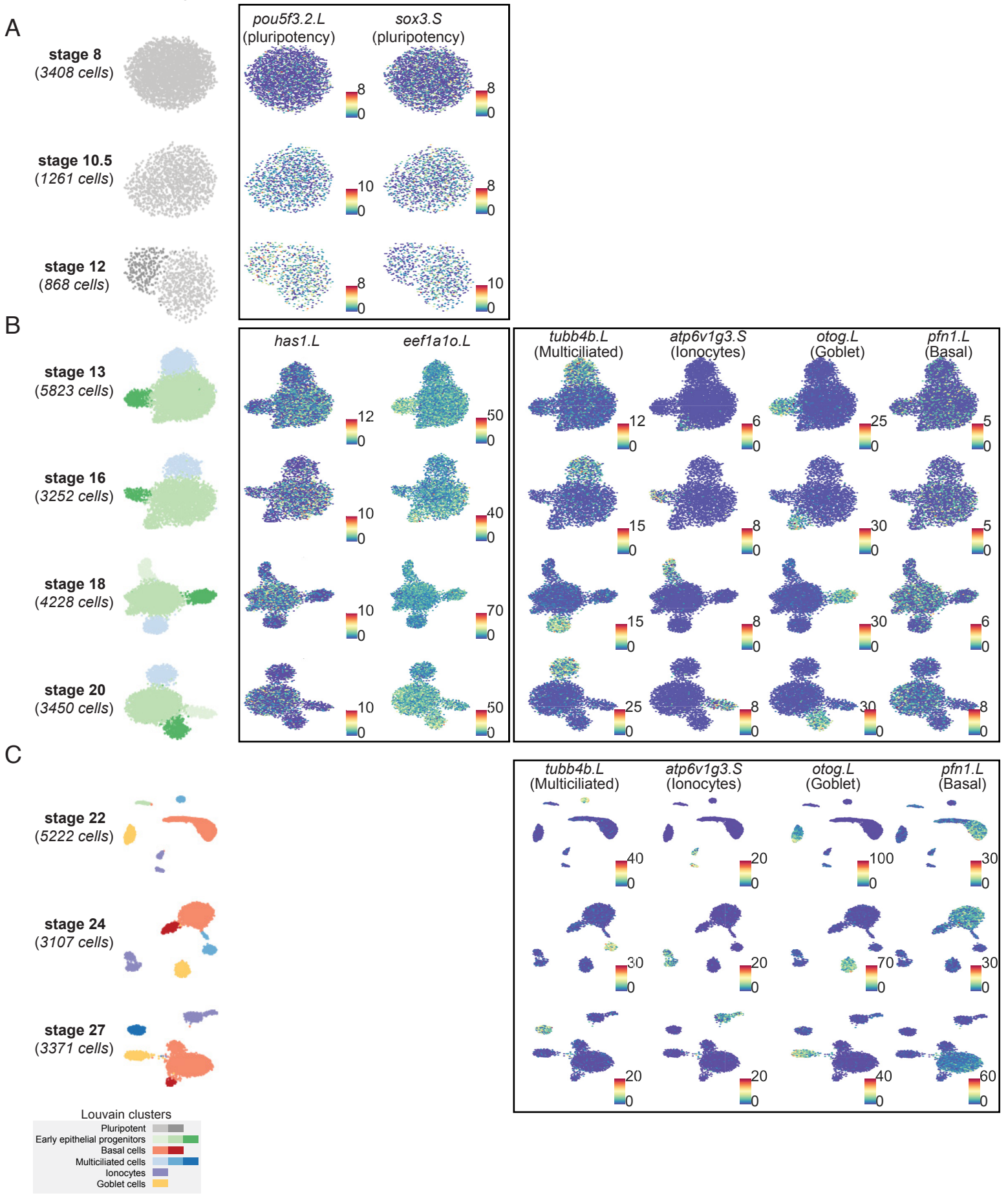


Fig. S2 Low dimensional scRNA-seq visualization of marker genes across stages

(**A**) UMAP plots showing blastula (stage 8) and gastrula stage (NF 10 and 12.5) cells marked by specific expression of pluripotency factors (*pou5f3.2.L* and *sox3.S*).

(**B**) Similar to **(A)**, but during neurula stages marking progenitor populations of multiciliated (*tubb4b.L*), ionocytes (*atp6v1g3.S*), goblet (*otog.L*) and basal cells (*pfn1.L*). The carbohydrate polymer hyaluronan synthase (*has1.L*) and elongation factor (*eef1a1o.L*) are expressed throughout neurula stages in all progenitors.

(**C**) Similar to **(A, B)**, but during early tailbud stages marking terminal cell types including multiciliated (*tubb4b.L*), ionocytes (*atp6v1g3.S*), goblet (*otog.L*) and basal (*pfn1.L*) cells.

The cells (left) are colored by cell types identified by Louvain clustering for individual stages. The scale bars indicate relative z-scaled expression across cells.

Supplementary figure 3

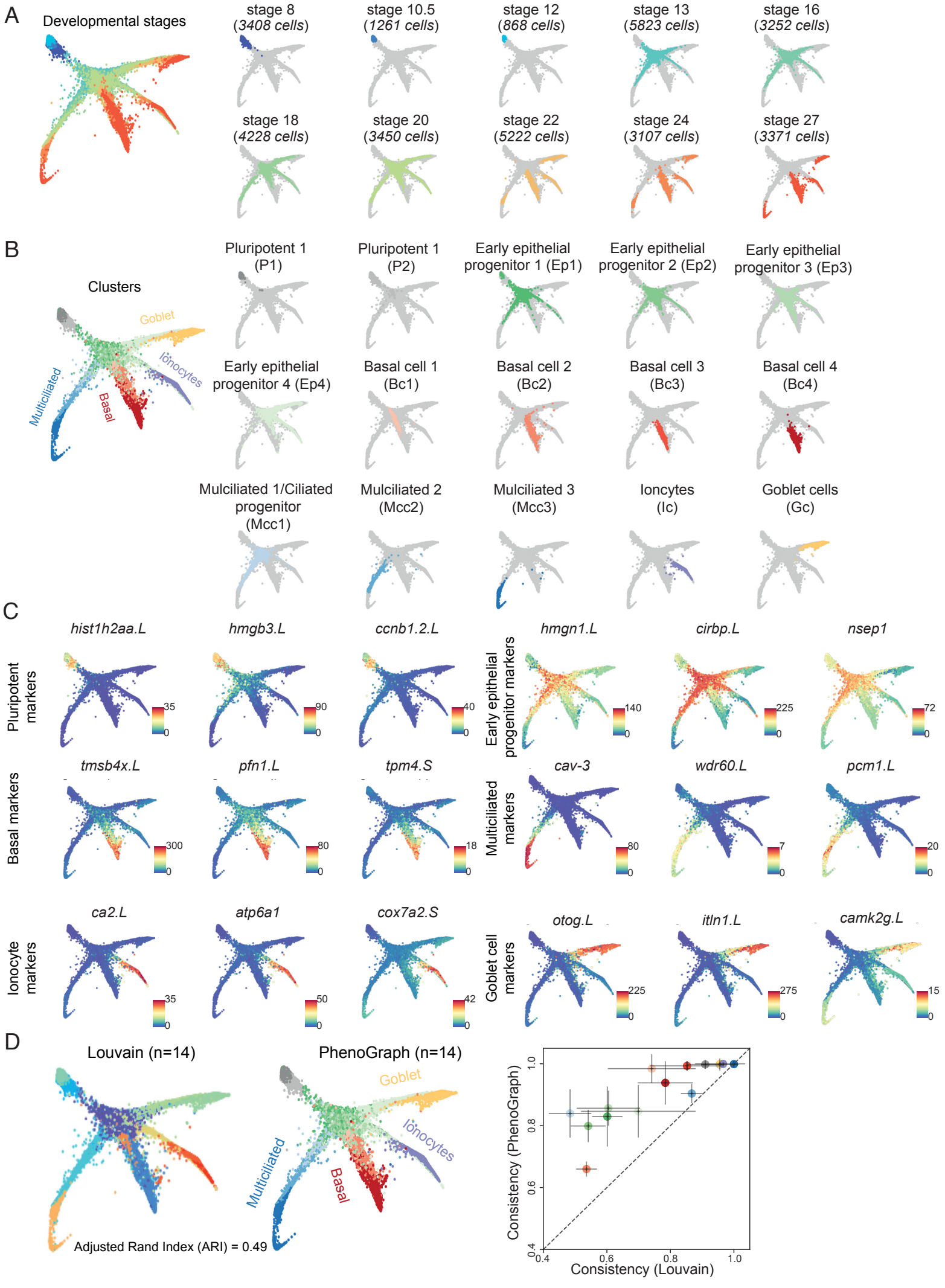


Fig. S3 knn-graph visualization of single-cells, stages, clusters and marker genes

(**A**) Embedding of single-cells during MCE development (MCE manifold) over a knn graph colored by stages (left). The number of cells across individual stages and their positions are overlaid and colored on knn-graph (right).

(**B**) Phenograph clusters highlighted over the MCE developmental manifold (left). The PhenoGraph clusters and respective cells are overlaid and colored on knn-graph (right).

(**C**) Expression of marker genes over MCE developmental manifold. The scale bars indicate the scaled imputed expression of respective markers.

(**D**) Adjusted Rand Index (ARI) highlighting the dissimilarity of the identified clusters between PhenoGraph and community-based Louvain clustering. Louvain resolution was set to achieve the same number of clusters for comparison (n), for this comparison. ARI is a measure of cluster similarity, which indicated only a 49% match between two clustering approaches. Consistency was computed using element-wise consistency (ECS), using random seeds and default resolution parameter.

Supplementary figure 4

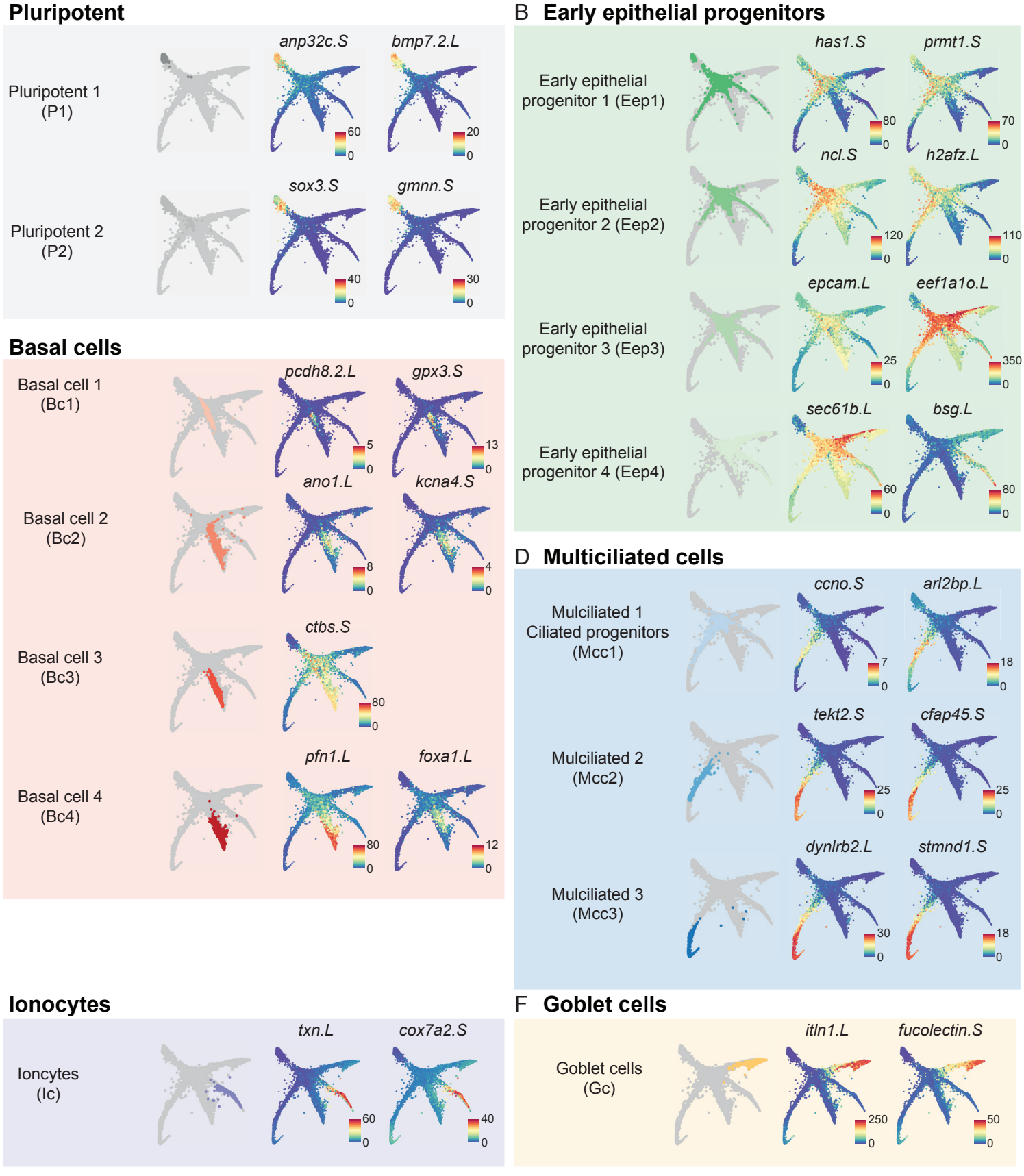


Fig. S4 Expression patterns of representative genes within subclusters

(**A**) MCE developmental knn graph showing expression of marker genes for each cell type and PhenoGraph cluster incl. pluripotent cells (A), early epithelial progenitors **(B)**, basal **(C)**, multiciliated **(D)**, ionocytes **(E)** and goblet cells **(F)**.

The scale bars indicated the scaled imputed expression of respective markers.

Supplementary figure 5

A

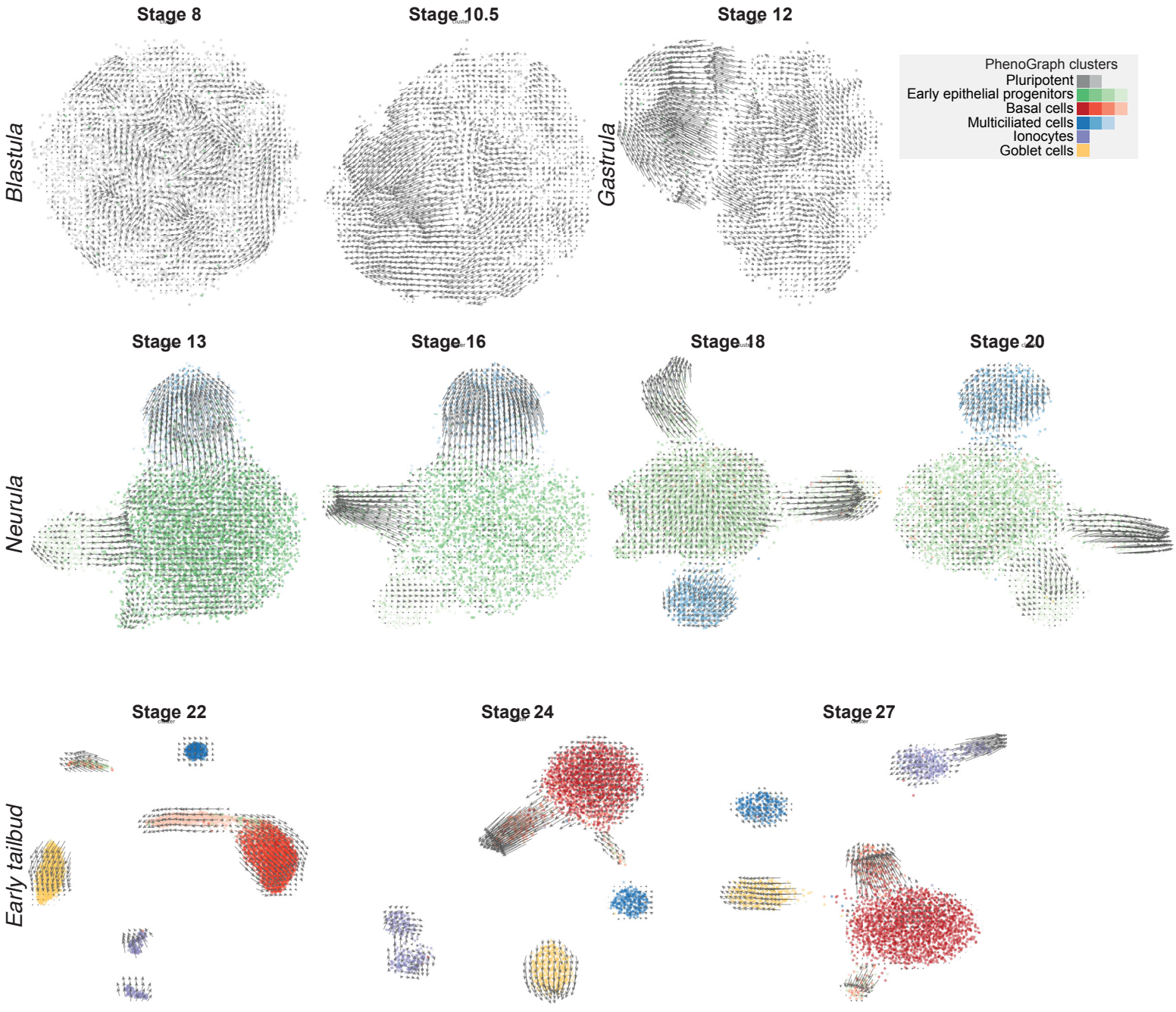


Fig. S5 RNA velocity across MCE developmental stages

(**A**) RNA velocity inferred over each MCE developmental stage and visualised over UMAP to highlight developmental propensities.

Supplementary figure 6

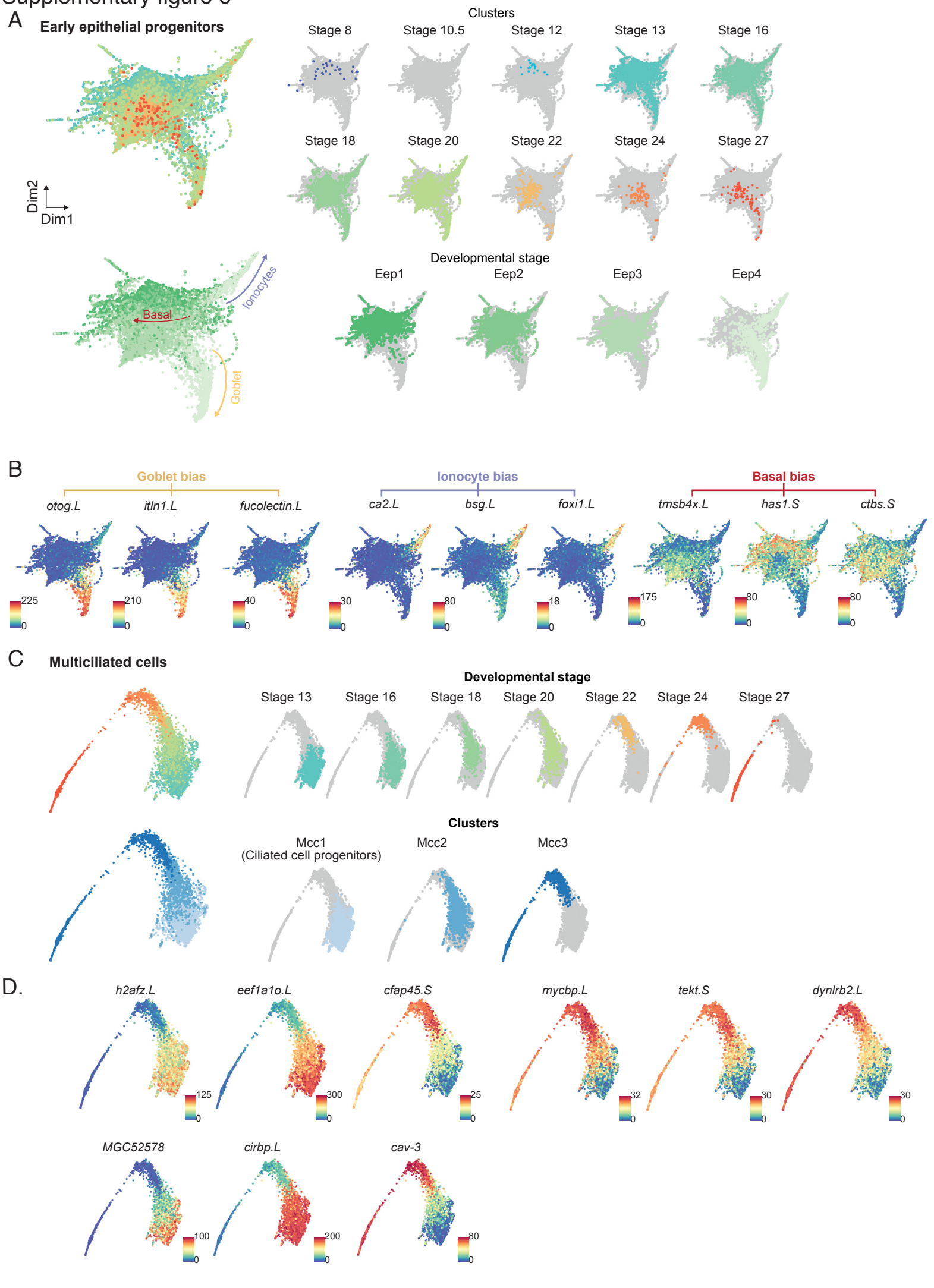


Fig. S6 Marker gene visualization in early epithelial progenitors and multiciliated cells

(**A**) Low dimensional graph embedding and visualisation of early epithelial progenitor cells (left), overlaid with cells from individual MCE developmental stages (top) and PhenoGraph clusters (bottom). The multi-lineage bias of early epithelial progenitors is indicated by arrows.

(**B**) Low dimensional visualisation of early epithelial progenitor cells indicating lineage bias, and overlaid with the expression of cell type markers including goblet (*otog.L*, *itln1.L*, *fucolectin.L*), ionocytes (*ca2.L*, *bsg.L*, *foxi1.L*) and basal cells (*tmsb4x.L*, *has1.S*, *ctbs.S*).

The scale bars indicate the scaled imputed expression of respective markers.

(**C**) Low dimensional visualisation of multiciliated cells alone (left), overlaid with cells from individual MCE developmental stages (top) and PhenoGraph clusters (bottom).

(**D**) Low dimensional visualisation of multiciliated cells overlaid with specific markers (*h2afz.L*, *eef1a10.L*, *cfap45.S*, *mycbp.L*, *tekt.S*, *dynlrb2.L*, *MGC52578*, *cirbp.L*, *cav-3*).

The scale bars indicate the scaled imputed expression of respective markers.

Supplementary figure 7

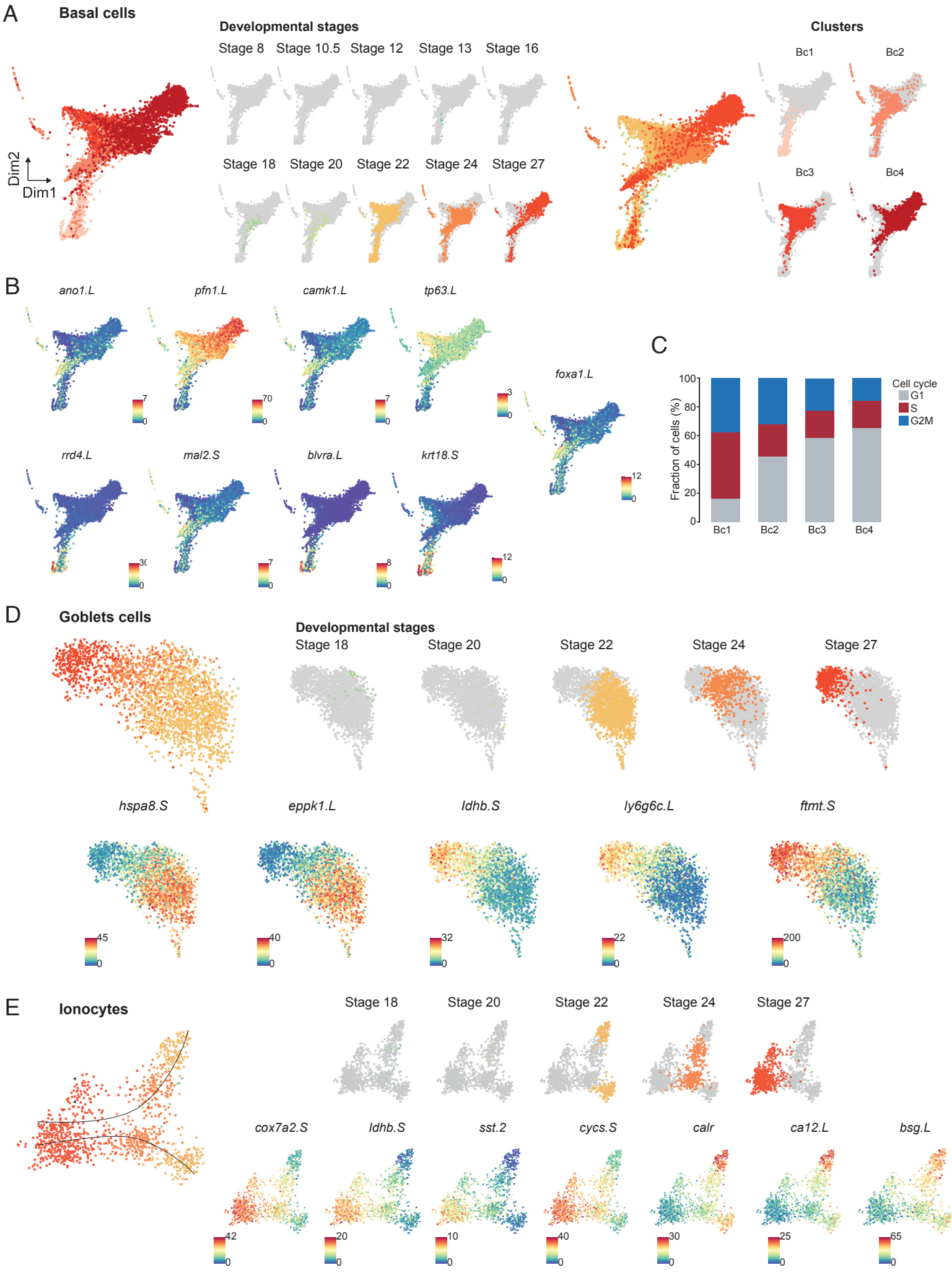


Fig. S7 Marker gene visualization in Basal, goblet cells and ionocytes

(**A**) Low dimensional visualisation of basal cells alone (left), overlaid with cells from individual MCE developmental stages (top) and PhenoGraph clusters (bottom).

(**B**) Low dimensional visualisation of basal cells overlaid with specific markers (*ano1.L*, *pfn1.L*, *camk1.L*, *rrd4.L*, *mal2.S*, *blvra.L*, *tp63.L*, *krt18.S*), including small secretory cell marker (SSC: *foxa1.L*).

(**C**) The basal cell subclusters can also be differentiated based on the proportion of cells in different cell cycle stages.

(**D**) Low dimensional visualisation of goblet cells alone (left), overlaid with cells from individual MCE developmental stages (top) and overlaid with specific markers (*hspa8.S*, *tmsb4x.L*, *ly6g6c.L*, *Idhb.S*, *eppk1.L*, *ftmt.S*).

(**E**) Low dimensional visualisation of ionocyte cell subpopulations alone (left), overlaid with cells from individual MCE developmental stages (top) and overlaid with specific markers (*tmsb4x.L*, *cox7a2.S*, *ca12.L*, *cycs.S*, *calr*, *Idhb.S*, *sst2*, *bsg.L*). Two ionocyte subtypes can be observed based on marker expression over developmental stages (black arrows)

The scale bars in panels (B and E) indicate the scaled imputed expression of respective markers.

Supplementary figure 8

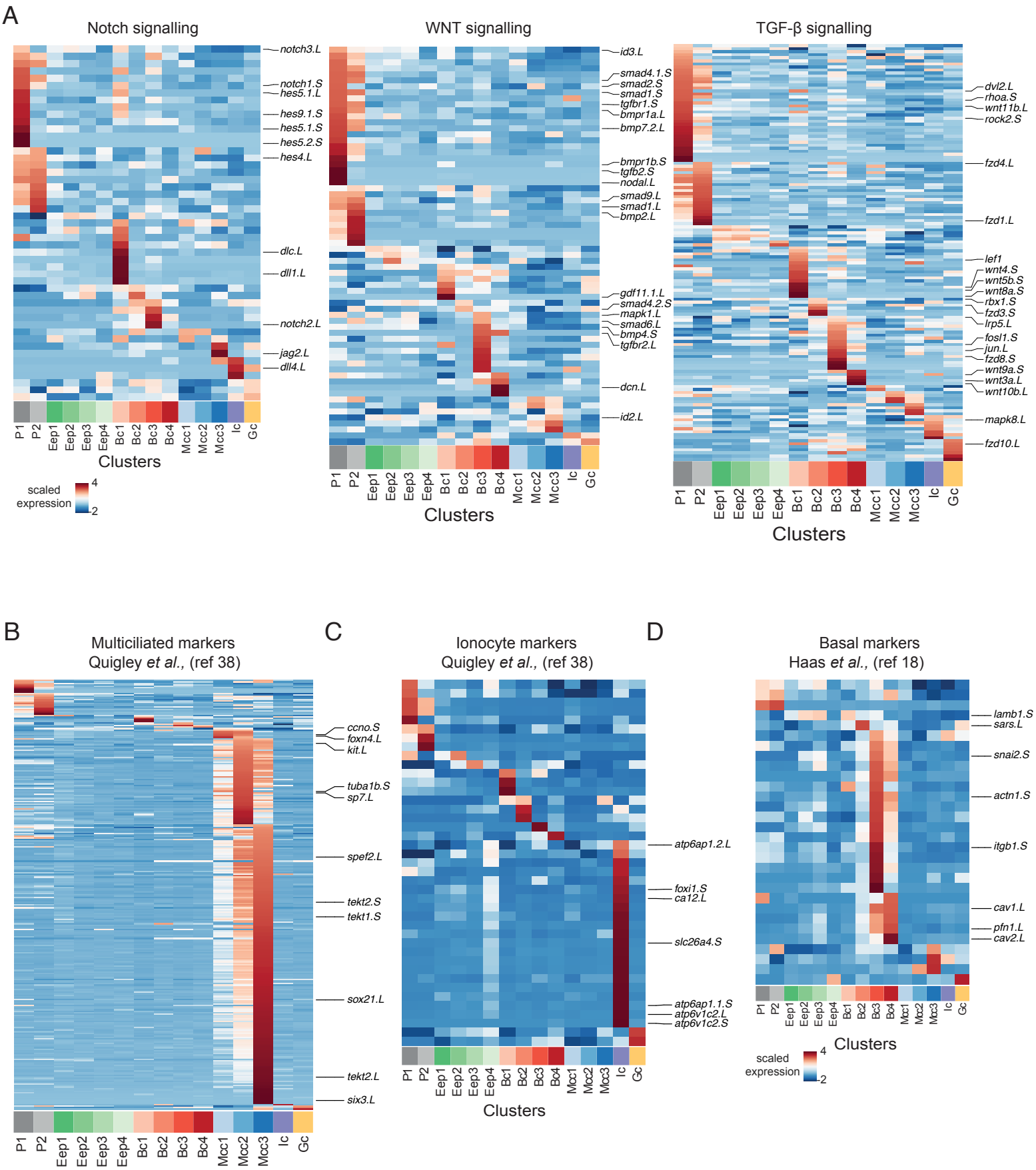


Fig. S8 Expression pattern of signalling genes and comparison with public datasets

(**A**) Heatmap showing the expression pattern of Notch, Wnt and TGF-B signalling pathways over MCE developmental stages and PhenoGraph clusters respectively.

(**B**) Heatmap showing the expression pattern of multiciliated core genes (y-axis), identified across MCE PhenoGraph clusters (x-axis). Step-wise expression and maturation of MCC core genes can be observed from Mcc1 to Mcc2 to Mcc3. While most multiciliated markers are indeed expressed in MCC subclusters, many bulk genes are heterogeneously expressed in basal, goblet cells and ionocytes.

(**C**) Heatmap showing the expression pattern of ionocyte core genes (y-axis) over MCE PhenoGraph clusters (x-axis).

(**D**) Heatmap showing the expression pattern of basal cells core genes (y-axis) over MCE PhenoGraph clusters (x-axis).

The scale bars in panel (A-D) indicate the z-scaled mean expression of respective markers.

Supplementary figure 9

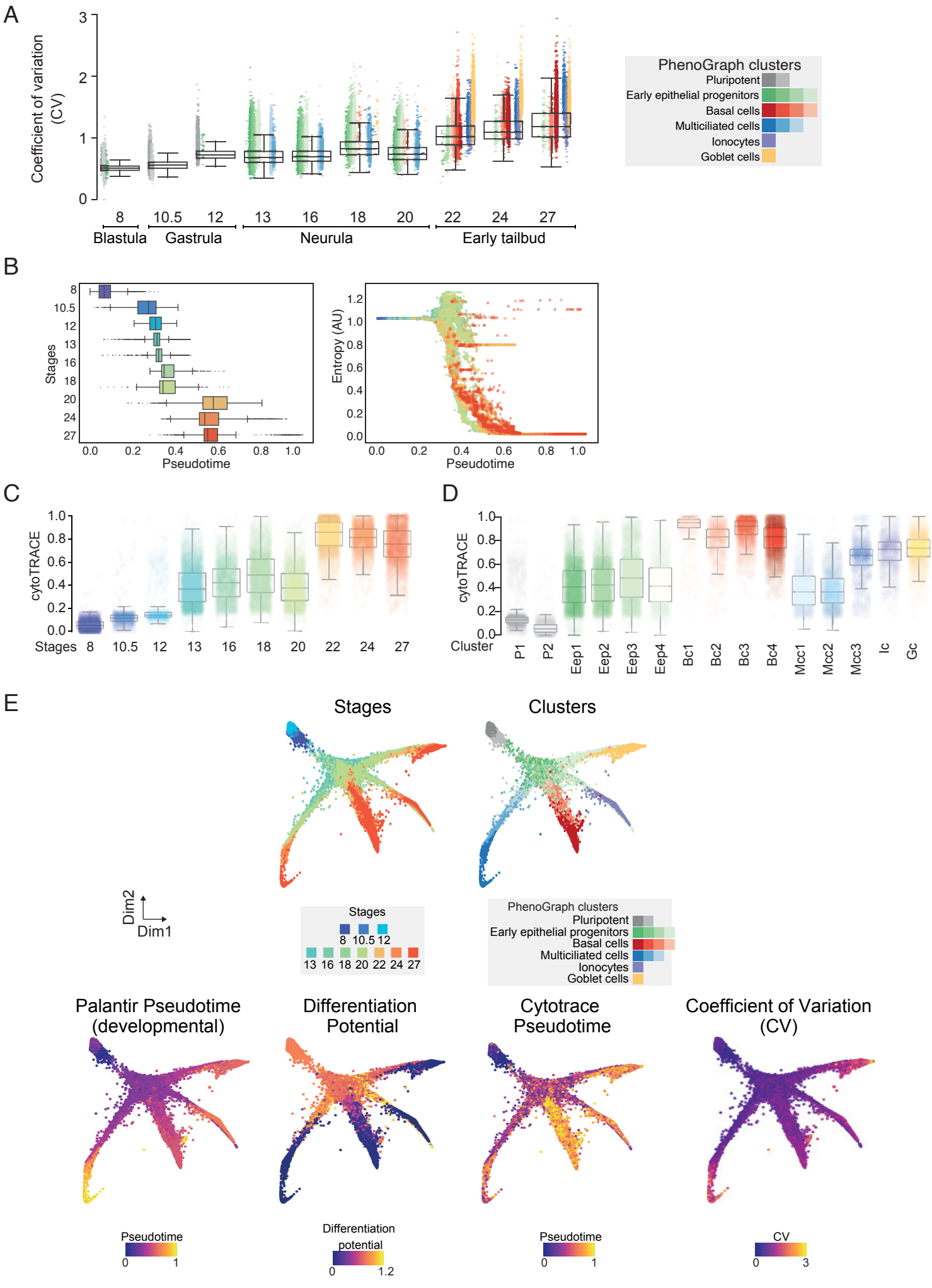


Fig. S9 Molecular features contributing to MCE developmental and cell-fate specification

(**A**) Boxplots comparing MCE developmental stages (x-axis) and subclusters (colored points) with the coefficient of variation (CV) for MCE PhenoGraph clusters. The CV increases over development (highest across early progenitors and basal cells), indicating heterogeneity and stochasticity as a hallmark to drive noise-induced differentiation.

(**B**) Boxplots comparing MCE developmental stages (left) and single-cell entropy (right) with developmental pseudotime ordering (x-axis).

(**C**) Boxplots comparing MCE developmental stages (x-axis) with cytoTrace pseudotime scores.

(**D**) Boxplots comparing PhenoGraph clusters (x-axis) with cytoTrace pseudotime scores.

The low cytoTrace scores (pluripotent, early epithelial and ciliated progenitors) indicate high stemness scores, while late-stage cell types have high cytoTrace scores indicating a differentiated state.

(**E**) Overview of different molecular features contributing to MCE developmental and cell-fate specification, including MCE developmental knn graph overlaid with MCE developmental stages, PhenoGraph clusters, Palantir developmental Pseudotime, Differentiation potential, CytoTrace Pseudotime and Coefficient of variation (CV).

Supplementary figure 10

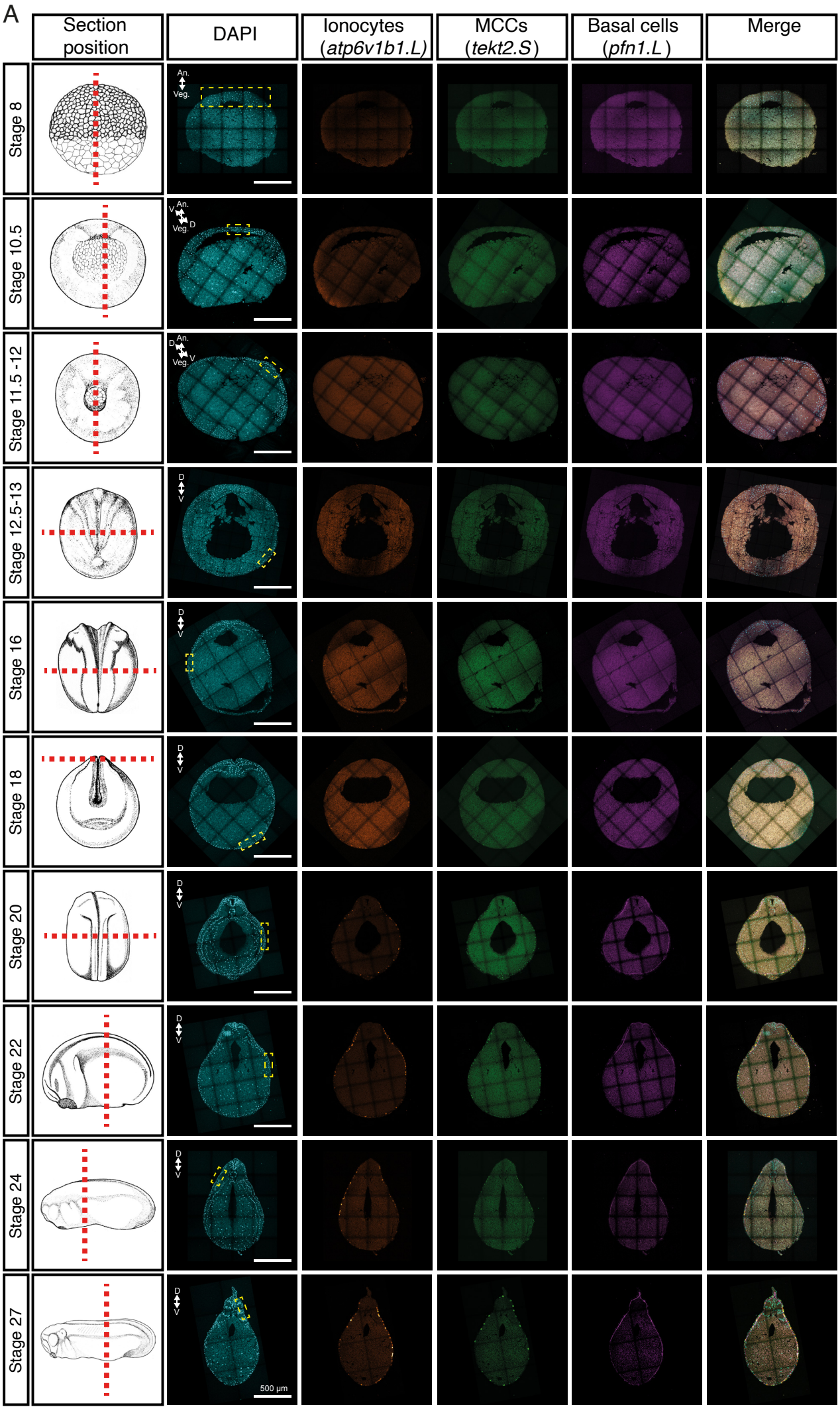


Fig. S10 *in-situ* HCR for cell types over MCE development stages

(**A**) *in-situ* hybridization chain reaction (HCR) and validation of lineage inference over 10 development stages of the embryonic epidermis in embryos, marking ionocytes (*atp6v1b1.L*, orange), multiciliated (*tekt2.S*, green) and basal cells (*pfn1.L*, magenta). The nuclei are marked by DAPI staining (cyan). The yellow rectangles indicate zoomed-in regions shown in Figure 6A. Images represent maximum intensity projections of transverse sections. Scale bars: 500μm

Supplementary figure 11

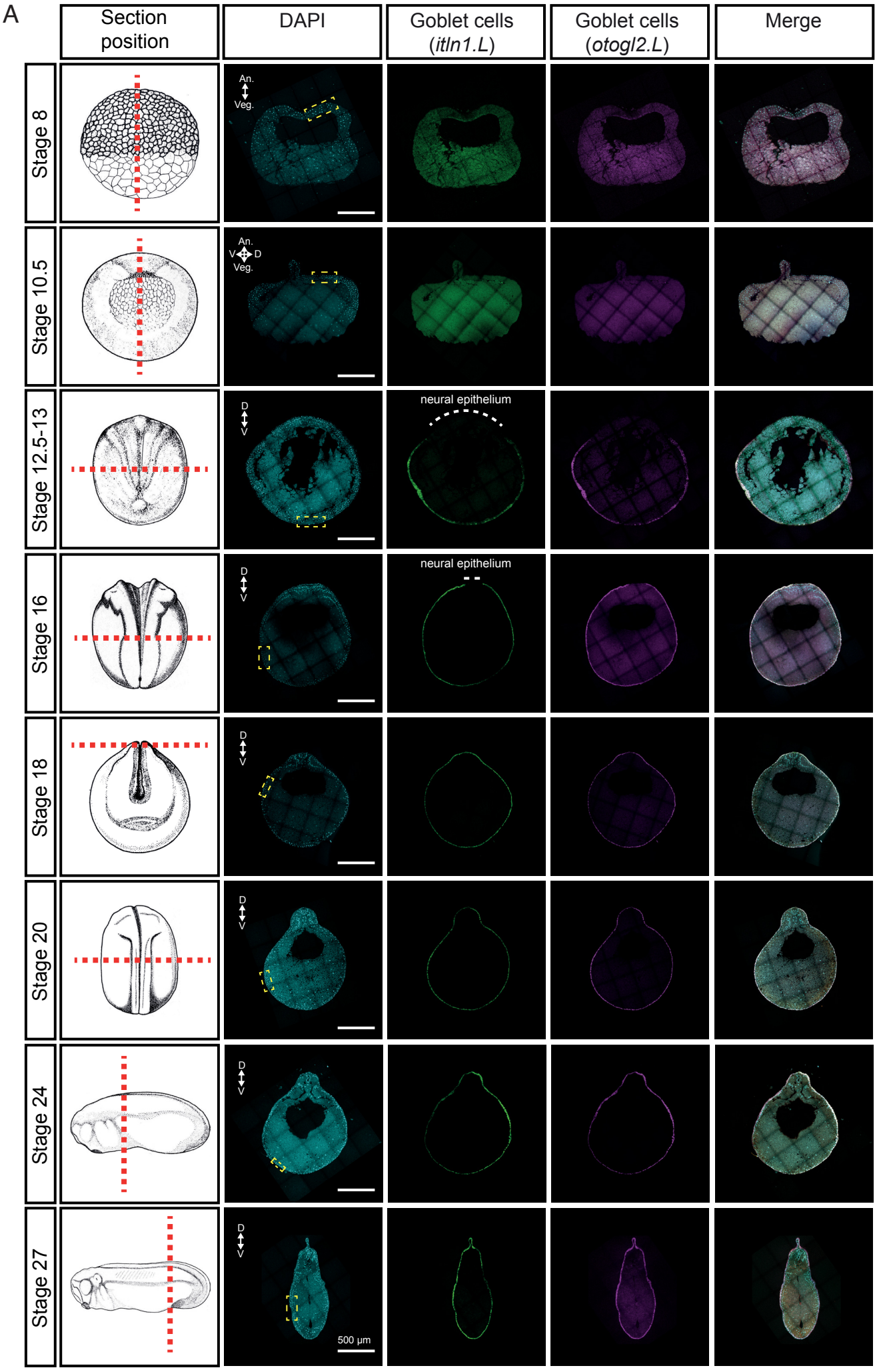
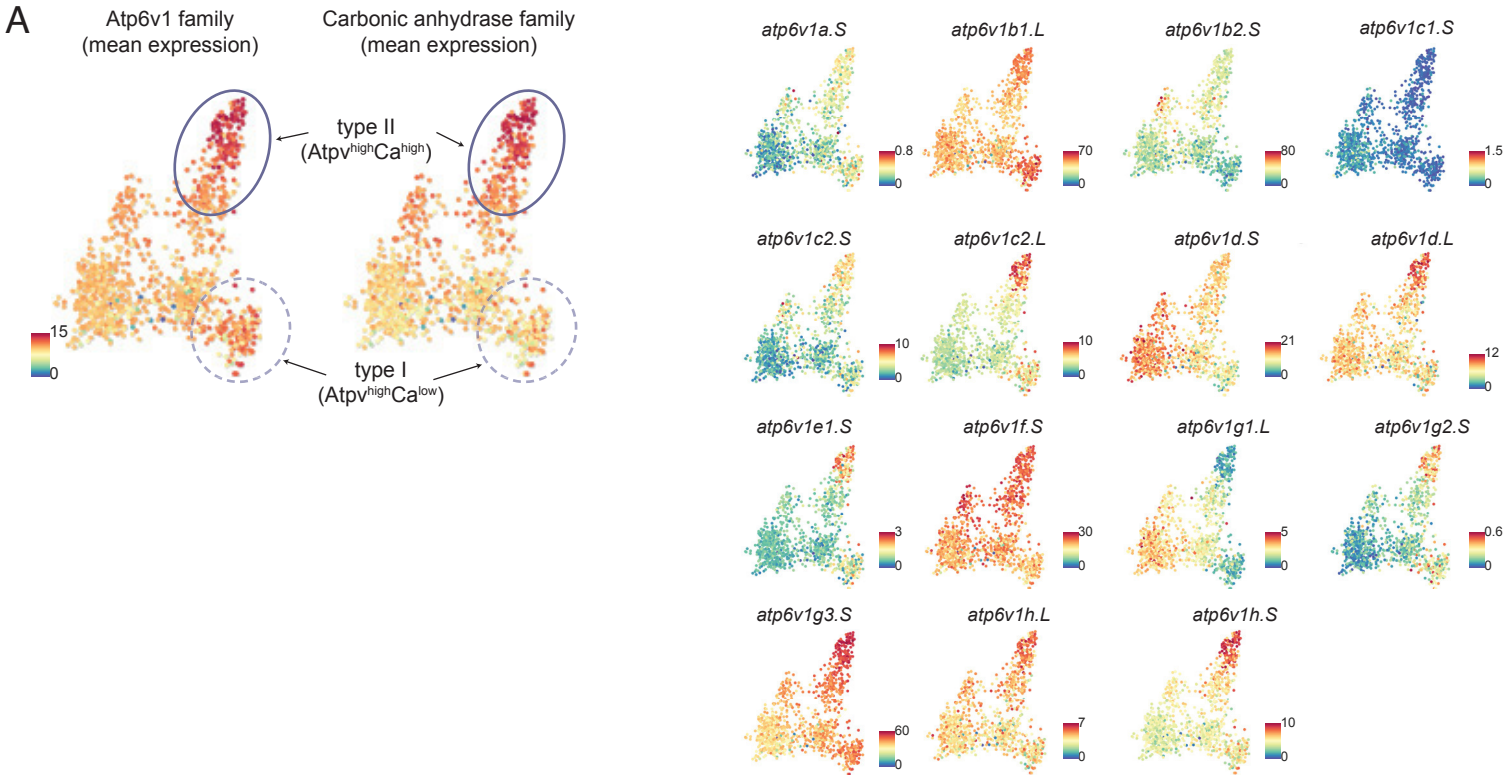
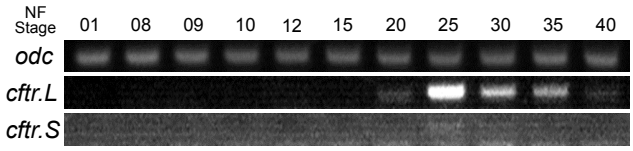


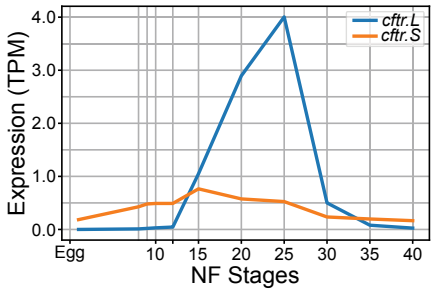
Fig. S11 *in-situ* HCR for goblet cell markers over MCE developmental stages

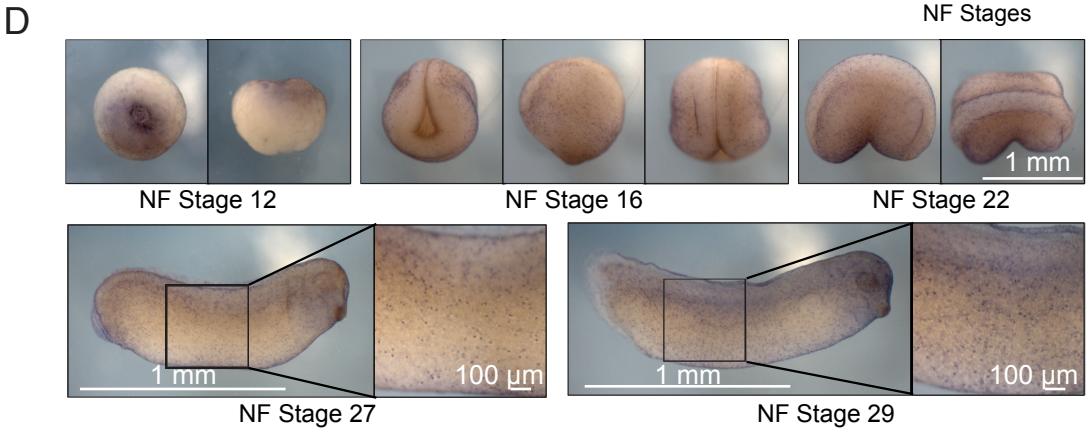
(**A**) *In-situ* hybridization chain reaction (HCR) of 8 developmental stages indicating goblet cell differentiation marked by intelectin (*itln1.L*, green) and otogelin (*otog.L/mucXS/otogl2.L*, magenta); nuclei are marked by DAPI staining (cyan). The yellow rectangles indicate regions zoomed-in in Fig. 7B. Images represent maximum intensity projections of transverse sections. Scale bars: 500μm

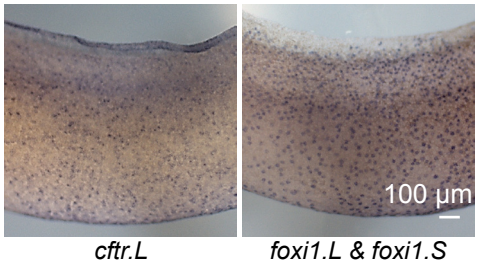
Supplementary figure 12











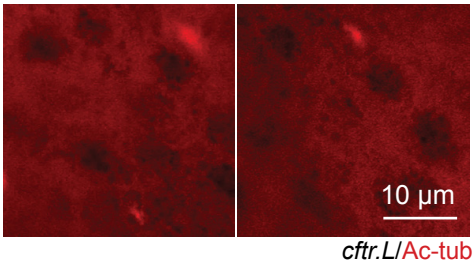


Fig. S12 Ionocyte subpopulations and marker gene expression

(**A**) Low dimensional visualisation of ionocyte subpopulations, overlaid with the mean expression of Atp6v1 and carbonic anhydrase family members (same as Fig. 5D), indicating type-I ($Atp6v1^{high}Ca^{low}$) and type II ionocytes ($Atp6v1^{high}Ca^{high}$). Individual expression of Atp6v1 and carbonic anhydrase family members (right panel). The scale bars indicate the scaled imputed expression of respective markers.

(**B**) Expression of *cftr.L* and *cftr.S* during MCE developmental stages, indicating peak expression at tailbud stages.

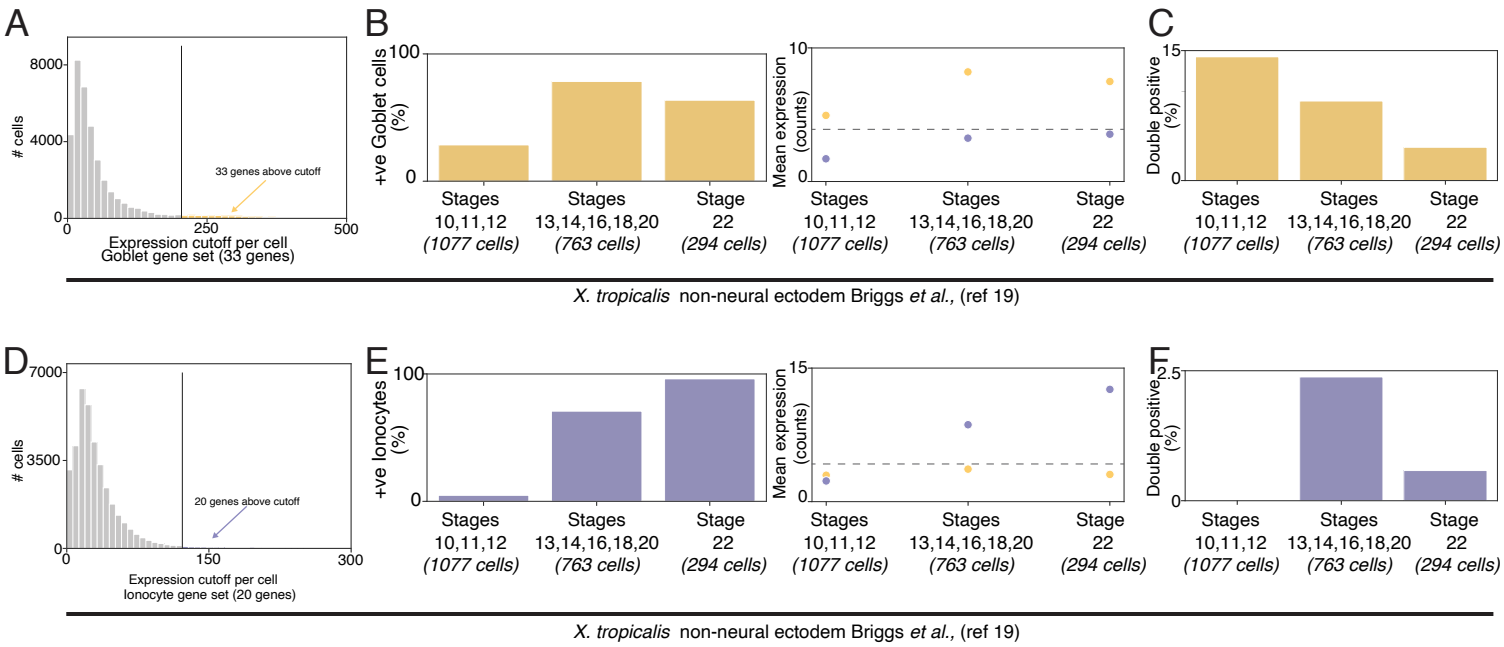
(**C**) Bulk expression from RNA-seq (TPM: transcripts per million reads) of *cftr.L* and *cftr.S* during MCE differentiation. RNA-seq captures the initial increase in expression during neurula stages and peaking at tailbud stages.

(**D**) Whole-mount *in-situ* staining for *cftr.L* across NF stages 12, 16, 22, 27 and 29, with zoomed-in views of stage 27 and 29. Scale bars: 1mm and 500 μm respectively.

(**E**) Zoomed-in view of whole-mount *in-situ* staining for *cftr.L* and *foxi1.L*-marked ionocytes in tailbud stage embryos. Scale bars: 100μm

(**F**) Distinct marking of ciliated cells (red, acetylated-tubulin (Ac-tub)) and ionocytes (black, *cftr.L*) in tailbud stage embryos. Scale bars: 10μm

Supplementary figure 13



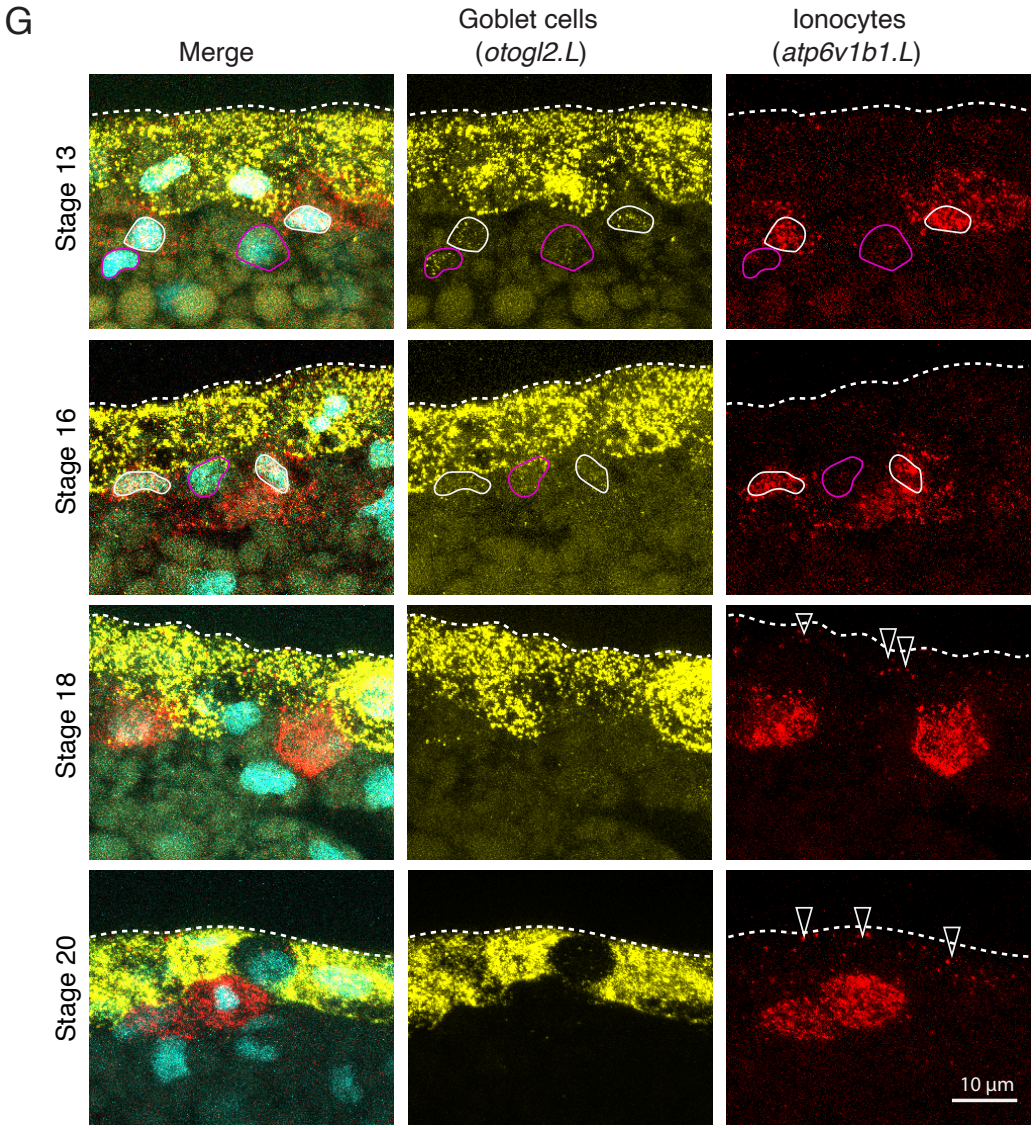


Fig. S13 Comparison with *X. tropicalis* atlas and secretory cell subpopulations

(**A**) Expression cutoff to select annotated goblet cells robustly expressing goblet gene signature (33 genes) across *X. tropicalis* development atlas (non-neural ectoderm lineage)

(**B**) The number, percentage and mean expression (counts) of goblet cells increases over *X. tropicalis* development atlas (non-neural ectoderm lineage).

(**C**) The number, and percentage of goblet cells expressing ionocyte signature genes (double-positive cells).

(**D**) Similar to (A), but for ionocyte gene signature (20 genes), across *X. tropicalis* development atlas (non-neural ectoderm lineage)

(**E**) Similar to (B), but representing the increase in the number, percentage and mean expression (counts) of ionocytes gene signature (20 genes)

(**F**) Similar to (C), but representing the number, percentage of ionocytes expressing goblet cell signature genes (double-positive cells).

(**G**) *In-situ* HCR for goblet cell (*otogl2.L*) and ionocyte marker (*atp6v1b1.L*) across neurula stages. White outlines mark nuclei of double-positive cells within the sensorial layer; pink outlines mark nuclei of cells with low levels of *atp6v1b1.L* and *otogl2.L* expression; white arrowheads mark double-positive cells within the superficial layer. Nuclei are marked by Dapi (cyan). A white dotted line marks the outer boundary of the superficial layer. Scale bars: 10μm.

Supplementary figure 14

A

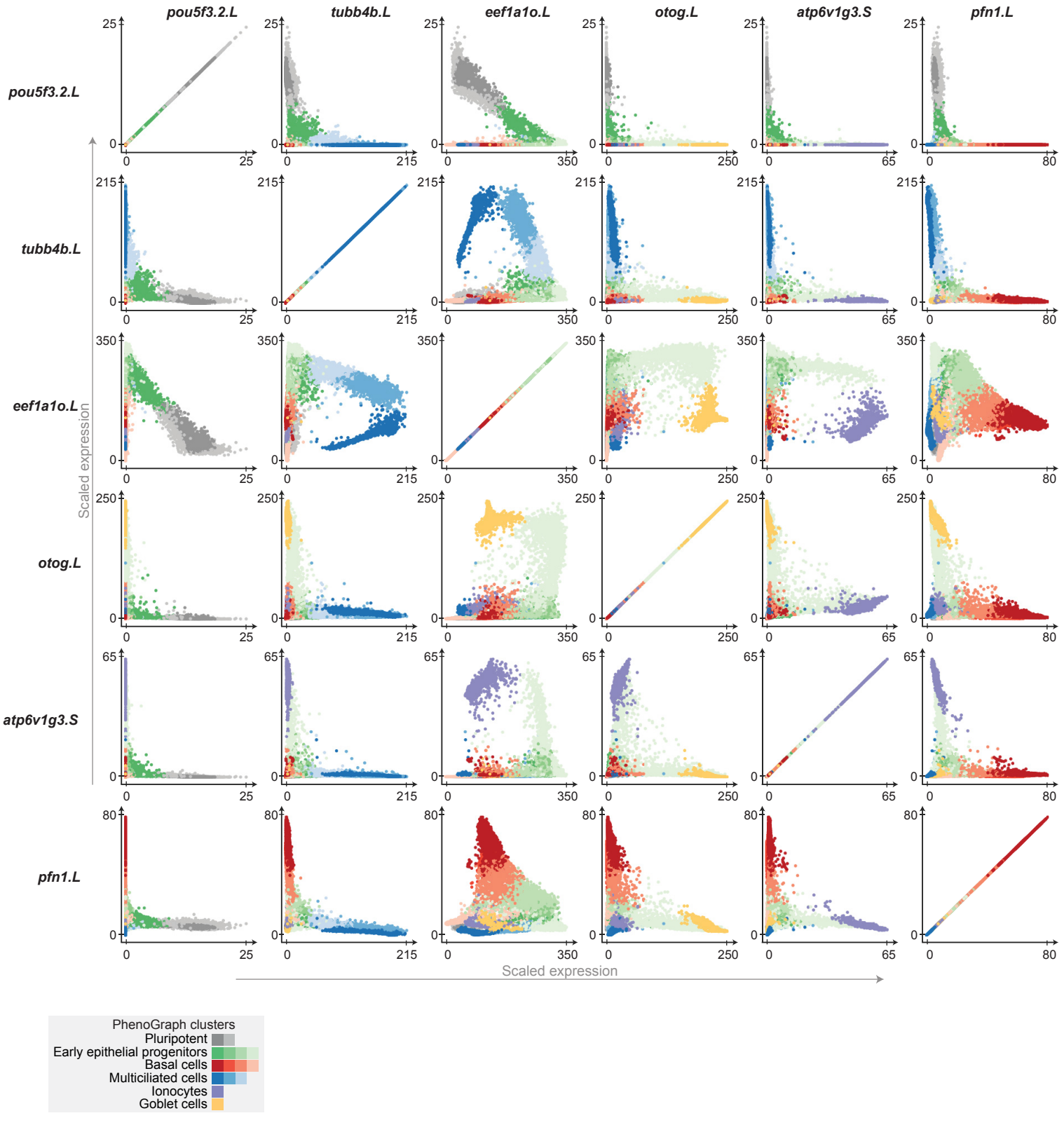


Fig. S14 Marker gene correlations over MCE developmental clusters
(**A**) Cell type marker correlation across individual cells and clusters over MCE development. The early progenitor clusters express multiple cell-types markers at lower levels (notably secretory genes), indicating multi-lineage bias. The x-axis and y-axis indicate the scaled imputed expression of respective markers.

Supplementary figure 15

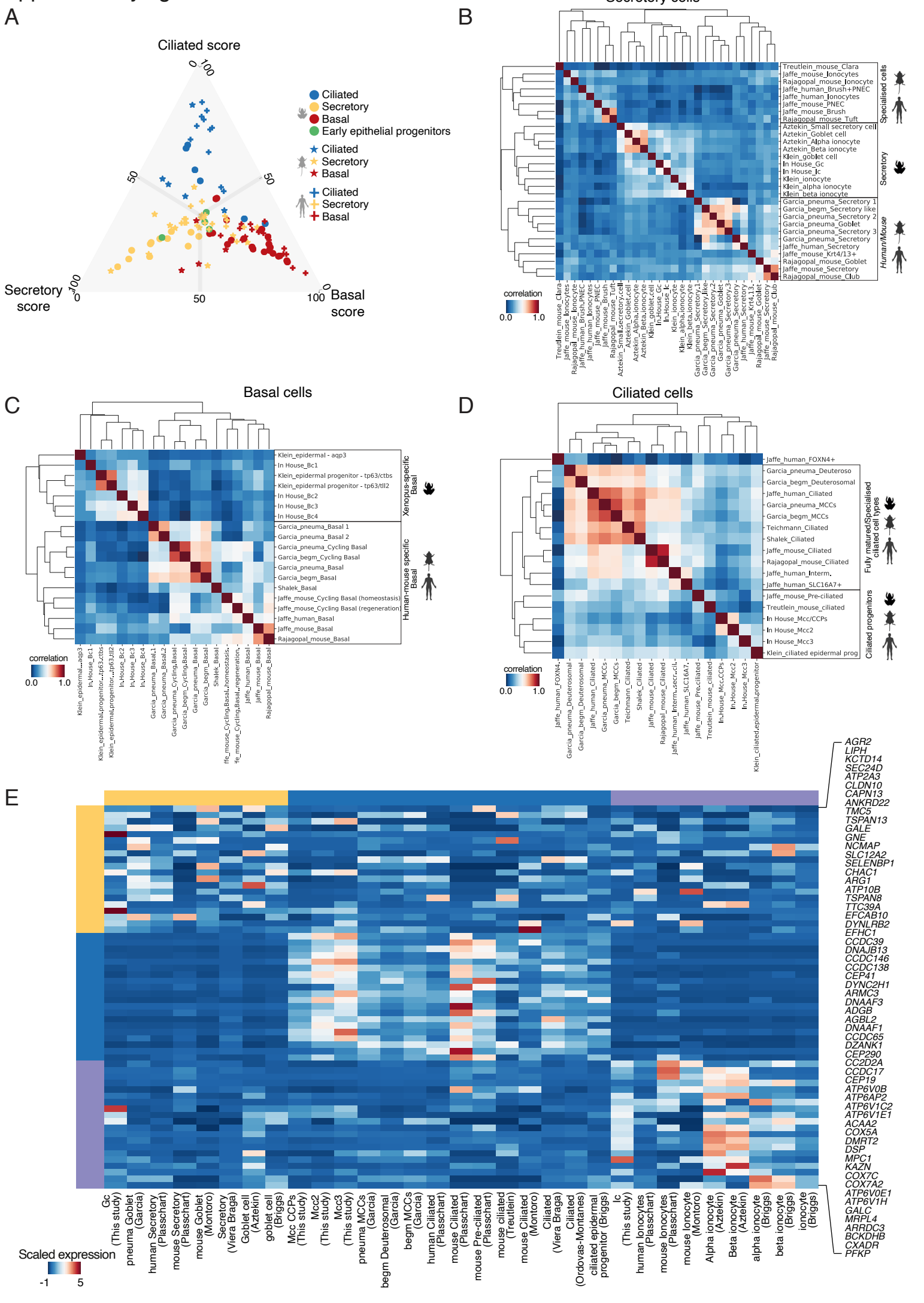


Fig. S15 Classification and correlation of cell types across single-cell atlases

(**A**) Classification and scoring of the different individual cell types from *Xenopus*, mouse and human single-cell studies, based on expressed and orthologous gene set. We plot the ciliated, basal and secretory enrichment scores (Fig. 10A) to annotate cell type specificity and accuracy to score respective cell types. The colors indicate the different cell types, while shapes indicate the different species.

(**B**) Correlation of secretory cell types from *Xenopus*, mouse and human single-cell datasets. The *Xenopus* secretory cell types are grouped separately from their higher vertebrate counterparts, likely due to a lack of mucins and other secretory molecules. The specialised secretory cells from higher vertebrates also form a separate cluster.

(**C**) Correlation of basal cell types from *Xenopus*, mouse and human single-cell datasets. The *Xenopus* basal cell types are grouped separately from the higher vertebrate basal cell subtypes, owing to a distinct mode of specification and specialised function across higher vertebrate basal cell types

(**D**) Correlation of ciliated cell types from *Xenopus*, mouse and human single-cell datasets. The ciliated progenitors (irrespective of species) are grouped together between vertebrates and distinctly separated from mature ciliated cell types, indicating a conserved expression module driving ciliogenesis.

(**E**) Marker gene expression (human ortholog DE genes, y-axis) in the different MCE cell types across species (x-axis). Cell type annotations are highlighted in groups (yellow, blue and purple bars across rows and columns).

The scale bars in panels (B-E) indicate the z-scaled mean expression of respective markers.

Supplementary tables

Table S1: List of valid cell barcodes and associated metadata.

Table S2: Statistics of cells passing good quality threshold, mapping percentage and number of differentially enriched genes per stage.

Table S3: List of highly variable genes by stages.

Table S4: List of highly variable genes by clusters.

Table S5: Two dimensional embedding coordinates for individual cells.

Table S6: List of TFs by stages and clusters.

Table S7: List of signalling genes by stages and clusters.

Table S8: List of core genes (published gene lists) by stages and clusters.

Table S9: List of different computational comparison by stages and clusters.

Table S10: Statistics on cell cycle status within stages, clusters and contingency matrix.

Table S11: List of reagents, antibodies and experimental resources used in this study.

Table S12: List of datasets used for evolutionary comparison, alongside species profiled and single-cell technologies utilized in respective studies.