Zebrahub – Multimodal Zebrafish Developmental Atlas Reveals the State Transition Dynamics of Late Vertebrate Pluripotent Axial Progenitors

Merlin Lange,[1,*] Alejandro Granados,[1,#] Shruthi VijayKumar,[1,#] Jordao Bragantini,[1,#] Sarah Ancheta,[1,#] Sreejith Santhosh,[2] Michael Borja,[1] Hirofumi Kobayashi,[1] Erin McGeever,[1] Ahmet Can Solak,[1] Bin Yang,[1] Xiang Zhao,[1] Yang Liu,[1] Angela Detweiler,[1] Sheryl Paul,[1] Honey Mekonen,[1] Tiger Lao,[1] Rachel Banks,[1] Adrian Jacobo,[1] Keir Balla,[1] Kyle Awayan,[1] Samuel D'Souza,[1] Robert Haase,[3] Alexandre Dizeux,[4] Olivier Pourquie,[5] Rafael Gómez-Sjöberg,[1] Greg Huber,[1] Mattia Serra,[2] Norma Neff,[1] Angela Oliveira Pisco,[1] & Loïc A. Royer [1,*]

[1]Chan Zuckerberg Biohub, San Francisco, USA.
[2]University of California, San Diego, USA.
[3]Cluster of Excellence "Physics of Life", TU Dresden, Germany.
[4]Institute of Physics for Medicine Paris, ESPCI Paris-PSL, Paris, France.
[5]Harvard Medical School, Boston, USA.

*Correspondence: merlin.lange@czbiohub.org, loic.royer@czbiohub.org
[#] Equal contribution.

ABSTRACT

Elucidating the developmental process of an organism will require the complete cartography of cellular lineages in the spatial, temporal, and molecular domains. We present Zebrahub, a comprehensive dynamic atlas of zebrafish embryonic development that combines single-cell sequencing time course data with light-sheet microscopy-based lineage reconstructions. Zebrahub is a foundational resource to study developmental processes at both transcriptional and spatiotemporal levels. It is publicly accessible as a web-based resource, providing an open-access collection of datasets and tools. Using this resource we shed new light on the pluripotency of Neuro-Mesodermal Progenitors (NMPs). We find that NMPs are pluripotent only during early axis elongation before becoming exclusively mesodermal progenitors. We attribute this restriction in NMP cell fate to emerging morphodynamic features that compartmentalize tissue motion.

INTRODUCTION

The development of a fertilized egg into a multicellular organism requires the coordination of various processes including genetic regulation, cellular organization, and tissue morphodynamics. A major challenge for developmental biology has been to understand the lineages of individual cells as they organize at various scales to form tissues and organs. The investigation of cell lineages in embryos started over a century ago with the use of light microscopes to visually track cells[1,2]. A key milestone was the publication in 1983 of the complete cell lineage of nematode[3,4]. However, the reconstruction of complete lineage trees is still limited to organisms with highly stereotypic cell fates and good optical access[5]. Advances in fluorescence light-sheet microscopy have enabled *in-toto* imaging of embryonic development at cellular resolution using zebrafish, mice, and drosophila, allowing for the reconstruction of specific lineages based on cell positions over time[6–10]. Single-cell RNA sequencing (scRNAseq) technologies have provided complementary means of tracking cell lineages based on their transcriptomes over developmental time[11–20]. Here, we combine light-sheet microscopy with scRNAseq of individual developing zebrafish embryos to construct a comprehensive multimodal atlas of cell lineages and molecular states at unique spatiotemporal resolution. We provide the data along with the analytical tools as an integrated and publicly available resource: Zebrahub. Using Zebrahub we investigated the Neuro-Mesodermal Progenitors (NMPs) – a conserved population of pluripotent axial elongation progenitors that generate skeletal muscles and posterior neural cells[21–26]. These cells play a crucial role in vertebrate axis elongation; however, their developmental dynamics remain unclear. Using a combination of multimodal techniques, we have elucidated the differentiation dynamics of NMPs throughout the whole process of vertebrate axis elongation.

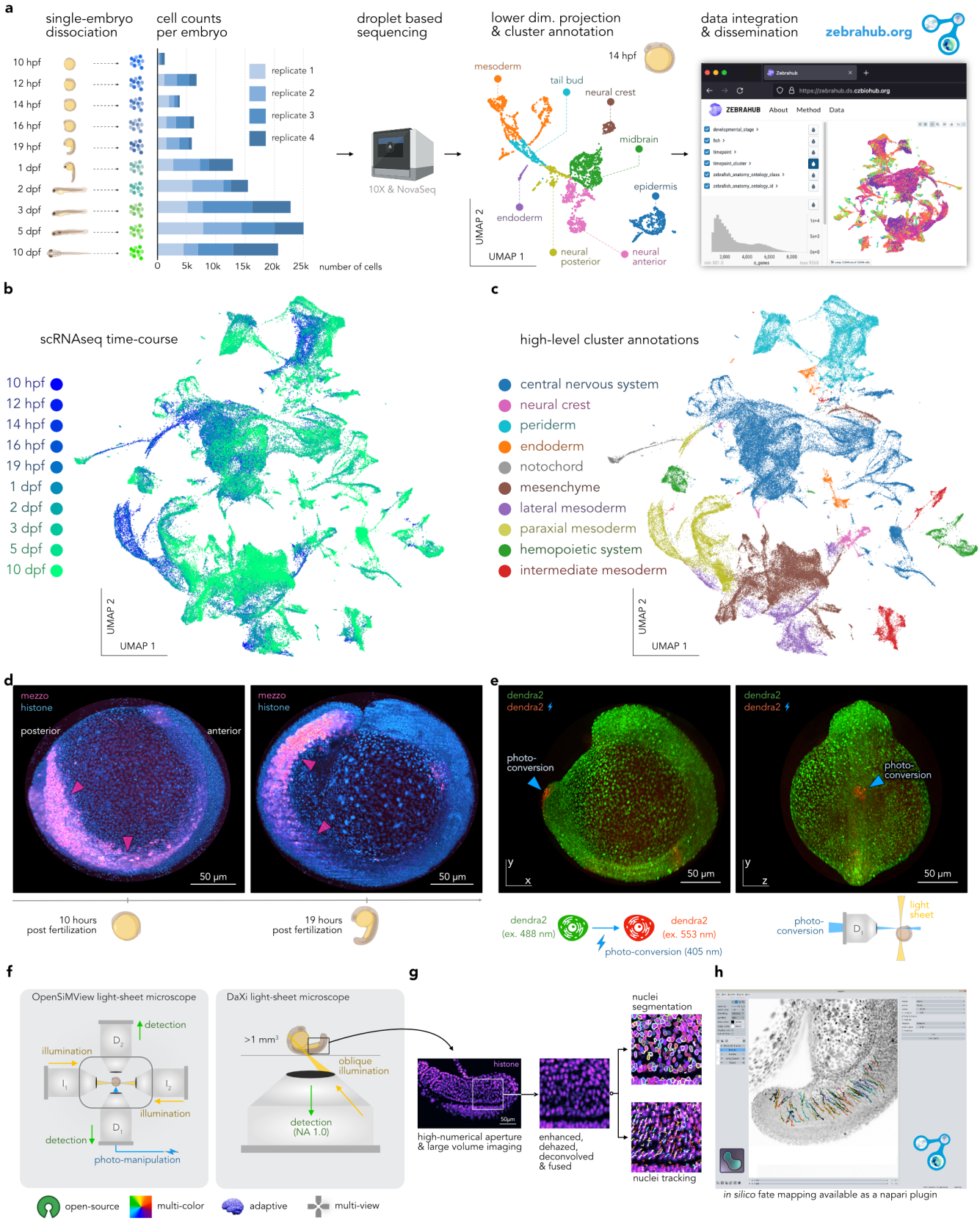


Figure 1. Zebrahub – Multimodal Zebrafish Developmental Atlas. (a) Zebrahub's scRNAseq experimental pipeline for single-zebrafish-embryo single-cell RNA sequencing across developmental stages (dpf: day post fertilization, hpf: hours post fertilization). From left to right we show: First, the number of cells per time point and replicate. Second, an example UMAP for a

(Figure 1 continued on next page)

14 hpf embryo transcriptome colored by cell type (3862 cells total). Third, Zebrahub's web portal allows the exploration of scRNAseq data online. **(b)** UMAP representation of the entire dataset of 40 single-embryo transcriptomes colored by developmental stage (120k total cells). **(c)** Same but colored by annotated cell types. **(d)** Maximum intensity projections of a 12-hour light-sheet live-imaging of a zebrafish embryo expressing *histone*-mCherry (cyan) and *mezzo*-GFP (magenta). Arrows indicate the *mezzo*-positive cells. **(e)** Similar imaging but on an embryo expressing the photoconvertible protein Dendra2. Photoconverted cells are in red (blue arrows). **(f)** Overview of the two light-sheet microscopes used in this study: OpenSiMView –two detections and two illuminations multi-view light-sheet microscope (left), and DaXi – a high-resolution, large field of view and multi-view single-objective light-sheet microscope (right). **(g)** Computational pipeline for image processing and nuclei tracking applied to images of a histone-mCherry zebrafish embryo acquired with our DaXi microscope (see Methods). **(h)** Screenshot of the *in silico* fate mapping napari plugin available as part of Zebrahub (Methods and Supp. Fig. 15).

RESULTS

Time-resolved single-embryo deep scRNAseq dataset of zebrafish development. Zebrahub's sequencing resource consists of a dataset containing 120,444 cells from 40 individual zebrafish embryos and larvae spanning 10 developmental stages from end-of-gastrulation embryos to 10-day larvae (10-, 12-, 14-, 16-, 19-, 24-hours post-fertilization, as well as 2-, 3-, 5- and 10-days post-fertilization). To capture single-cell RNA sequencing variability across individuals, the cell dissociation protocol was optimized to avoid embryo pooling thus achieving non-pooled single-embryo single-cell RNA sequencing (Fig. 1a, Supp. Fig. 1 & Methods). Four fish were sequenced for each of the ten developmental stages thus obtaining 40 individually sequenced embryos. Using strict quality control criteria (Supp. Fig. 2 and Methods) we obtained a minimum of one thousand high-quality single-cell transcriptomic profiles for the early developmental time points and more than 20 thousand cells at later stages (Fig. 1a – left). Importantly, we were able to obtain a minimum of 3,000 genes per cell (25,000 UMIs per cell) for developmental stages before 24 hpf – highlighting the depth of our dataset.

Cell type clustering and annotation. To visualize the transcriptomes for these 120,444 cells we generated uniform manifold projections (UMAP[27]) for each developmental stage (Fig. 1a – center and Supp. Fig. 3) and the whole dataset (Fig. 1b). We observed branching patterns that recapitulate the progression of embryonic development, with cells taken from early embryos located at the beginning of branches and cells obtained from older juveniles positioned towards the outer areas (Fig. 1b and Supp. Video 1). A total of 529 cell clusters (Leiden clustering) were annotated by examining enriched genes and cross-referencing with the published literature as well as the zebrafish database ZFIN[28] (Supp. Table 1). The different cell types identified were then grouped into higher-level tissue types (Fig. 1c and Supp. Video 2). The resulting projections for the different developmental stages reveal an increase in the number of cell types over time consistent with cell differentiation during development (Supp. Fig. 4). Our entire dataset, including projections and annotations, is available for download and online browsing through a user-friendly portal at zebrahub.org (Fig. 1a – right).

Single-cell image-based fate mapping during zebrafish development. To complement the sequence-derived transcriptomic state information, the individual cellular dynamics in developing embryos were imaged with an adaptive multi-view light-sheet microscope[29,30] (Fig. 1f, left). To accelerate the dissemination of this technology both the hardware design and control software are open-source (Supp. Fig. 5, Methods). This platform lets us image the spatiotemporal cellular dynamics among cells of interest using reporter transgenic lines, as illustrated using a pan-mesodermal marker (transgenic line *mezzo*:eGFP[31], see Fig. 1d and Supp. Video 3). However, transgenic reporter lines for all lineages of interest are not readily available nor easily engineered. To follow groups of cells and tissues of interest in the absence of reporter lines, our instrument can also label specific subsets of cells by photo-converting fluorophores such as Dendra2 (Fig. 1e, f, left).

High-resolution light-sheet imaging for lineage reconstruction. To image and accurately track the movement of cells in densely populated and rapidly developing embryos, we employed a single-objective light-sheet microscope (DaXi) capable of high-resolution imaging (1.0 NA) over large volumes (> 1 mm^3) without the need for time-consuming sample preparation protocols[32] (Fig. 1f - right, Supp. Fig. 6). An ubiquitous histone reporter line driving the expression of mCherry[33] was used to image the nuclei of cells. Following acquisition, images were processed and enhanced to facilitate the identification and tracking of cell nuclei (Fig. 1g and Supp. Video 4). Our cell-tracking pipeline uses a neural network to detect cell nuclei in the images and to delineate boundaries between cells. This information is then used to obtain nuclei segmentations and the corresponding tracks (Fig. 1g, Supp. Video 4 and methods). To help users navigate and use these very large datasets we developed an *in silico* fate mapping plugin for the multidimensional image viewer napari (Fig. 1h and Supp Fig. 15). This software allows users to conduct virtual fate mapping experiments by marking territories and following the trajectories of cells and their descendants. All imaging datasets and tools are accessible online at zebrahub.org and together with the scRNAseq datasets constitute a foundational multimodal resource to study vertebrate development.

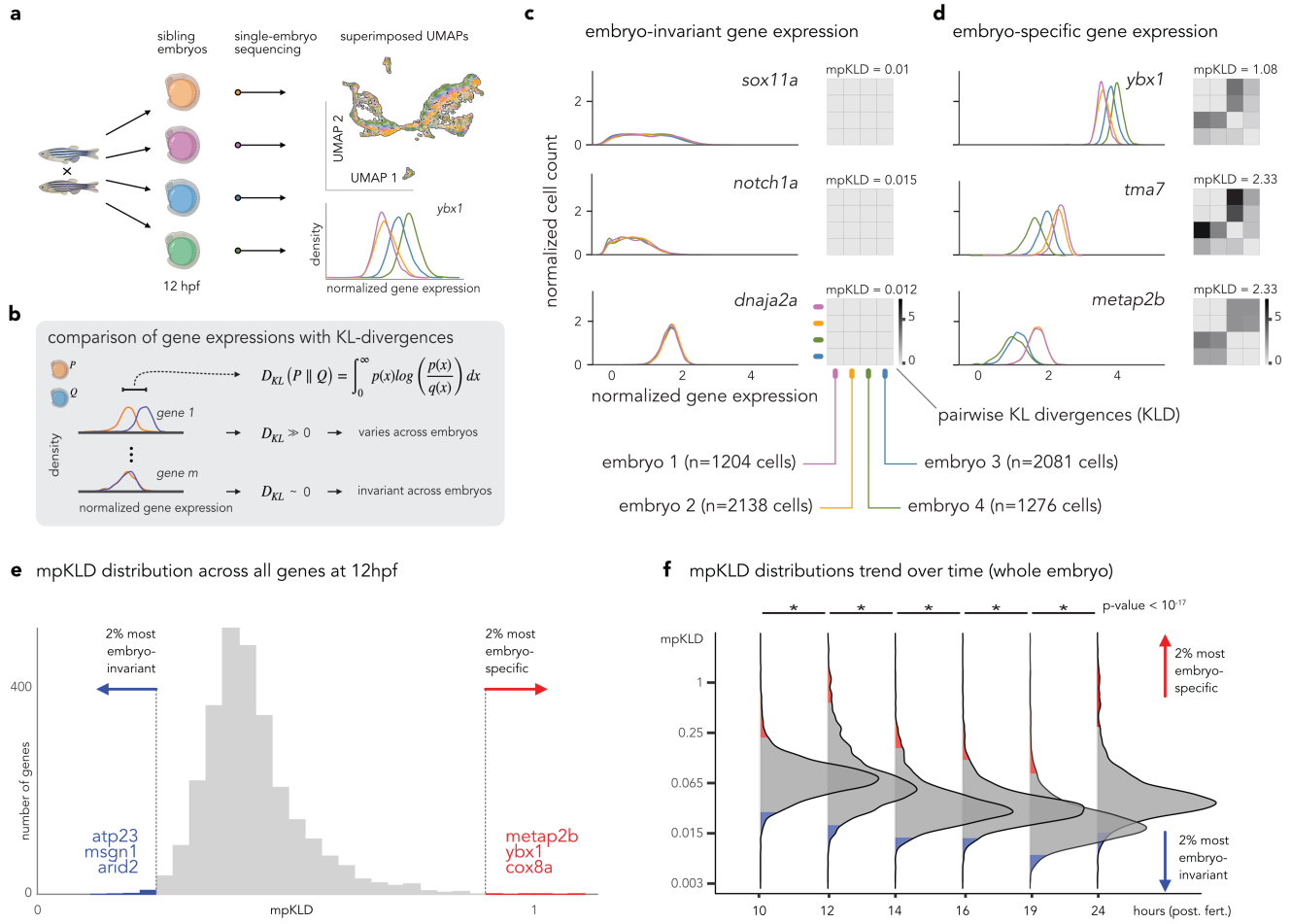
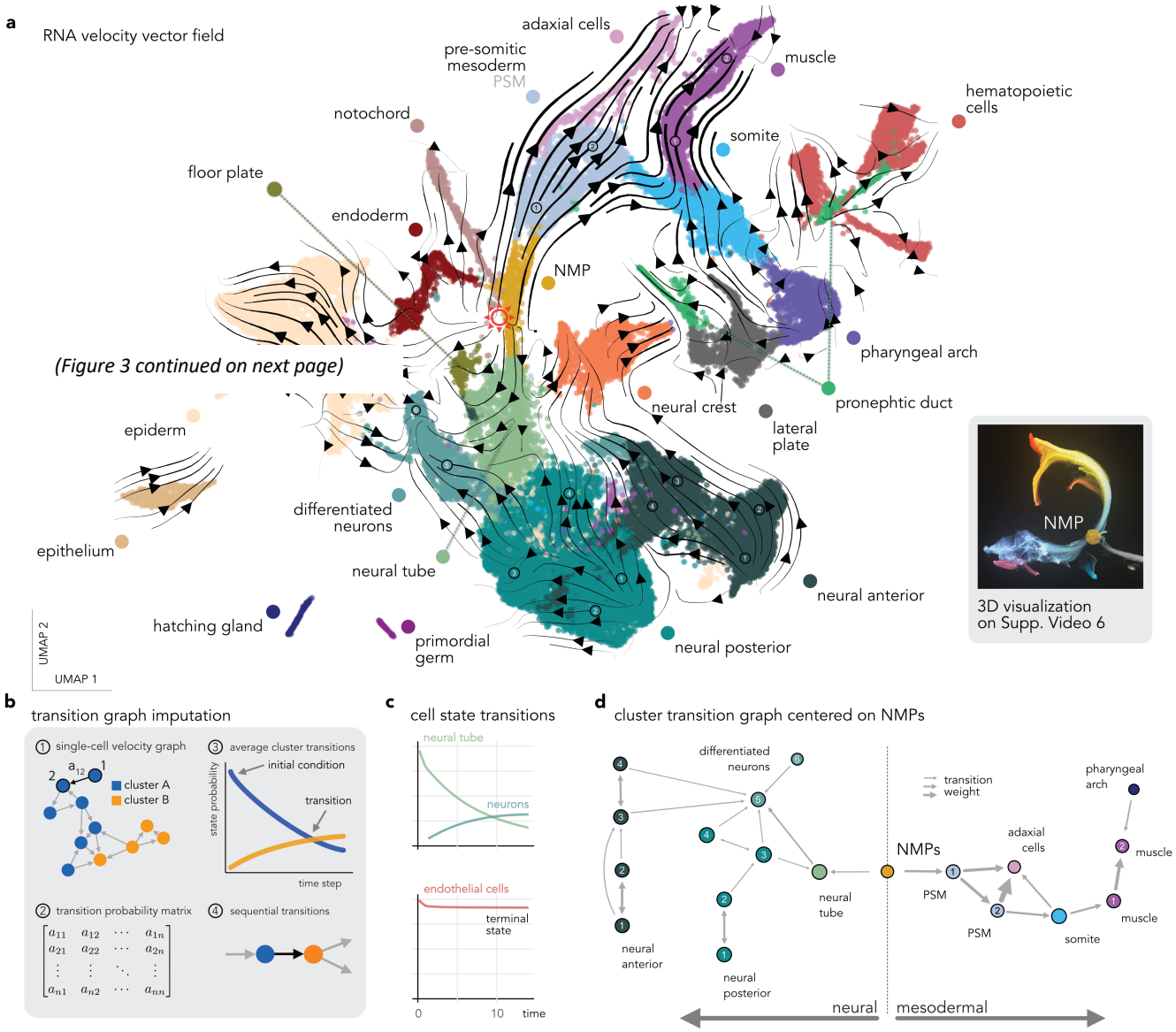


Figure 2. Inter-individual transcriptomic variability analysis. (a) Sequencing of sibling embryos enables the analysis of inter-individual gene expression variance. A UMAP of the single-cell data without batch correction (12 hpf, four embryos, 6699 cells total). Example of normalized expression of gene *ybx1* for four embryos (color coded). **(b)** We used the Kullback-Leibler divergence (KLD) to quantify the pairwise similarity between probability distributions, illustrating variant (top, KLD ≫ 0) and invariant (bottom, KLD ≈ 0) gene expression across embryos. **(c)** Example distributions for embryo-invariant genes (KLD ≈ 0) at 12 hpf: *sox11a*, *notch1a*, and *dnaja2a*. The x-axis indicates the log-normalized expression values, and the y-axis shows the density (relative number of cells). Heatmaps show all pairwise KLD values. We summarized these heatmaps per gene by the median pairwise KLD (mpKLD, values above the heatmap). **(d)** Example distributions for embryo-variant genes (KLD ≫ 0) at 12 hpf: *ybx1*, *tma7*, *metap2b*. **(e)** Histograms of mpKLD values scores for all cells at 12 hpf. A few examples of high and low mpKLD genes are shown. **(f)** Distributions of mpKLD values across all genes at 10, 12, 14, 16, 19, and 24 hpf. In **e** and **f** the top and bottom 2% of genes are indicated as colored areas (p-values < 10^{-15} are indicated with an asterisk).

Inter-individual transcriptomic variance between sibling embryos. The single-embryo resolution and high sequencing depth of our scRNAseq datasets provide a unique opportunity to study the variability of gene expression during development by comparing the single-cell transcriptomes of siblings (Fig. 2a). To quantify the differences in gene expression patterns between siblings, we developed a framework comparing gene expression distributions across cells, utilizing the Kullback-Leibler divergence[34] (Fig. 2b). A Kullback-Leibler divergence of 0 indicates that the two distributions are identical, while values greater than zero indicate increasing divergences between the distributions (Fig. 2b). For example, at 12 hpf, multiple genes showed seemingly identical expression patterns between siblings (*e.g., sox11a*, *notch1a* and *dnaia2a* in Fig. 2c), whereas other genes showed embryo-specific distributions (*e.g., ybx1*, *tma7* and *meta2b* in Fig. 2d). We computed all pairwise Kullback-Leibler divergences for each gene and summarized these using the median value (mpKLD for median pairwise Kullback-Leibler divergence, Fig. 2c and Fig. 2d gray heatmaps - Methods). At 12 hpf, most genes showed similarity across embryos, with a few genes exhibiting low mpKLDs – indicating nearly identical gene expression across embryos (bottom 2% of genes colored in blue in Fig. 2e). However, we also observed a tail of genes with high

mpKLDs – indicating varying gene expression across siblings (top 2% of genes colored in red on Fig. 2e). Applying the same method, we observed genes with identical or embryo specific expression at different time points and in specific tissues (Supp. Fig. 7). We confirmed that the mpKLD between siblings could not be simply explained by sampling error by showing that they are significantly different from those obtained from a null distribution (see Methods and Supp. Fig. 8a). Comparable distributions of the mpKLDs were observed for the other early developmental stages (Supp. Fig. 8b). These results show that the single fish sequencing allows the comprehensive analysis of inter-individual transcriptomic variability between siblings.



(Figure 3 continued on next page)

Figure 3. Whole embryo RNA velocity and derived cell state transition map. (a) Projection (UMAP) of zebrafish embryo transcriptomes (early development, 10 to 24 hpf time points), color- coded by cell type. Arrow streams correspond to the average RNA velocities of single cells. **(b)** Workflow to reconstruct state transitions from RNA velocity graph. ① Weighted directed graph represents the expected transitions between transcriptional states from single cells. ② This graph defines a Markov process that can be represented as a matrix. ③ We can then estimate cluster transitions by simulating successive state transitions for groups of cells through the graph. ④ Finally, we summarize cluster-level transitions as a coarse-grained graph. **(c)** Examples of average transition probabilities for neural tube and endothelial cells. **(d)** A coarse-grain graph of cell state transitions centered around NMPs, which shows transitions between all cell states. The width of the arrows is proportional to the transition rates. Colors and numbers correspond to the annotation in **a**. Nodes with the same color indicate clusters of cells belonging to the same broad cell types.

Time-resolved analysis of transcriptomic variability between individuals shows that gene expression reaches the highest level of similarity among siblings at 19 hpf. To gain a better understanding of transcriptomic variability, we extended the analysis to six other time points (10 to 24 hpf) from early development. Interestingly, we observed an overall decrease in mpKLD, as shown by a shift in the distributions (Figure 2f – Two-Sample Kolmogorov-Smirnov Test across consecutive time points, p-value < 10^{-15}). The gene expression distributions across siblings reach a point of minimal gene expression variability at 19 hpf, with a rebound in variability at 24 hpf (Fig. 2f). This pattern was also observed when examining specific cell types (Supp. Fig. 9a). We speculate that the higher degree of similarity at 19 hpf reflects the tight control of gene expression at the end of the segmentation period[36].

RNA velocity recapitulates lineage formation and suggests the existence of late pluripotent progenitors. To gain a deeper understanding of gene expression dynamics during early developmental stages (10 hpf to 24 hpf), we used RNA velocity[37]. The computed vectors were projected on the annotated UMAP (Fig.3a and Supp. Video 5) and recapitulate the progression of embryonic time (Supp. Fig. 10a). For example, the RNA velocities clearly show the mesodermal differentiation trajectory, starting from pre-smitic mesoderm, through somites, to muscles (Fig. 3a, upper right). Interestingly, the cluster at the origin of the mesodermal branch, indicated by a red star, is also at the origin of a branch going to the neural tissue. The cells composing this cluster are enriched for both neural and mesodermal transcription factors (*sox2* and *tbxta*, respectively, see Supp. Fig. 10ab). These transcription factors are the molecular signature of the Neuro-Mesodermal Progenitors (NMPs) – a recently characterized population of late pluripotent axial progenitors that are believed to contribute to mesoderm and neural tissue post-gastrulation[38].

Cell state transition graph analysis confirms that the NMP cluster contributes cells to both mesoderm and neural tissue. To substantiate quantitatively these findings, we coarse-grained our RNA velocity data into a cell state transition graph representing the dynamics of differentiation between cell types and tissues (Fig. 3b). Modeling the single-cell transcriptome transitions from the RNA velocities as a Markov process allowed us to explore a variety of cell type trajectories[39] (see Methods and Fig. 3c). Cell states related to progenitor populations quickly transitioned into their corresponding differentiated states (e.g., neural tube to neurons in Fig. 3c – top). In contrast, matured or fully differentiated cell states did not transition into any other state (e.g., endothelial cells in Fig. 3c – bottom). The cell state transitions were recapitulated in a coarse-grain graph (Fig. 3d), supporting the notion that NMPs transition to both mesodermal and neural states. Next, we examined their molecular and spatiotemporal dynamics.

***Sox2/tbxta* gradient predicts the future transcriptomic state of NMPs.** We zoomed in on the NMPs and the adjacent clusters and re-projected the data on a new UMAP (fig. 4a, Supp. Fig. 10c). In this projection, the NMPs stay positioned between the presomitic mesoderm (PSM) and neural tube (Fig. 4a and Supp. Fig. 11). The RNA velocity analysis on the new projection showed the same mesodermal and neural flows originating from the NMP population (Fig. 4b and Supp. Video 6). Flows, originating from the NMP cluster, were observed towards the notochord (mesodermal tissue) and floor plate (neural tissue - Fig. 4b) – two cell populations located along the midline of the vertebrate's tail. Notably, we found a *sox2/tbxta* expression gradient for NMPs: cells expressing more *tbxta* were closer to mesodermal lineages whereas cells expressing more *sox2* were closer to neural cells (Fig. 4c). This result suggests the importance of *tbxta* and *sox2* transcription factors in pushing the NMPs toward the mesodermal or neural fates, respectively.

NMP population size decreases during early development. We quantified the relative proportion of sequenced NMP cells between 10 hpf and 24 hpf and observed a decrease over time with a stronger decline between 12 hpf and 16 hpf (Fig. 4d). To corroborate this observation, we used Hybridization Chain Reaction (HCR, see Methods) to stain *sox2* and *tbxta* mRNA in fixed embryos. We then imaged the samples, segmented the cells, and counted those that co-expressed *sox2* and *tbxta* (see Supp. Fig. 12a and Methods). This analysis was consistent with the scRNAseq-deriver quantification, confirming a reduction of NMPs between 12 hpf and 16 hpf (see Fig. 4e, f, g, and Supp. Fig. 12b)

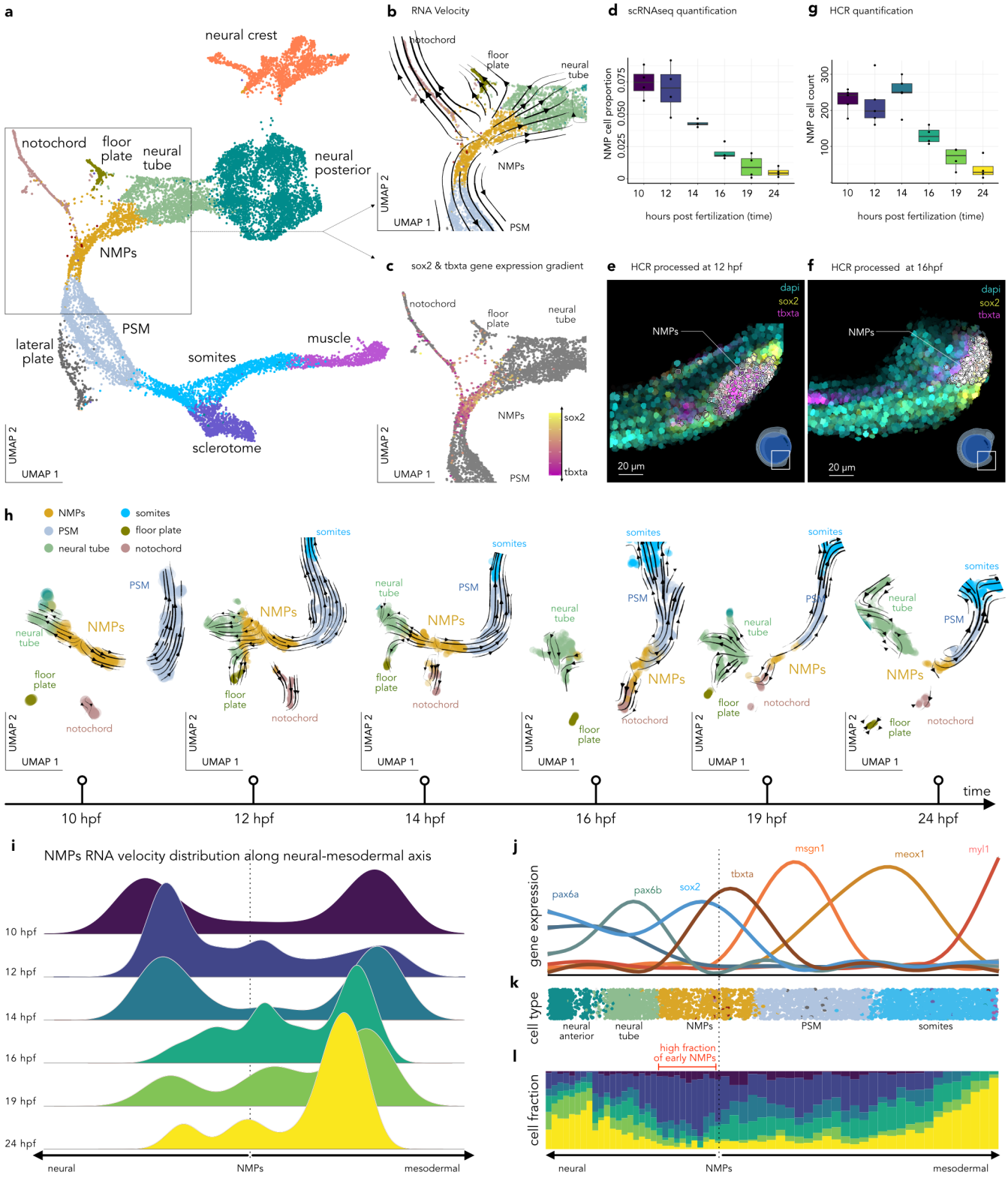


Figure 4. Transcriptional dynamics of NMPs during early development. (a) Reprojection (UMAP) of the NMPs and adjacent clusters, color-coded by annotated cell types. **(b)** Zoom in on the NMPs with color-coded cell types and superimposed RNA velocity streamlines. **(c)** Same projection but with cells color-coded by their *sox2/tbxta* expression ratio, stronger yellow indicates

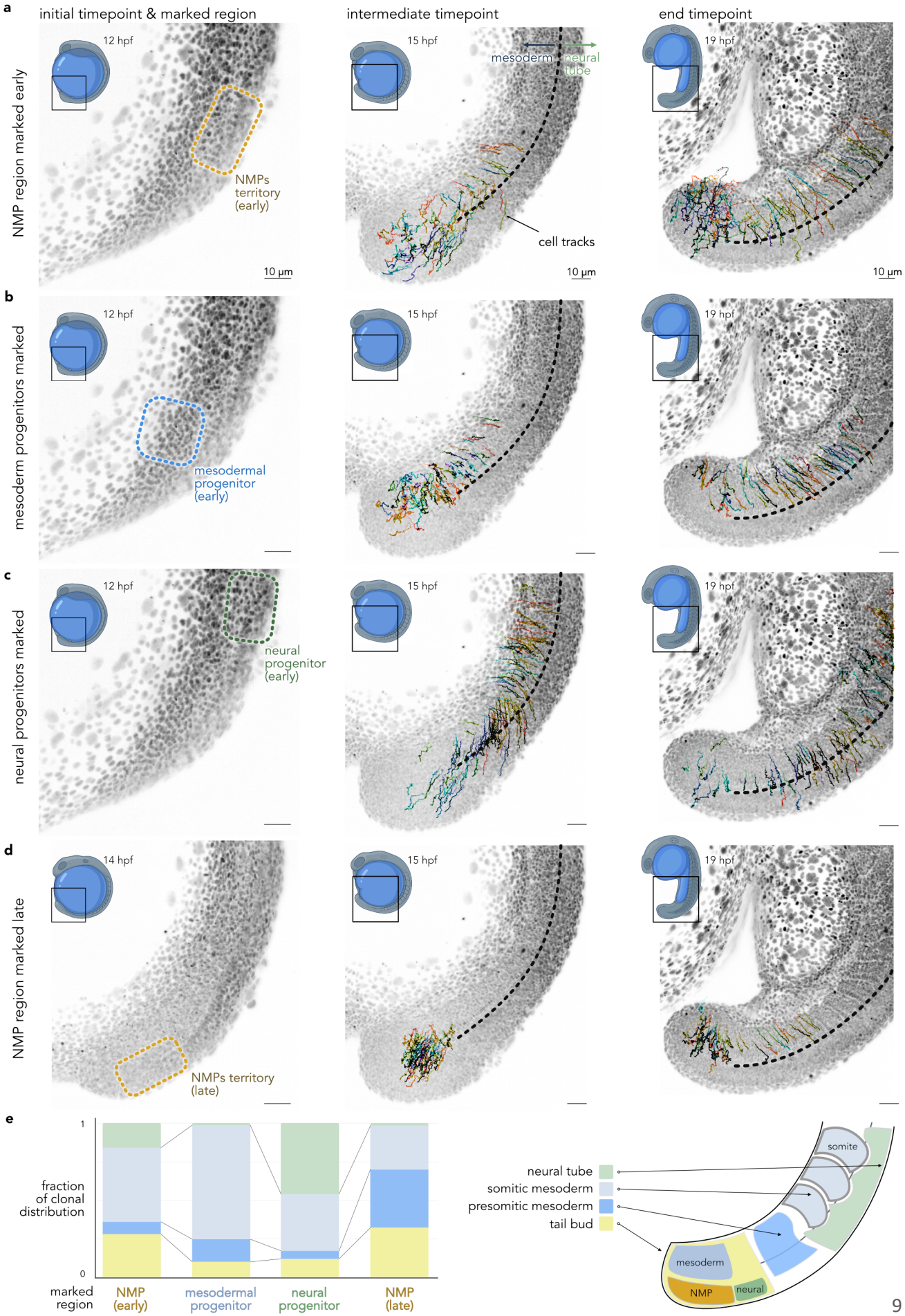
(Figure 4 continued on next page)

higher *sox2*, and stronger magenta indicates higher *tbxta*. **(d)** Quantification of the proportion of cells in the NMP clusters relative to the total number of cells, for each developmental stage. **(e)I** Expression of *sox2* (yellow) and *tbxta* (red) for segmented cells (12 hpf) derived by hybridization chain reaction (HCR). Cells co-expressing *sox2* and *tbxta*, presumptive NMPs, have their boundary in white. **(f)** Same as **e** but at 16 hpf. **(g)** Quantification of presumptive NMPs (cells co-expressing *sox2* and *tbxta*), obtained from the HCR image analysis. **(h)** Velocity trajectories are projected into a time series of aligned UMAPs centered around the NMPs and their adjacent clusters, colored by cell type. **(i)** Time-resolved quantification of the distribution of RNA velocity for NMPs along the neural-mesodermal axis. These are the normalized distributions of the magnitudes of individual RNA velocity vectors pointing toward either the neural or mesodermal tissues. **(j)** Pseudotime resolved scaled expression of genes involved in neural development (left axis) and mesodermal development (right axis). **(k)** Annotated cell-type visualized on the pseudotime axis rooted in the NMPs, recapitulating the neural (left) and the mesodermal (right) developmental transcriptomic trajectories. **(l)** Fraction of cells from different developmental stages visualized along the pseudotime axis.

Dissecting the molecular heterogeneity of NMPs. Next, we investigated the expression dynamics of genes associated with molecular pathways that are important in axial development. Amongst the transcripts specifically enriched in NMPs (Supp. Table 2) were genes involved in the control of body elongation (*cdx1a*, *lin28a*, *cdx4, gdf11*) [40–42] as well as genes related to FGF, RA, and Wnt signaling pathways, important for mesodermal and neural induction[38] (Supp. Fig. 13a). Interestingly, enriched genes showed a peak of expression at early stages of development (Supp. Fig. 13a 10, 12, 14 hpf). Remarkably, the well-known pluripotency factor *oct4*[43] was among the enriched genes at 12 hpf (encoded by *pou5f1*, arrow in Supp. Fig. 13a). We confirmed the specific expression of *oct4* in NMPs by comparing it with other tissues and confirmed that after 10 hpf, *oct4* expression is maintained at a high level in NMPs compared to other cell types (Supp. Fig. 13d). We also plotted genes that are important for the cell cycle[44] and observed an upregulation of these genes at 12 hpf, suggesting an increase in proliferation at this stage (Supp. Fig. 13b). As expected, we observed a striking co-linear activation of the hox genes indicating a temporal maturation of the NMPs as the body extends (Supp. Fig. 13c). We clustered the cells annotated as NMPs in a new UMAP and observed a small population of cells located between the early and the late cell clusters (Supp. Fig. 13e mixed population), with a high expression of *tbxta*/*sox2* and enriched for genes specific to the notochord lineages (*noto*, *ntd5*, *shha*). This observation echoes studies describing a population of progenitors in the vertebrate posterior territory that can contribute to the notochord[45–47]. Taken together our results indicate that NMPs are enriched in pro-mesodermal, pro-neural genes pathways and a pluripotency factor with the highest expression early in development.

Gene-expression dynamics suggest that early NMPs are pluripotent and late NMPs mostly contribute to the mesoderm. To gain a better understanding of the temporal dynamics of NMPs during early development, we re-analyzed the RNA velocity data at individual time points (10, 12, 14, 16, 19, and 24 hpf). At 10 hpf (the end of gastrulation, beginning of axial extension – Fig. 4h) the NMP population appears closer to the neural posterior cluster with RNA velocities directed towards both mesodermal and neural clusters. At 12 and 14 hpf, we observed a similar pattern: the two main RNA velocity flows originating from the NMPs are directed towards both PSM (mesoderm) and neural posterior clusters. However, at 16 hpf we observed a shift: the NMPs attached only to the mesoderm with their RNA velocities extending exclusively toward the mesoderm. A similar configuration is observed at 19 hpf and 24 hpf, with the RNA velocity of NMPs directed toward the mesoderm. Interestingly, at almost every stage the cell flow originating at the NMPs is also partly directed to the notochord (mesoderm) and a lesser extent to the floor plate (neural). To quantify this observation and prevent ambiguities due to the UMAP projections, we computed, at each developmental stage, the distribution of RNA velocities for NMPs along the neural-mesodermal axis (Fig. 4i). This analysis confirmed that during early stages (10, 12, and 14 hpf), NMP velocities point towards both mesoderm and neural tissue, whereas during later stages (16, 19, and 24 hpf), they primarily point towards the mesoderm.

Pseudotime lineage reconstruction supports NMPs' fate restriction. To further understand the developmental progression of NMPs along the neural-mesodermal axis, we reconstructed pseudotime developmental trajectories[48]. These trajectories recapitulate the expression of genes that play crucial roles in neural and mesodermal differentiation[49] (Fig. 4j). We also identified the key cell types in these trajectories (Fig. 4k) and visualized their distribution in developmental time (Fig. 4l). These findings were consistent with our previous observations, indicating that the majority of NMPs moving in the neural direction are early-stage cells, while the cells moving in the mesodermal direction are from all stages (Fig. 4k, l – red bar). Overall, our time-resolved scRNAseq dataset indicates that the fate of NMPs changes over time, with early NMPs, predicted to have the potential for both mesodermal and neural fates, whereas after 16 hpf they are restricted to the mesodermal lineage. Next, we reasoned that to confirm these transcriptomic insights we needed detailed spatiotemporal cell tracking information.



(Figure 5 continued on next page)

Figure 5. Reconstruction of the zebrafish embryonic posterior progenitor's territory via *in silico* fate mapping. (a) *In silico* fate mapping of the pluripotent early NMP territory. Left - selected territory at the initial timepoint, middle - intermediate timepoint showing tracks for selected cells, right – final time point. **(b)** *In silico* fate mapping of the axial mesodermal posterior progenitors. We observe that these cells mostly integrate into the mesoderm (74%). (c) *In silico* fate mapping of the axial neural posterior progenitors. These cells partly integrate the neural tube (46%). **(d)** *In silico* fate mapping of the late presumptive NMP territory. These cells integrate mostly the mesodermal tissue (70%). **(e)** Left – the proportion of the clonal distribution after photoconversion of the different progenitor territories: NMPs early (total number of cells = 75), mesodermal progenitors (total number of cells = 81), neural progenitors (total number of cells = 76), NMPs late (total number of cells = 53). Right – a schematic representation of the different axial progenitor territories.

High-resolution *in silico* fate mapping predicts time-dependent NMP fate restriction. We applied our nuclei segmentation and tracking algorithm to reconstruct high-resolution fate maps from single-cell tracks (Fig. 1g, and Supp. Fig. 14). To facilitate the visualization and analysis of this data, we developed a napari plugin that lets users select and follow groups of cells to determine their fate (Fig. 1h, Supp. Fig. 15 and Methods). We refer to this technique as *in silico* fate mapping. To determine the presumptive NMP domain, we mapped the spatiotemporal localization of NMPs by HCR based on their co-expression of *sox2* and *tbxta* (Fig. 3d, e and Supp Fig. 12). Consistent with the transcriptomic dynamics described earlier, this virtual fate mapping of the early NMP territory (<14 hpf) confirms that NMPs give rise to mesodermal and neural progeny (Fig. 5b and Supp. Video 7). We quantified the fraction of progeny and observed that 48% of cells went to the somites, 16% went to the neural tube, 28% went to the tail bud, and 8% went to the pre-somitic mesoderm (Fig. 5f). The ventrolateral domain posterior to the NMPs is the mesodermal progenitor territory (Fig. 5b and Supp. Video 8), the dorsomedial territory posterior to the NMP area is the neural progenitor zone (Fig. 5c and Supp. Video 9). Our *in silico* fate maps confirm the model of zebrafish posterior developmental specification domains[21]. At later time points (after 14 hpf, see Fig. 5d and Supp. Video 10) the fate maps of NMPs change, and they contribute only to mesodermal tissue. Thus, most of the progeny of the late NMPs integrates the pre-somitic mesoderm and somitic mesodermal tissues (respectively 38% and 28%, see Fig. 5f), or stay in the tail bud (32%). In conclusion, our virtual fate mapping of axial progenitors corroborates our finding that NMPs are effectively pluripotent only during the beginning of axial elongation.

***In vivo* fate-mapping via photo-conversion confirms time-dependent NMP fate restriction.** To complement our *in silico* fate mapping experiments, we performed photoconversion followed by live imaging of developing zebrafish embryos (Fig. Supp Fig. 16, Supp Fig. 17, and Methods). Using embryos ubiquitously expressing the photoconvertible protein Dendra2[50] we photoconverted the presumptive territory of NMPs at different developmental stages based on our previous observations (Supp Fig. 12). Early in development (10 hpf – Fig. 6a and Supp. Video 11) progeny from the photoconverted NMP territory were distributed in mesodermal and neural tissues (Fig. 6b and Supp. Video 11). However, when we photoconverted the NMPs' presumptive territory at 14 hpf (Fig. 6c and Supp. Video 12), the progeny only integrated into the mesoderm (Fig. 6d and Supp. Video 12). We computed the distribution of the clones by segmenting our 3D images and quantifying the NMP progeny in the manually annotated territory (mesodermal tissue and neural tissue, Fig. 6e and Fig. 6f). Early photoconversion resulted in 46.2% of progeny in neural tissues, and 53.8% of progeny in the mesodermal tissues (Fig. 6g). In contrast, photoconversion of NMPs at 14 hpf resulted only in 3.5% in the neural tube, and a large majority (96.5%) in the mesoderm (Fig. 6h).

Automated tracking of an NMP followed by manual curation. To confirm that NMPs are truly pluripotent cells, versus just a combination of lineage-specific progenitors (i.e., 'salt and pepper' distribution) we tracked single photoconverted putative NMPs. We then computationally backtracked cells that integrated at later stages the neural tube and mesoderm, searching for a common ancestor originating in the presumptive NMP territory. We then manually curated the resulting tracks – a tedious process because this tissue is highly crowded and dynamic (Fig. 6i). We were able to reconstruct the track of early NMP that divided and gave progeny that integrated into both mesodermal and neural tissues (Fig. 6j and Supp. Video 12). However, when we repeated this search for late photoconversion experiments, we could not identify any similarly pluripotent lineage originating from the presumptive NMP territory.

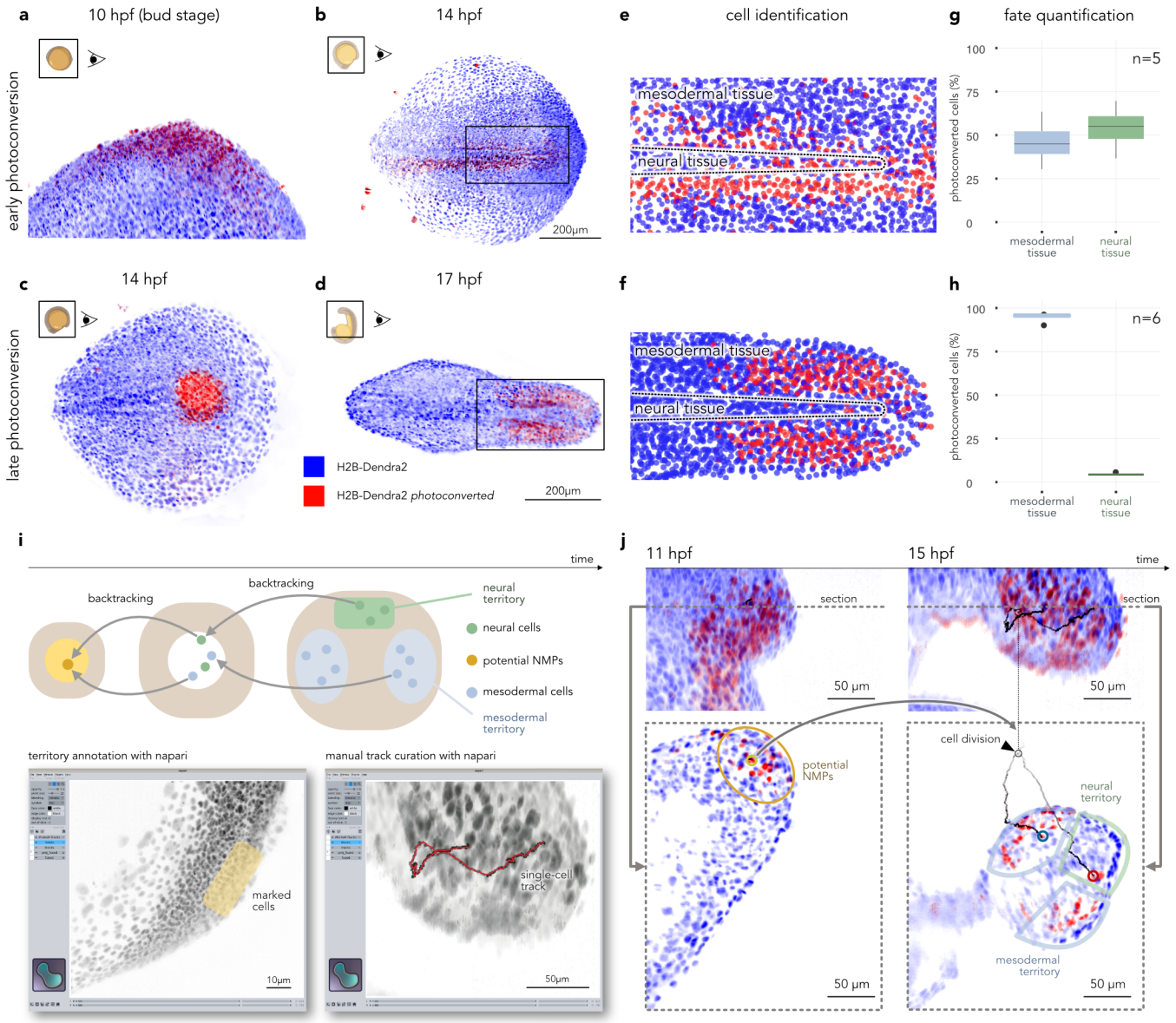


Figure 6. *In vivo* fate mapping of NMPs via photo-conversion and live imaging. (a) Photoconversion of the early presumptive NMP territory at 10 hpf. **(b)** Distribution of the photoconverted cells and their progeny at 14 hpf. **(c)** Photoconversion of the late NMP territory at 14 hpf. **(d)** Distribution of the photoconverted cells and their progeny cells at 17 hpf. **(e)I** Segmentation and identification of the photoconverted cells, in the area selected in **b** (rectangle). **(f)** Segmentation and identification of the photoconverted cells, in the area selected in **d** (rectangle). **(g)** Distribution of early photoconverted cells and their progeny with mesodermal (left) and neural (right) fates (n= 5 animals). **(h)** Same, but for late photoconverted cells and their progeny (n=6 animals). **(i)** Identification of NMPs from territory annotations and backtracking inferred from automatically generated tracking data (top panel). Screenshot of the napari plugin for *in silico* fate mapping, showing territory annotation (left), and 3d cell track curation (right). **(j)** Illustration of a single early NMP (left) dividing into two daughter cells. After cell division, the first daughter cell integrates the neural territory, while the second integrates into the mesodermal territory (right). The top panels are two maximal-intensity projections of the posterior zebrafish embryo. The bottom panels are section visualization.

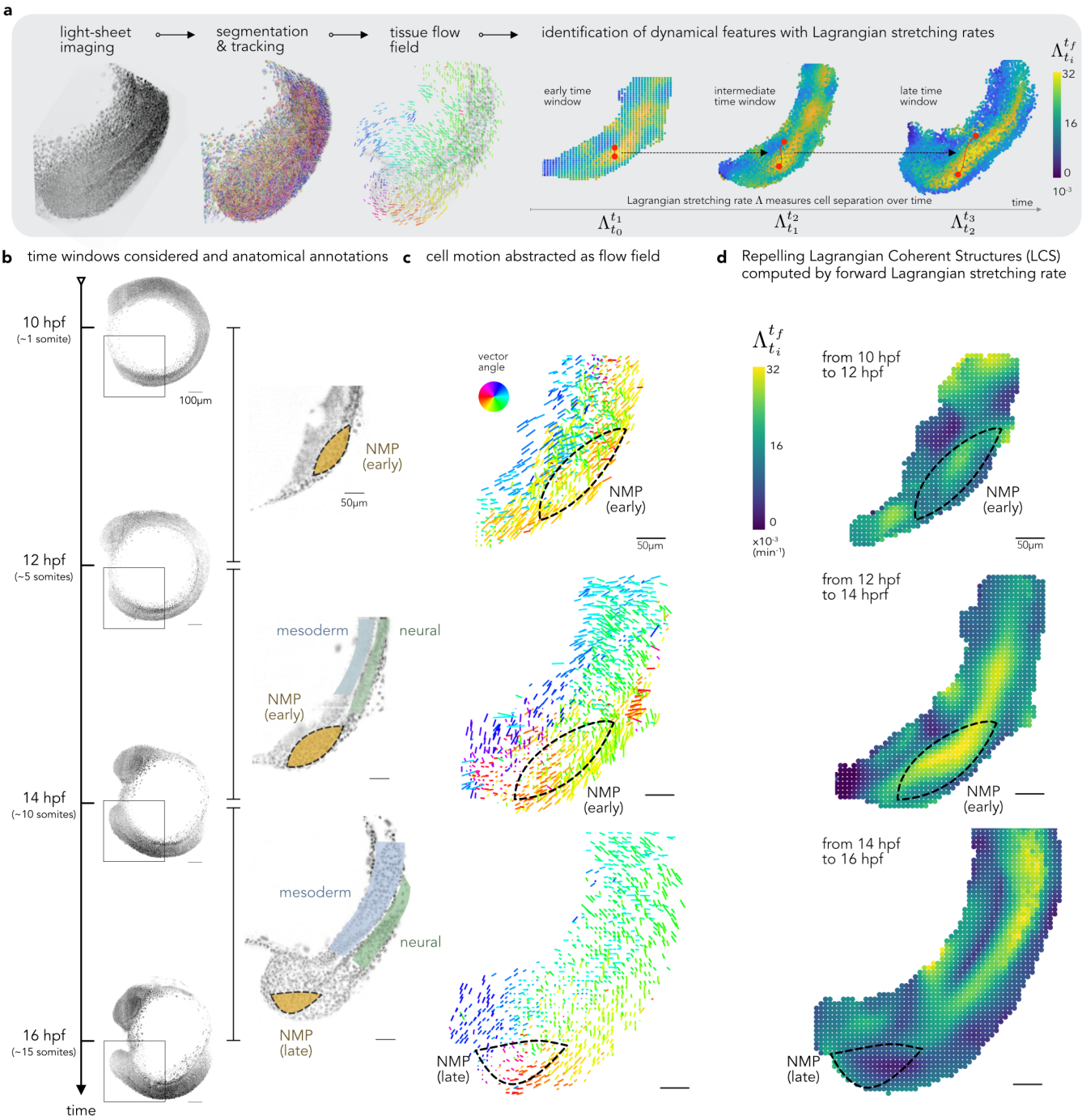


Figure 7. Tissue morphodynamics constrain the state transition of NMPs during vertebrate body axis elongation. (a) Sketch of the dynamic morphoskeleton approach used to identify robust dynamical features from multicellular flows. We focus on repellers – a type of Lagrangian coherent structure locating divergent cell trajectories. We identify repellers by computing the Lagrangian stretching rate ($\Lambda_{t_0}^{t_f}$, see Supp. Theory Note). **(b)** Illustration of the time intervals selected to compute the stretching rates, and anatomical annotations of the tail's territories (NMP, neural and mesodermal tissues). **(c)** Vector fields that represent the cells' motion during these time intervals, and dashed lines represent the presumptive NMP territory. **(d)** Lagrangian stretching rate (brighter shades correspond to more stretching i.e., divergent cell trajectories) and the emergence of repellers in the early NMP presumptive territory. This territory was ascertained by HCR and is outlined by a dashed line.

Tissue morphodynamics constrain the fate of NMPs. In addition to the molecular signatures identified, we tested whether spatial positions and distinct tissue deformation environments could affect NMPs' fate specification. To this end, we applied the recently developed dynamic morphoskeletons (DMs) framework[51]. This framework aims to identify Lagrangian Coherent Structures (LCS) which are distinct structures present in the embryonic cell flow that organize and compartmentalize cellular trajectories[51]. Of relevance for NMP specification are repelling LCSs (or simply repellers) that mark regions where cells distinctly separate over time. We identify repellers by computing the Lagrangian stretching of cellular flows derived from our tracking data (Fig. 7a, Methods, and Supp. Information). Our analysis revealed that during the early stages, from 10 to 12 hpf, the NMPs (localized via HCR, see Supp. Fig. 18) are located on a repeller (see Fig. 7b, c, and d). This repeller is characterized by neighboring cells rapidly separating from each other (Fig. 7a, see vectors Fig. 7c). The repelling structure was strongest between 12 and 14 hpf, a time during which the NMPs are effectively pluripotent (Fig. 7d - middle). This is consistent with the antagonist cell flows observed at the NMP territory (Fig. 7c, middle). In contrast, during the next time window when the NMPs' fate is restricted to the mesoderm (after 14 hpf), no repelling structure was present (Fig. 7d - bottom). At that point, the cell flows are coherent and solely directed toward the mesoderm (Fig. 7c - bottom). Therefore, cells in that region are now effectively prevented from reaching the neural tube. The same was true for the subsequent time window (Supp. Fig. 19). These results suggest that tissue compartmentalization is correlated with, and can potentially explain, the observed fate restriction identified via both sequencing and imaging.

DISCUSSION

We introduced Zebrahub, a collection of single-embryo single-cell temporally resolved transcriptome and image datasets for the comprehensive analysis of zebrafish development. All datasets, designs, and software are accessible for download, sequencing and imaging datasets can be explored interactively, and image-based lineages can be visualized using our napari[52] plugins. This resource includes a wide range of multidisciplinary innovations such as a protocol for high-efficiency single-embryo cell dissociation, a framework for studying the transcriptomic variability of individual embryos, an open-source design for adaptive multi-view light-sheet imaging and photo-manipulation, as well as computational tools for *in silico* fate mapping from light-sheet imaging data.

Furthermore, in contrast to previous scRNAseq studies[11–16,19], our resource encompasses a broader developmental period, attains non-pooled resolution at the single-embryo level, and our gene coverage per embryo is sufficient to measure transcriptome variability between individuals. In addition, we used state-of-the-art technology, including a high-resolution (1.0 NA) and large-volume (> 1mm3) light-sheet microscope[32], automated whole embryo cell tracking, and a theoretical framework[51], to investigate the lineages and cellular movements during late stages of vertebrate embryonic development. We could achieve this in highly dynamic and densely populated tissues, such as those found in the developing tail.

This unique combination of technologies helped us to comprehensively characterize the transition dynamics of axis elongation progenitors. We propose a model of the NMPs' fate dynamics that integrates information from transcriptomic states, image-derived lineages, and tissue kinematics (Fig. 8). During axis elongation, NMPs change their state from contributing to multiple lineages, principally mesodermal and neural, to exclusively contributing to the mesoderm. We found a peak of expression of key molecular pathways for axial progenitors' commitment (Wnt, FGF, and RA pathway) and the presence of the pluripotency factor *oct4* in early pluripotent NMPs. *Oct4* has been linked to the maintenance and differentiation of axial progenitors[43,53–55] and the reactivation of pluripotency in the mouse neural crest[56]. By combining *in silico* and *in vivo* fate mapping, we confirmed the fate restriction from pluripotent to exclusively mesodermal progenitors at around 15 hpf.

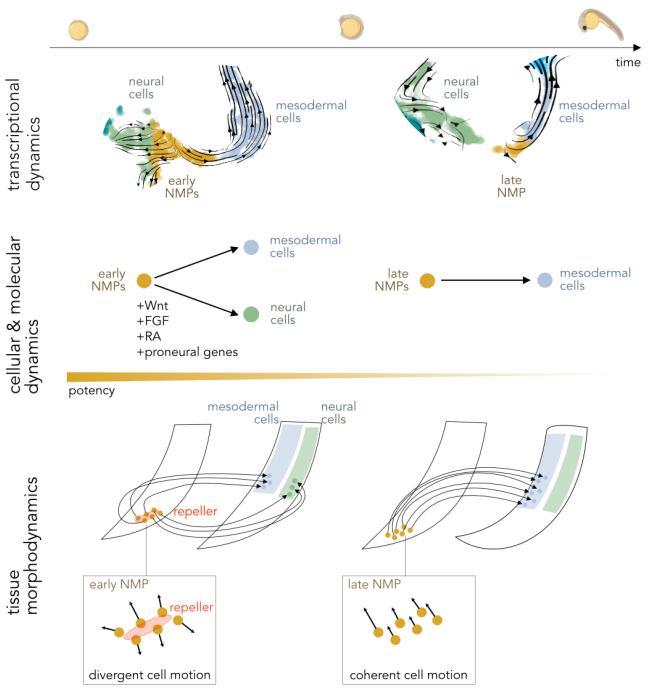


Figure 8. Summarizing model for our finding on the NMP fate dynamic during body axis elongation, using transcriptomic, molecular states, single-cell behavior, and tissue morphogenesis.

At the tissue level, we observed a repeller structure indicating rapidly diverging cell flows. The establishment of this repeller

structure, possibly caused by rapid cell proliferation, suggests the need for a pluripotent population of buffer progenitors with delayed differentiation. This delay would prevent these cells from prematurely differentiating and thus integrating a mismatched tissue. In contrast, when the axial progenitor territory is more stable (no repeller), and cell flow is coordinated, the NMPs remain pluripotent competent, but give only rise to mesoderm. This suggests the importance of tissue kinematics in modulating effective NMP differentiation fates. Interestingly, our data, and in particular our RNA velocity study, also suggests that the NMPs might be at the origin of other axial tissues, namely notochord and floor plate. Altogether, our results challenge the classical view of lineage commitment after gastrulation and support the idea of late and broad differentiation plasticity.

Zebrahub's collection of accessible datasets and tools is a foundational resource to the community. By combining imaging and sequencing datasets we joined the strengths of both modalities. Beyond NMPs, we expect that this resource will help the community to shine a light on other key developmental processes. Zebrahub is poised to expand with more developmental stages, novel multiomic datasets, and ultimately a consensus digital lineage reconstruction of multiple embryos. Integrating existing datasets, including from other species, will enable the creation of a comprehensive atlas of vertebrate embryogenesis, initiating a new era for developmental and evolutionary biology.

ACKNOWLEDGEMENTS

We thank the whole Data Science and IT teams of the Chan Zuckerberg Biohub for helpful discussion and support. We are grateful for the support, feedback, and proofreading received from S. Schmid, J. DeRisi, and S. Quake. We thank P. Keller for the critical reading of the manuscript and for help with our variant of the SiMView microscope design. We thank C. Guillot, L. Mahadevan, and S. Darmanis for initial discussions. We thank J. Weissman for his support and encouragement in pursuing zebrafish as the lab's model organism. We thank A. York and Calico Life Science LLC for developing and loaning the AMS-AGY V2.0 objective. We also would like to thank J. Maitin-Shepard and V. Jain for their generous help with Neuroglancer. Funding for this work was provided by Chan Zuckerberg Biohub San Francisco (CZB SF). We thank the CZB SF donors, Priscilla Chan and Mark Zuckerberg for their generous support.

AUTHOR CONTRIBUTIONS

M.L. and L.A.R. conceived the project and designed the experiments. L.A.R. supervised the project. M.L. led the experimental and computational work. S.V., M.B., and M.L. developed the cell dissociation protocol and performed the library prep. S.V. and M.L. performed the HCR experiments. A.De., S.P. and H.M. performed the sequencing. A.G., S.A., A.M., and A.O.P. performed computational analysis for scRNAseq data. A.G. and S.A. designed the single fish computational pipeline. A.G. performed the cell cluster transition graph. M.L., T.L., R.B., A.J., K.B., S.V. and X.Z. annotated the cells. B.Y., Y.L., and H.K. developed and built the microscopes. M.L. performed light-sheet imaging experiment. S.V. performed confocal imaging experiments with help from M.L.; J.B., A.C.S., and L.A.R. wrote the light-sheet image processing code. R.G.S. designed and built the imaging chamber. J.B. developed the *in silico* fate mapping software. R.A., H.K., and L.A.R. wrote the code for microscope control. M.S. and S.S. wrote the code for the dynamic morphoskeletons framework and performed the analysis with J.B., M.L., and L.A.R. J.B. analyzed the imaging data. M.L., A.G., S.A., A.O.S., and L.A.R. analyzed the sequencing data. K.A., S.D., A.C.S., and L.A.R. developed the data portal. O.P. and G.H. provided feedback on the project conception. A.O.S., N.N., M.L., and L.A.R. supervised method development. A.Di., M.L., J.B., and L.A.R. generated the 3D visualizations for imaging and scRNAseq data. M.L. drafted the first version of the manuscript with input from A.G. and S.A., and the final version was written by M.L. and L.A.R. with feedback from all co-authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by Loic A. Royer. (loic.royer@czbiohub.org)

DATA AVAILABILITY

Raw sequencing data is available at NCBI's SRA BioProject PRJNA940501. Raw imaging data are available at zebrahub.org.

CODE AVAILABILITY

All relevant code repositories are listed on zebrahub.org.

INSTRUMENT AVAILABILITY

Blueprints, parts lists, control codes, and instructions can be found at repositories listed on zebrahub.org.

METHODS

Animals. The care and experimental procedures for zebrafish were conducted in compliance with the protocols approved by the institutional animal care and use committee at the University of California San Francisco (UCSF). The fish were bred and maintained at 28.5°C and were fed three times per day using an automatic feeder[57]. Embryos were raised at 28.5°C and staged based on hours post fertilization (hpf), with the number of somites used as temporal landmarks[58] during somitogenesis. We employed wild-type EKW embryos for the scRNAseq study. For the light-sheet experiments, we used either the tg(*h2afva:h2afva*-mcherry)[33] line or the same line crossed with tg(*mezzo*:eGFP)[31] (a gift from J. Huisken, University of Gottingen). The tg(*bactin2*:memb-Cerulean-2A-*H2B*-Dendra2) line[50] was used for the photo-conversion experiments (obtained from the European Zebrafish Resource Center).

Hybridization Chain Reaction (HCR) and imaging. We investigated the spatiotemporal localization of *sox2* and *tbxta* mRNA transcripts across six developmental stages (10, 12, 14, 16, 19 hpf, and 1 dpf). A probe set designed for the target mRNAs was obtained from Molecular Instruments. To prepare the embryos for *in situ* hybridization, we dechorionated them using a 1mg/ml pronase solution for no longer than 90 seconds, then washed them thoroughly with egg water. Next, the embryos were fixed with 4% paraformaldehyde solution overnight at 4°C. We performed RNA *in situ* hybridization according to the Molecular Instruments RNA FISH protocol for zebrafish embryos (MI-Protocol-RNAFISH-Zebrafish-Rev9, adapted from Choi et al.[59]) using the probe set. Following amplification, the embryos were incubated with DAPI (1:3000 dilution) for 20 minutes during the first wash step and gradually transferred to an 80% glycerol solution for storage. Finally, the embryos were carefully mounted whole in 80% glycerol on a glass slide and imaged using a confocal microscope (Andor Dragonfly).

HCR image processing and quantification. We utilized the HCR images to quantify the proportion of presumptive NMPs based on the co-expression of *sox2* and *tbxta* mRNA transcripts. The images were first manually cropped to isolate the region of interest. Next, we denoised the images using Aydin's auto-tuned Butterworth filter. Aydin[60] is a powerful and user-friendly image-denoising tool that can significantly improve the image quality of microscopy datasets (http://aydin.app). Images were then downscaled along the z-axis by a factor of two to obtain isotropic images. The cells were then segmented using the DAPI channel in two steps. Firstly, we detected cells against the background, and secondly, we split the detected regions into individual cell instances. For the first step, we filtered the DAPI channel using a morphological opening operator[61] and then removed the gaps between nuclei using a morphological area-closing operator. We then thresholded the images using the Otsu method to identify cells against the background. In the second step, individual cells were split using the previously opened images, which were inverted, normalized, and provided as input to the watershed algorithm[62].

NMP cell counts from HCR images. Before the gene expression measurement of individual cells was extracted, we homogenized the illumination of the z-slices. For every channel, each z-slice of each image was corrected using the DAPI channel as a reference, and the background fluorescence was removed using an area-opening morphological operator per z-slice. We then extracted the total intensities of the *sox2* and *tbxta* probes for each cell and classified these as presumptive NMP by manually adjusting the threshold. The threshold was chosen empirically to include all visible mRNA transcripts while avoiding background and fluorescence artifacts.

Single-embryo cell dissociation. For scRNAseq experiments, a single embryo was used per experiment, with four replicate experiments (siblings) per developmental stage. The cell dissociation protocol is summarized in Supp. Figure 1. First, embryos were dechorionated manually with forceps. Next, a single embryo of the required stage was dissociated in either 50 µl of FACSmax solution (Gentalis - T200100) for embryos between 10 hpf to 1 dpf, or 50 µl of 0.25% Trypsin-collagenase (100 mg/ml) solution for embryos between 2 dpf to 10 dpf. The dissociation was performed by pipetting and flicking the Eppendorf gently to prevent bubble formation, and the resulting solution was centrifuged at speeds varying gradually based on the stage of interest, with lower speeds of 300 rcf for 2 mins being used for earlier stages and higher speeds up to 750 rcf for 7 mins for later stages, all while maintaining a temperature of 4°C. The dissociation media was carefully removed, and the pellet was washed with 1X PBS + 1% BSA solution. For embryos between 2 to 10 dpf, the solution was passed through a 20 µm mesh to remove undissociated debris, and then the sample was centrifuged again under the same conditions at 4°C. The supernatant was discarded, and the pellet was dissolved in 25 µl of 1X PBS+ 0.1% BSA + 18% optiprep. The cell concentration was measured using a cellometer auto T4 cell counter (Nexcelom Biosciences). The cell count was determined using a dilution (1:4) of the cell suspension and trypan blue solution.

Library preparation and sequencing. The single-cell suspensions were processed into droplet emulsions using the Chromium Single Cell Controller (10x Genomics), and scRNA-seq libraries were generated with the Chromium Next GEM Single Cell 3' Reagent Kit v3.1 (10 hpf to 1 dpf) and Single Cell 3' Reagent Kit v3.1 HT (2dpf to 10 dpf). Before loading single-cell GEMs, the single-cell suspensions were assessed using the Cellometer Auto T4 Cell Counter to determine the optimal cell count, which ranged from 700 to 1200 cells/ul for loading onto the Chromium Controller. Single cells were loaded into each channel with the appropriate amount of diluted

RT master mix based on the measured cell concentration and targeted cell recovery per sample. Fragmentation, ligation, and PCR amplification were performed using the BioRad C1000 Touch Thermal Cycler with thermal profiles specified by the 10X Genomics User Guide CG000204. Amplified cDNA and final libraries were assessed using the Agilent 4200 TapeStation/High Sensitivity D5000 Screentape and qPCR with the Kapa Library Quantification kit for Illumina. Each library was normalized to 4 nM and pooled for each NovaSeq sequencing run. The pools were sequenced with Novaseq run kits, using either single or dual index plates (Illumina 20028313), with a 1% PhiX control library spiked in. Sequencing was performed on the NovaSeq 6000 Sequencing System (Illumina).

Read alignment and single-ell filtering. After sequencing, the obtained sequences were de-multiplexed using bcl2fastq (v2.20.0.422). To align the reads to the reference genome, a reference genome was first created using CellRanger v.4.0.0. Subsequently, the CellRanger v.5.0.1 software was used for read alignment to the reference zebrafish genome.

Filtering and clustering of single-cell RNA sequencing data. To filter out low-quality cells, we performed quality control by removing cells with counts below 10,000 or with a fraction of mitochondrial genes greater than 5%. We then identified 2000 highly variable genes for each time point using Scanpy (v1.9, see scanpy.readthedocs.io/) and conducted Principal Component Analysis (PCA). We selected the first 100 principal components to cluster the single cells. To visualize the cells in a two-dimensional space, we applied the UMAP algorithm[27] (see umap-learn.readthedocs.io) to explore lower-dimensional representations of our gene expression data and facilitate the annotation of cell types. The UMAP projection showed that data from different embryos did not exhibit significant batch effects, as shown in Fig. 3.

Integration of multiple time points. To create a comprehensive UMAP embedding that included all time points, we employed the integration pipeline from Seurat[63] (v3, see satijalab.org/seurat/), a software tool used for single-cell genomics. Typically, data integration pipelines are utilized to rectify batch effects, which can lead to an over-correction of actual biological differences. However, we anticipate that some differences between time points would correspond to genuine biological changes during development. To prevent overcorrection, we utilized the integration pipeline with reciprocal PCA. This procedure generated a PCA projection of all cells from all time points. Finally, we used the first 30 principal components of the integrated PCA space to produce 2-dimensional and 3-dimensional projections using the UMAP algorithm (as depicted in Fig. 1e, Supp. Video 2, and 3).

Clusters annotation. To ensure a comprehensive representation of cellular diversity, we chose a higher clustering resolution than is typically required. We identified cell types based on the expression of cluster-specific enriched markers obtained from literature search and ZFIN database queries (see zfin.org). For each developmental stage, individual UMAPs were annotated using a primary zebrafish expert and then verified the annotations independently with a second expert. A third expert used zebrafish ontology terms from the database to homogenize cell-type naming across all time points. For the integrated UMAP centered around the NMPs population, we reannotated cell clusters following the same procedure. To aid interpretation, we also used coarser annotations of broader cell types, which are summarized in Supp. Table 1.

Gene expression variance across sibling embryos. We devised a framework to quantify gene expression differences across individual embryos. Our framework relies on estimating probability distributions of gene expression across single cells for each embryo and applying an information-theoretic approach to quantify the differences between those distributions. First, for each gene, we took the log normalized counts and grouped the counts by embryo. We filtered out genes with low counts; specifically, we discarded genes for which 80% of the cells had zero counts. We also removed genes with a mean expression less than 0.1 to prevent skewed distributions towards zero values. Additionally, we excluded genes that had less than 10 non-zero counts for each embryo. After quality control, we fitted a probability density function for the expression of each gene (n = 4 embryos per time point) using a Gaussian kernel (function scipy.stats.gaussian_kde v. 1.9.1, with the default 'scott' method to calculate the estimator bandwidth, see Fig. 2c, d, and Supp. Fig. 7). Finally, we used the Kullback-Leibler divergence to estimate, for each gene and per stage, all pairwise divergences between embryo gene count distributions (scipy.stats.entropy v. 1.9.1, and see Fig. 2c, d heatmaps, and Supp. Fig. 7). The KL divergence is a statistical distance that measures the difference between two probability distributions over the same range of values. We sampled 1000 values over the same range (min(x,y) - 1, max(x,y) + 1) for each pair of distributions P(x), P(y). Since the KL divergence is not symmetric, we computed the pairwise distances in both directions (KL(Px,Py) and KL(Py,Px), for $x \neq y$, $n = 4^2-4$ values). We then summarized all 12 pairwise KL scores with the median to get the mpKLD. Repeating this process for all genes, we obtained a distribution of mpKLDs for the whole transcriptome. We performed this analysis on different cell populations, i.e., for cell types and the entire embryo. The analysis was performed independently at each time point. To ensure our distribution estimates were robust, we only analyzed cell types with at least 80 cells per embryo.

A null model for median pairwise KL divergences (mpKLD) through embryo ID randomization. To establish a negative control and obtain a baseline for differences in gene expression between

embryos, we performed a bootstrap-like method by randomly reassigning the embryo-specific ID for each cell. We thus obtained expected mpKLD values as if the differences between embryos were due to chance alone. We repeated this process 20 times, which produced the expected mpKLD values for each gene based on the randomized embryo IDs (see Supp. Fig. 8a, red and blue histograms). To evaluate the significance of the observed differences between embryos, we calculated a z-score for each gene by comparing the observed distribution of mpKLD values to the scrambled distribution. We observed a strong linear relationship between the mpKLD scores and the z-scores (see Supp. Fig. 8a, scatterplot). This null model analysis allowed us to assess the significance of observed differences in gene expression between individual embryos.

RNA velocity. RNA velocity is a method that allows for the estimation of the rate and direction of changes in gene expression for individual cells. To perform RNA velocity analysis, we used scVelo (v0.2.5, scvelo.readthedocs.io). We started with the CellRanger alignment and identified spliced and unspliced variants for all genes using Velocyto[37] (v0.17, see velocyto.org). Next, using scVelo's dynamic modeling approach[64], we computed the RNA velocity for each gene and cell. This lets us estimate the rate and direction of changes in gene expression for each cell in our dataset. Finally, we projected the average velocity vectors as a stream plot onto the integrated UMAP to visualize the changes in gene expression over time. We repeated this process for each time point to investigate how RNA velocity vectors change over time. The resulting analysis was presented in Figure 3b.

RNA velocity graph. The RNA velocity analysis allowed us to obtain a probability transition matrix between single-cell states, which describes the likelihood of a cell transitioning from one state to another based on the dynamics of gene expression derived from RNA velocity. To construct this matrix, we used the kernel function provided by CellRank[39] (see cellrank.readthedocs.io), which takes into account both the velocity vectors and the similarity between single cells. The resulting matrix and its associated graph can be used to define a Markov process[39,64], which lets us estimate the transition rates between individual cell states and identify potential developmental trajectories. In other words, this approach models the temporal dynamics of gene expression in individual cells and predicts their transcriptional changes as they differentiate and as the tissue develops over time. The resulting graph can be visualized to provide a comprehensive overview of the transcriptional trajectories of cells during development.

Cell cluster transition graph. We assumed that developmental dynamics follow a Markov process such that the weighted graph defines the stochastic probability matrix[39]. To calculate the transitions between cell clusters, we considered independent realizations of the Markov process where each cell cluster (cell state or type) is fixed as the initial condition. Progenitor clusters, for example, will transition into their corresponding intermediate states, whereas terminal populations will not transition. Therefore, we defined the initial states of the Markov process such that all cells in each cluster start with probability 1/n, where n is the number of cells in the cluster. By applying iterative multiplication of the stochastic matrix, we obtained the distributions of cell fates at different steps. At each Markov step, we sum up the probabilities of single cells within each cluster to obtain average cell state probabilities (Fig. 3b) and repeated this process for all 39 cell clusters. To generate a coarse-grained velocity graph, we defined edges as the immediate transitions between cell clusters (Fig. 3b – velocity figure cartoon with dynamics). We defined immediate transitions between pairs of clusters according to their average transition probabilities within the first 10-time steps, when most cell states have left their initial state (except when they are terminal).

RNA velocity quantification along the mesodermal-neural axis. We used the velocity graph as previously described. For each NMP cell, we computed the graph's weighted sum of outcoming velocity magnitudes. Finally, to distinguish between the neuronal and mesodermal lineages, we defined velocity magnitudes toward neurons as “negative”, and those towards mesoderm as “positive”, and plotted the distribution of velocity magnitudes per time point for all NMPs (Fig 4i).

Pseudo-time trajectories. Based on the integrated embedding (Fig. 4a), we found potential developmental trajectories for NMPs using Slingshot[48] (see github.com/kstreet13/slingshot). Briefly, Slingshot fits a principal curve in a high-dimensional space such that single cells map to a one-dimensional projection called pseudo-time. As input for Slingshot, we used a 3-dimensional UMAP projection of the data and defined the root as the cluster with the highest number of NMPs. In agreement with our results from RNA velocity, Slingshot identified two main trajectories, one towards mesodermal cell types and another towards neuronal cell types. We verified that these two trajectories were composed of the expected cell populations with the corresponding differentiation markers (Fig. 3j, k). As a control, we checked the correspondence of pseudo-time to real-time and verified that the most differentiated cells present at the end of the trajectories came from the latest time-point (24 hpf, see Fig. 3l). Finally, to find differential gene expression across pseudo time, we used tradeSeq[65] (see github.com/statOmics/tradeSeq) and tested for differences between the neuronal and mesodermal lineages. We also found genes specifically associated with each lineage (Fig. 3j).

Aligned UMAPs of individual time points. The stochastic nature of the UMAP algorithm leads to variations in the orientation, scale,

and layout of UMAP projections. This impedes the comparison of datasets, such as developmental stages (e.g., Fig. 3h), as it is challenging to distinguish between (i) genuine differences in the data, and (ii) differences caused by algorithmic stochasticity. To address this issue, we employed the aligned UMAP algorithm, which aims to align projections by utilizing pairwise anchors between different datasets. For each point (cell), the nearest neighbors in the previous and next developmental stages were used as anchors. To prevent over-alignment artifacts, we relaxed the strength of the alignment regularization parameter to 0.001. The outcome is a series of UMAP projections, each corresponding to a time point, that exhibit conserved relative orientation and proportions with one another (Fig. 3h). We confirmed that the aligned UMAPs maintained the biological relationships of known lineages as indicated by the expression of known markers and the distribution of cell type annotations (Fig 3h).

Adaptive multi-view multi-color light-sheet imaging and processing. The open-source repository at github.com/royerlab/opensimview provides all information for building a 2 detection and 2 illumination simultaneous adaptive multi-view multi-color microscope[29,30]. We provide all blueprints, a parts list, the control software, and image processing tools. Once acquired the dataset consists of multiple views which were fused, deconvolved, and stabilized using DEXP[66] (see github.com/royerlab/dexp). DEXP is an image processing and visualization library based on napari, CuPy, Zarr, and DASK designed for managing and processing light-sheet microscopy datasets. It consists of light-sheet imaging specialized image processing functions (equalization, denoising, dehazing, registration, fusion, stabilization, deskewing, deconvolution), visualization functions (napari-based viewing, 2D/3D rendering, video compositing and resizing, mp4 generation), as well as dataset management functions (copy, crop, concatenation, tiff conversion). Almost all functions are GPU accelerated via CuPy but also have a numpy-based fallback option for testing on small datasets. In addition to a functional API, DEXP offers a command line interface that makes it very easy for non-coders to pipeline large processing jobs from a large multi-terabyte raw dataset to fully processed and rendered video in MP4 format.

Sample preparation for multi-view light-sheet Imaging. Embryos were first gently dechorionated with a pair of sharp forceps under a binocular dissecting microscope. Embryos were gently transferred in FEP tubes (Zeus Inc. custom order) embedded in 0.1% low-gelling-temperature agarose (Sigma, A0701). The FEP tubes were placed in a custom-made sample capillary holder previously filled with 0.7% low-gelling temperature agarose to support the embryo and then placed in the chamber using a holder insertion tool and section pipette tip (sealed with low toxicity silicone – Kwik-Sil^{tm} from World Precision Instrument). A detailed procedure for the mounting is shown in Supp. Fig. 6a. CAD drawings of the custom-made FEP holder and the mounting chamber are available in Supp. Fig. 5. For recording after 15 hpf, fresh and filtered embryo water was circulating in the chamber with 0.016% tricaine.

Photo-conversion experiments. For the photo-conversion, we used a python-based custom-made software (see photo-conversion arm in github.com/royerlab/opensimview), that allows the users to select a region of interest and photoconvert this area by controlling a two-axis galvo mirror actuator. We selected the embryo area to photoconvert based on our HCR mapping of the presumptive NMP territory (Supp. Fig. 12). Photoconversion was achieved using a 405 nm laser. Time and laser power were manually accessed based on the sample fluorophore state transition evaluation.

Photo-conversion image analysis and quantification. The estimation of the proportion of neural and mesodermal nuclei on the photo-manipulated light-sheet data requires three steps: segmentation of the nuclei; selection of which nuclei were photo-manipulated; nuclei label assignment to neural or mesodermal tissue. Before segmenting the nuclei, for each time-lapse, an individual frame (i.e., a 3D image) was selected at the time point of interest. The photo-manipulated channel fluorescence can be weak in some nuclei; thus, we used both channels (red and green) for segmentation. We removed the background using morphological filtering and thresholding. The images were normalized, so their intensities were between 0 and 1. These normalized images were forwarded into a custom-trained 3D U-Net[67] where their nuclei and boundaries were detected. The nuclei detections were binarized with a threshold of 0.5 and the green, and red channels detections were combined using a logical OR operator, the boundaries detection were combined using addition, resulting in a single binary nuclei detection and boundary per image. Finally, the nuclei were segmented using the hierarchical watershed[68] and selecting the segments by picking a single level of the hierarchy (i.e., hierarchy horizontal cut). We extracted the original images' total intensity of red and green fluorescence for each segment. The photo-manipulated nuclei were selected by manual thresholding of their total red fluorescence intensity – a single threshold per image. The regions of interest of each image were manually annotated into mesodermal or neural regions using napari[52], these annotations were propagated into overlapping nuclei, assigning their respective labels to the nuclei's segments.

High-resolution & large volume single-objective light-sheet imaging. We used our DaXi imaging platform[32] to perform high-resolution imaging (1.0 NA) of tail development. For each time point, we use two orthogonal oblique light sheets to improve coverage. Images were processed using DEXP[66] to (i) equalize brightness between views, (ii) remove background, (iii) correct for distortions between views, and (iv) temporally stabilize the

timelapse to remove artifacts caused by sample movement or vibrations. Instructions to build and operate this instrument can be found are: github.com/royerlab/daxi

Sample preparation for single objective light-sheet microscopy (DaXi). Embryos are first dechorionated in a glass-bottom cell culture dish (35mm Cell Culture Dish with Glass Bottom 20mm – Stellar Scientific) and then gently transferred into a 0.7% solution of low-gelling-temperature agarose (Sigma, A0701). Embryos are then positioned at the correct position and angle for imaging using a custom-made capillary. When solid, the agarose surrounding the tail was cut off and removed using a dissection knife and forceps to permit full development and tail elongation (Supp. Fig. 6b). The time lapses were done with a gentle flow of fresh and filtered embryo-medium solution with 0.016% tricaine to prevent embryo movement.

Whole embryo cell segmentation and tracking. As with the photo-converted images from the multi-view microscope, the images from the single-objective microscope were fused and stabilized using DEXP[66]. Once processed, the time-lapse was cropped to select the region of interest encompassing the embryo's tail. Once again, the cropped video was stabilized to further reduce the local movement between frames of the extremely dynamic process of tail development. Subsequent analyses were conducted on this cropped and stabilized time-lapse. To obtain the tracks, we used a hybrid approach combining multiple segmentation hypothesis tracking and the latest fluorescence imaging nuclei segmentation development using deep learning. We detect the nuclei and their boundaries using a 3D U-Net[67] and the watershed algorithm[68] to obtain the candidate segments. Instead of performing a horizontal cut to select segments, we use segment hierarchies to generate multiple segmentation hypotheses and let the tracking algorithm pick the most likely segments.

***In silico* fate mapping.** Despite recent advancements[69,70] cell tracking is far from being a solved problem and still requires extensive manual curation. Unfortunately, manual curation of tens or even hundreds of thousands of cell trajectories is simply infeasible. We estimate the local cell flow between frames to circumvent this issue and use this as a robust alternative for analyzing the cells' movements. For the local movement estimation, we use radial neighborhood regression[71]. For each time point, we fit a regression model that predicts the positions of cells in subsequent frames. The training data for these models are pairs of adjacent same-track coordinates at the time point of interest. Track endpoints (splits) are connected to the nearest starting point on the next frame. With these regression models, one for each time point, we can evaluate the movement of any cell at any location if they fall within the radius of an existing track. The tool is available at github.com/royerlab/in-silico-fate-mapping, and lets users predict the cells' fate by estimating their movement over time given a user-selected starting point. This is analogous to a photo-manipulation experiment but *in silico*.

Dynamic Morphoskeleton. To investigate cellular flows during morphogenesis in the zebrafish tailbud, we used the Dynamic Morphoskeletons (DM) method[51], which is based on the concept of Lagrangian Coherent Structures (LCS). The DM is a self-consistent and objective method that identifies repellers, attractors, and their domains along cell trajectories by computing the Lagrangian stretching. The Lagrangian stretching is computed from cell trajectories using a recently developed technique that identifies LCSs from sparse and noisy trajectories[72]. For all results presented in this study, we set the radius of neighboring trajectories to a fixed value and used a time interval of interest. The Lagrangian stretching facilitates the identification of repelling spatiotemporal regions, which are marked by high values of Lagrangian stretching, and enables the characterization of differential mechanical environments for cell fate specification. The DM method is ideal for quantifying morphogenesis as it reveals the intrinsic geometric feature of spatiotemporal trajectories and is robust to noise. For more details explanation see the Supp. Theory Note.

3D visualization & rendering. Visualization, interaction, and figure-making of the available data were done using napari[52] (see napari.org), particularly the napari-animation plugin (github.com/napari/napari-animation). We also used rendering routines in DEXP, some of which use the Spimagine (see github.com/maweigert/spimagine) volume rendering code. Additionally, the light-sheet datasets were stored using zarr[73] to enable out-of-memory loading. 3D RNA velocity visualizations were rendered using the Houdini software.

SUPPLEMENTARY VIDEOS CAPTIONS

Supplementary Video 1. Zebrahub's full dataset scRNAseq UMAP color-coded by annotations.
Supplementary Video 2. Zebrahub's full dataset scRNAseq UMAP color-coded by time points.
Supplementary Video 3. Light-sheet time-lapse of Tg(*h2afva:h2afva*-mcherry; *mezzo*:eGFP) embryo starting at 50% epiboly.
Supplementary Video 4. Image processing and analysis steps for nuclei segmentation and tracking.
Supplementary Video 5. Dynamic visualization of the RNA velocity flows overlayed on full UMAP color-coded by annotations.
Supplementary Video 6. RNA velocity-derived flows around the NMP cluster
Supplementary Video 7. *In silico* fate mapping of the early NMP presumptive territory.
Supplementary Video 8. *In silico* fate mapping of the mesodermal axial progenitors.
Supplementary Video 9. *In silico* fate mapping of the neural axial progenitors.
Supplementary Video 10. *In silico* fate mapping of the late NMP presumptive territory.
Supplementary Video 11. *In vivo* Photoconversion of the early NMP presumptive territory using the tg(*bactin2*:memb-Cerulean-2A-*H2B*-Dendra2) transgenic line.
Supplementary Video 12. *In vivo* Photoconversion of the late NMP presumptive territory using the tg(*bactin2*:memb-Cerulean-2A-*H2B*-Dendra2) transgenic line.
Supplementary Video 13. Tracking of an NMP with daughter cells integrating both mesodermal and neural tissues.

SUPPLEMENTARY TABLE CAPTIONS

Supplementary Table 1. Annotation ontology tracker.
Supplementary Table 2. Differential gene expression for annotated cells at 10, 12, 14, 16, 19, and 24 hpf.

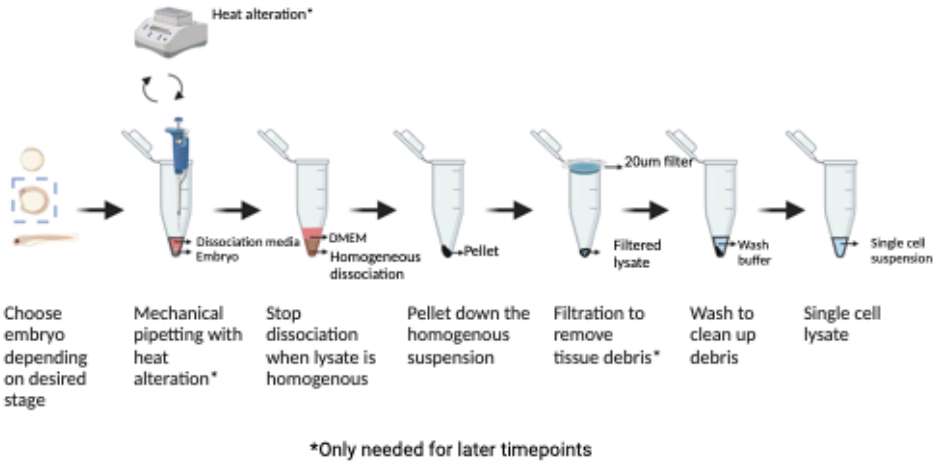
REFERENCES

1. Conklin, E. G. *The Organization and Cell-lineage of the Ascidian Egg*. (Academy of Natural Sciences, 1905).
2. Vogt, W. [Not Available]. *Wilhelm Roux Arch. Entwicklungsmechanik Org.* **120**, 384–706 (1929).
3. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64–119 (1983).
4. Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev. Biol.* **56**, 110–156 (1977).
5. Guignard, L. *et al.* Contact area–dependent cell communication and the morphological invariance of ascidian embryogenesis. *Science* **369**, eaar5663 (2020).
6. Huisken, J., Swoger, J., Del Bene, F., Wittbrodt, J. & Stelzer, E. H. K. Optical Sectioning Deep Inside Live Embryos by Selective Plane Illumination Microscopy. *Science* **305**, 1007–1009 (2004).
7. Keller, P. J., Schmidt, A. D., Wittbrodt, J. & Stelzer, E. H. K. Reconstruction of Zebrafish Early Embryonic Development by Scanned Light Sheet Microscopy. *Science* **322**, 1065–1069 (2008).
8. McDole, K. *et al.* In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level. *Cell* **175**, 859-876.e33 (2018).
9. Wolff, C. *et al.* Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb. *eLife* **7**, e34410 (2018).
10. Wan, Y. *et al.* Single-Cell Reconstruction of Emerging Population Activity in an Entire Developing Circuit. *Cell* **179**, 355-372.e23 (2019).
11. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
12. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
13. Farrell, J. A. *et al.* Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).
14. Farnsworth, D. R., Saunders, L. M. & Miller, A. C. A single-cell transcriptome atlas for zebrafish development. *Dev. Biol.* **459**, 100–108 (2020).
15. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
16. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
17. Packer, J. S. *et al.* A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science* **365**, eaax1971 (2019).
18. Saunders, L. M. *et al. Deep molecular, cellular and temporal phenotyping of developmental perturbations at whole organism scale*. http://biorxiv.org/lookup/doi/10.1101/2022.08.04.502764 (2022) doi:10.1101/2022.08.04.502764.
19. Zhai, J. *et al.* Primate gastrulation and early organogenesis at single-cell resolution. *Nature* **612**, 732–738 (2022).
20. Lohoff, T. *et al.* Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. *Nat. Biotechnol.* **40**, 74–85 (2022).
21. Martin, B. L. & Kimelman, D. Canonical Wnt Signaling Dynamically Controls Multiple Stem Cell Fate Decisions during Vertebrate Body Formation. *Dev. Cell* **22**, 223–232 (2012).
22. Guillot, C., Djeffal, Y., Michaut, A., Rabe, B. & Pourquié, O. Dynamics of primitive streak regression controls the fate of neuromesodermal progenitors in the chicken embryo. *eLife* **10**, e64819 (2021).
23. Tzouanacou, E., Wegener, A., Wymeersch, F. J., Wilson, V. & Nicolas, J.-F. Redefining the Progression of Lineage Segregations during Mammalian Embryogenesis by Clonal Analysis. *Dev. Cell* **17**, 365–376 (2009).
24. Attardi, A. *et al.* Neuromesodermal progenitors are a conserved source of spinal cord with divergent growth dynamics (doi: 10.1242/dev.166728). *Development* **146**, dev175620 (2019).
25. Koch, F. *et al.* Antagonistic Activities of Sox2 and Brachyury Control the Fate Choice of Neuro-Mesodermal Progenitors. *Dev. Cell* **42**, 514-526.e7 (2017).
26. Gouti, M. *et al.* A Gene Regulatory Network Balances Neural and Mesoderm Specification during Vertebrate Trunk Development. *Dev. Cell* **41**, 243-261.e7 (2017).
27. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018) doi:10.48550/ARXIV.1802.03426.
28. Bradford, Y. M. *et al.* Zebrafish information network, the knowledgebase for Danio rerio research. *Genetics* **220**, iyac016 (2022).
29. Tomer, R., Khairy, K., Amat, F. & Keller, P. J. Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy. *Nat. Methods* **9**, 755–763 (2012).
30. Royer, L. A. *et al.* Adaptive light-sheet microscopy for long-term, high-resolution imaging in living organisms. *Nat. Biotechnol.* **34**, 1267–1278 (2016).
31. Ruprecht, V. *et al.* Cortical Contractility Triggers a Stochastic Switch to Fast Amoeboid Cell Motility. *Cell* **160**, 673–685 (2015).
32. Yang, B. *et al.* DaXi—high-resolution, large imaging volume and multi-view single-objective light-sheet microscopy. *Nat. Methods* **19**, 461–469 (2022).

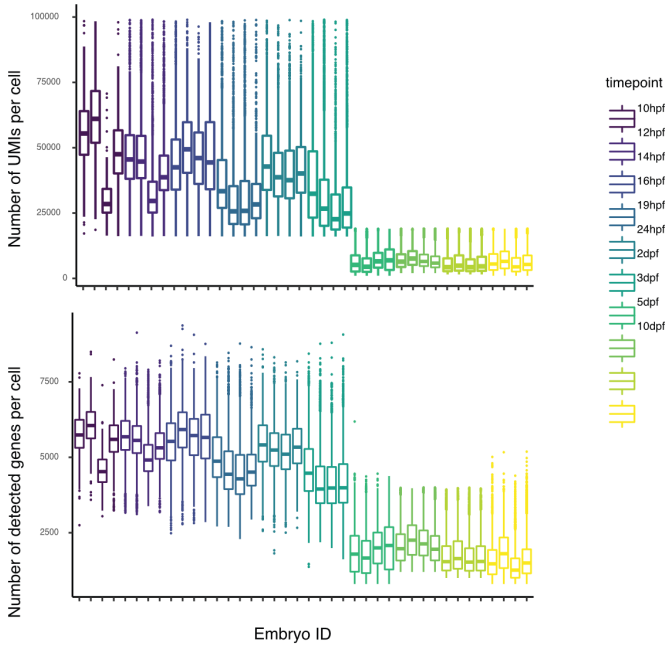
33. Shah, G. *et al.* Multi-scale imaging and analysis identify pan-embryo cell dynamics of germlayer formation in zebrafish. *Nat. Commun.* **10**, 5753 (2019).
34. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
35. Chari, T., Banerjee, J. & Pachter, L. *The Specious Art of Single-Cell Genomics*. http://biorxiv.org/lookup/doi/10.1101/2021.08.25.457696 (2021) doi:10.1101/2021.08.25.457696.
36. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203**, 253–310 (1995).
37. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
38. Wymeersch, F. J., Wilson, V. & Tsakiridis, A. Understanding axial progenitor biology *in vivo* and *in vitro*. *Development* **148**, dev180612 (2021).
39. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
40. Nowak, J. S., Choudhury, N. R., de Lima Alves, F., Rappsilber, J. & Michlewski, G. Lin28a regulates neuronal differentiation and controls miR-9 production. *Nat. Commun.* **5**, 3687 (2014).
41. Aires, R. *et al.* Tail Bud Progenitor Activity Relies on a Network Comprising Gdf11, Lin28, and Hox13 Genes. *Dev. Cell* **48**, 383-395.e8 (2019).
42. Robinton, D. A. *et al.* The Lin28/let-7 Pathway Regulates the Mammalian Caudal Body Axis Elongation Program. *Dev. Cell* **48**, 396-405.e3 (2019).
43. Aires, R. *et al.* Oct4 Is a Key Regulator of Vertebrate Trunk Length Diversity. *Dev. Cell* **38**, 262–274 (2016).
44. Riba, A. *et al.* Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nat. Commun.* **13**, 2865 (2022).
45. Yamanaka, Y., Tamplin, O. J., Beckers, A., Gossler, A. & Rossant, J. Live imaging and genetic analysis of mouse notochord formation reveals regional morphogenetic mechanisms. *Dev. Cell* **13**, 884–896 (2007).
46. Davis, R. L. & Kirschner, M. W. The fate of cells in the tailbud of Xenopus laevis. *Development* **127**, 255–267 (2000).
47. Row, R. H., Tsotras, S. R., Goto, H. & Martin, B. L. The zebrafish tailbud contains two independent populations of midline progenitor cells that maintain long-term germ layer plasticity and differentiate based on local signaling cues. *Development* dev.129015 (2015) doi:10.1242/dev.129015.
48. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).
49. Van den Berge, K. *et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).
50. Dempsey, W. P., Fraser, S. E. & Pantazis, P. PhOTO Zebrafish: A Transgenic Resource for In Vivo Lineage Tracing during Development and Regeneration. *PLoS ONE* **7**, e32888 (2012).
51. Serra, M., Streichan, S., Chuai, M., Weijer, C. J. & Mahadevan, L. Dynamic morphoskeletons in development. *Proc. Natl. Acad. Sci.* **117**, 11444–11449 (2020).
52. Sofroniew, N. *et al.* napari: a multi-dimensional image viewer for Python. (2022) doi:10.5281/zenodo.7276432.
53. DeVeale, B. *et al.* Oct4 is required ~E7.5 for proliferation in the primitive streak. *PLoS Genet.* **9**, e1003957 (2013).
54. Economou, C. *et al.* Intrinsic factors and the embryonic environment influence the formation of extragonadal teratomas during gestation. *BMC Dev. Biol.* **15**, 35 (2015).
55. Livigni, A. *et al.* A Conserved Oct4/POUV-Dependent Network Links Adhesion and Migration to Progenitor Maintenance. *Curr. Biol.* **23**, 2233–2244 (2013).
56. Zalc, A. *et al.* Reactivation of the pluripotency program precedes formation of the cranial neural crest. *Science* **371**, eabb4776 (2021).
57. Lange, M. *et al.* ZAF, the first open source fully automated feeder for aquatic facilities. *eLife* **10**, e74234 (2021).
58. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203**, 253–310 (1995).
59. Choi, H. M. T. *et al.* Third-generation *in situ* hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* **145**, dev165753 (2018).
60. Solak, A. C., Royer, L. A., Janhangeer, A.-R. & Kobayashi, H. royerlab/aydin: v0.1.15. (2022) doi:10.5281/zenodo.7222198.
61. Vincent, L. Morphological Area Openings and Closings for Grey-scale Images. in *Shape in Picture* (eds. O, Y.-L., Toet, A., Foster, D., Heijmans, H. J. A. M. & Meer, P.) 197–208 (Springer Berlin Heidelberg, 1994). doi:10.1007/978-3-662-03039-4_13.
62. Lotufo, R. A., Falcao, A. X. & Zampirolli, F. A. IFT-Watershed from gray-scale marker. in *Proceedings. XV Brazilian Symposium on Computer Graphics and Image Processing* 146–152 (2002). doi:10.1109/SIBGRA.2002.1167137.
63. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
64. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
65. Van den Berge, K. *et al.* Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.* **11**, 1201 (2020).
66. Royer, L. A., Bragantini, J., Solak, A. C., Haase, R. & Yang, B. royerlab/dexp: First zenodo release. (2022) doi:10.5281/zenodo.7039059.
67. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention –*

MICCAI 2015 (eds. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) vol. 9351 234–241 (Springer International Publishing, 2015).
68. Cousty, J. & Najman, L. Incremental algorithm for hierarchical minimum spanning forests and saliency of watershed cuts. in *10th International Symposium on Mathematical Morphology (ISMM'11)* vol. 6671 272–283 (2011).
69. Sugawara, K., Çevrim, Ç. & Averof, M. Tracking cell lineages in 3D by incremental deep learning. *eLife* **11**, e69380 (2022).
70. Malin-Mayor, C. *et al.* Automated reconstruction of whole-embryo cell lineages by learning from sparse annotations. *Nat. Biotechnol.* 1–6 (2022) doi:10.1038/s41587-022-01427-7.
71. Grisel, O. *et al.* scikit-learn/scikit-learn: scikit-learn 0.24.1. (2021) doi:10.5281/zenodo.4450597.
72. Mowlavi, S., Serra, M., Maiorino, E. & Mahadevan, L. Detecting Lagrangian coherent structures from sparse and noisy trajectory data. *J. Fluid Mech.* **948**, A4 (2022).
73. Miles, A. *et al.* zarr-developers/zarr-python: v2.4.0. (2020) doi:10.5281/zenodo.3773450.

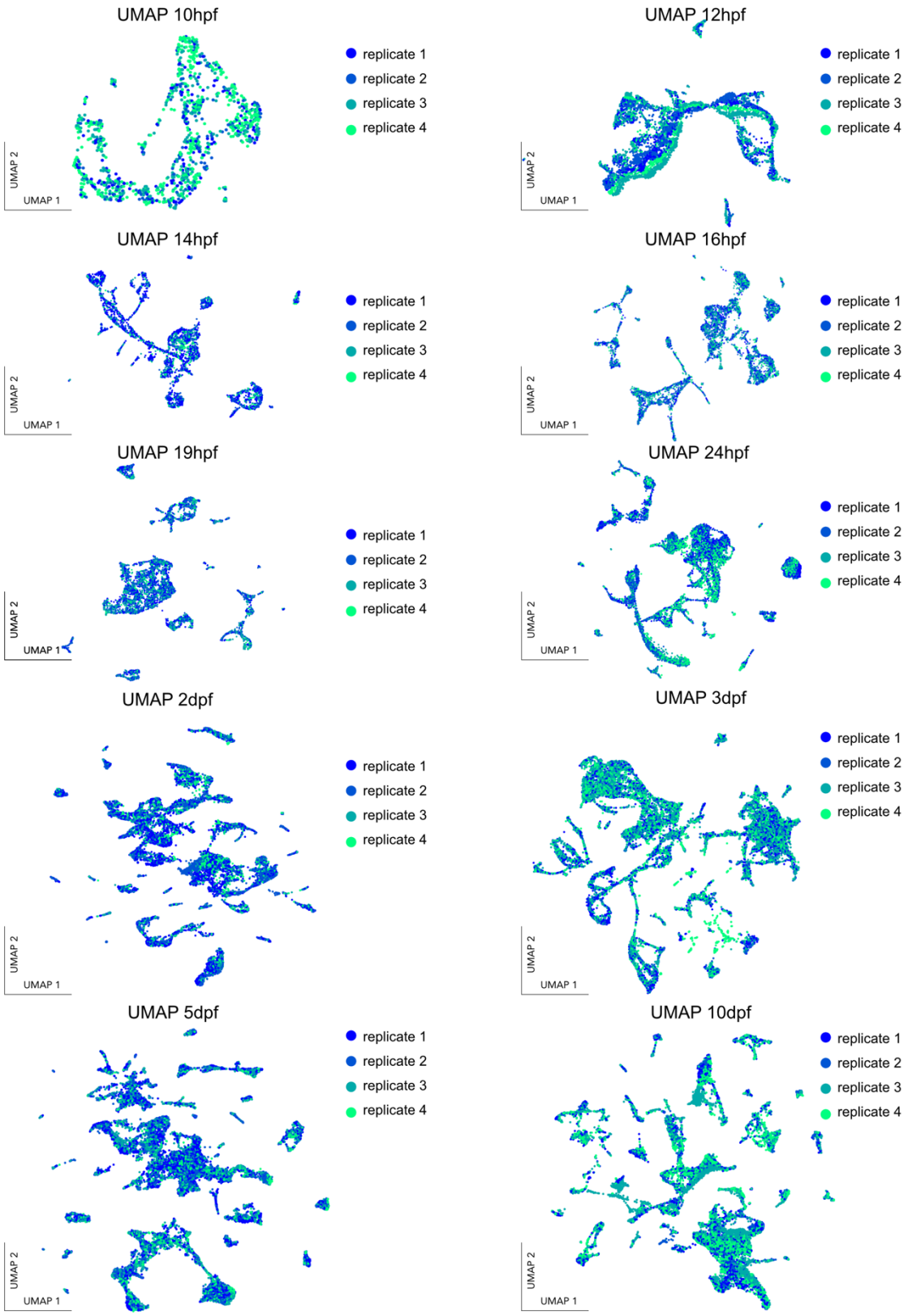
SUPPLEMENTARY FIGURES



Supplementary Figure 1. Schematic representation of the single embryo dissociation protocol for zebrafish embryos.

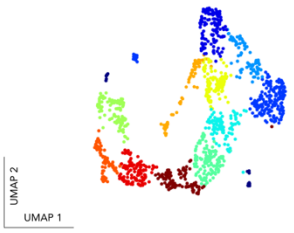


Supplementary Figure 2. Distribution of UMIs and detected genes per embryo. The number of detected UMIs per cell for each embryo. Individual embryos are shown in the x-axis sorted by developmental time point (n=40 embryos, 10 developmental time points). The distribution of UMIs per cell is shown as a boxplot. An equivalent plot for the number of detected genes per cell.



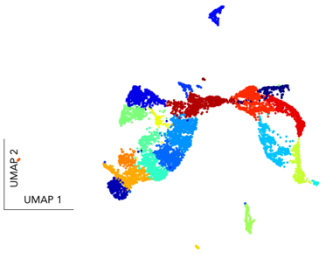
Supplementary Figure 3. UMAPs for each timepoint colored coded by the individual replicate sequenced embryo, showing no batch effect across experiment.

UMAP 10hpf - bud stage



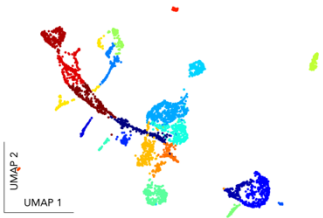
- anterior axial hypoblast
- anterior neural plate
- ectodermal cell
- forebrain neural keel
- hindbrain neural keel
- hindbrain neural plate
- mesoderm
- midbrain neural keel
- notochord
- paraxial mesoderm
- segmental plate
- tail bud

UMAP 12hpf - 5 somites



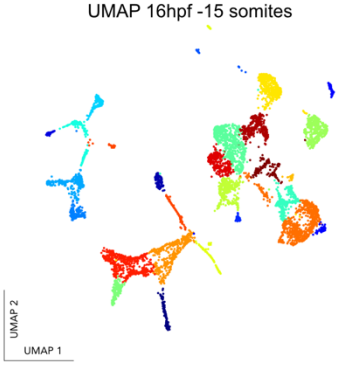
- adaxial cell
- anterior neural keel
- ectodermal cell
- endoderm
- floor plate
- hindbrain neural keel
- hindbrain neural plate
- lateral mesoderm
- lateral plate mesoderm
- midbrain hindbrain boundary neural keel
- neural crest
- neurogenic placode
- notochord
- paraxial mesoderm
- periderm
- polster
- presumptive diencephalon
- presumptive telencephalon
- primordial germ cell
- segmental plate
- somite
- tail bud
- yolk syncytial layer

UMAP 14hpf -10 somites



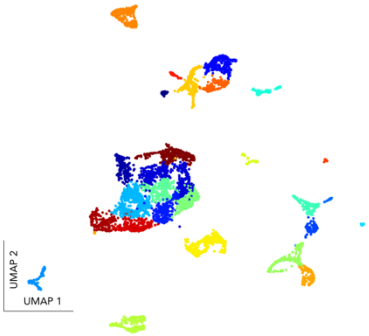
- central nervous system
- ectodermal cell
- endoderm
- epidermal cell
- floor plate
- ionocyte progenitor cell
- lateral plate mesoderm
- midbrain hindbrain boundary neural keel
- neural crest
- neural plate
- notochord
- optic vesicle
- otic placode
- paraxial mesoderm
- periderm
- polster
- posterior lateral mesoderm
- presumptive diencephalon
- presumptive telencephalon
- primary neuron
- primordial germ cell
- pronephric mesoderm
- segmental plate
- somite
- tail bud

UMAP 16hpf -15 somites

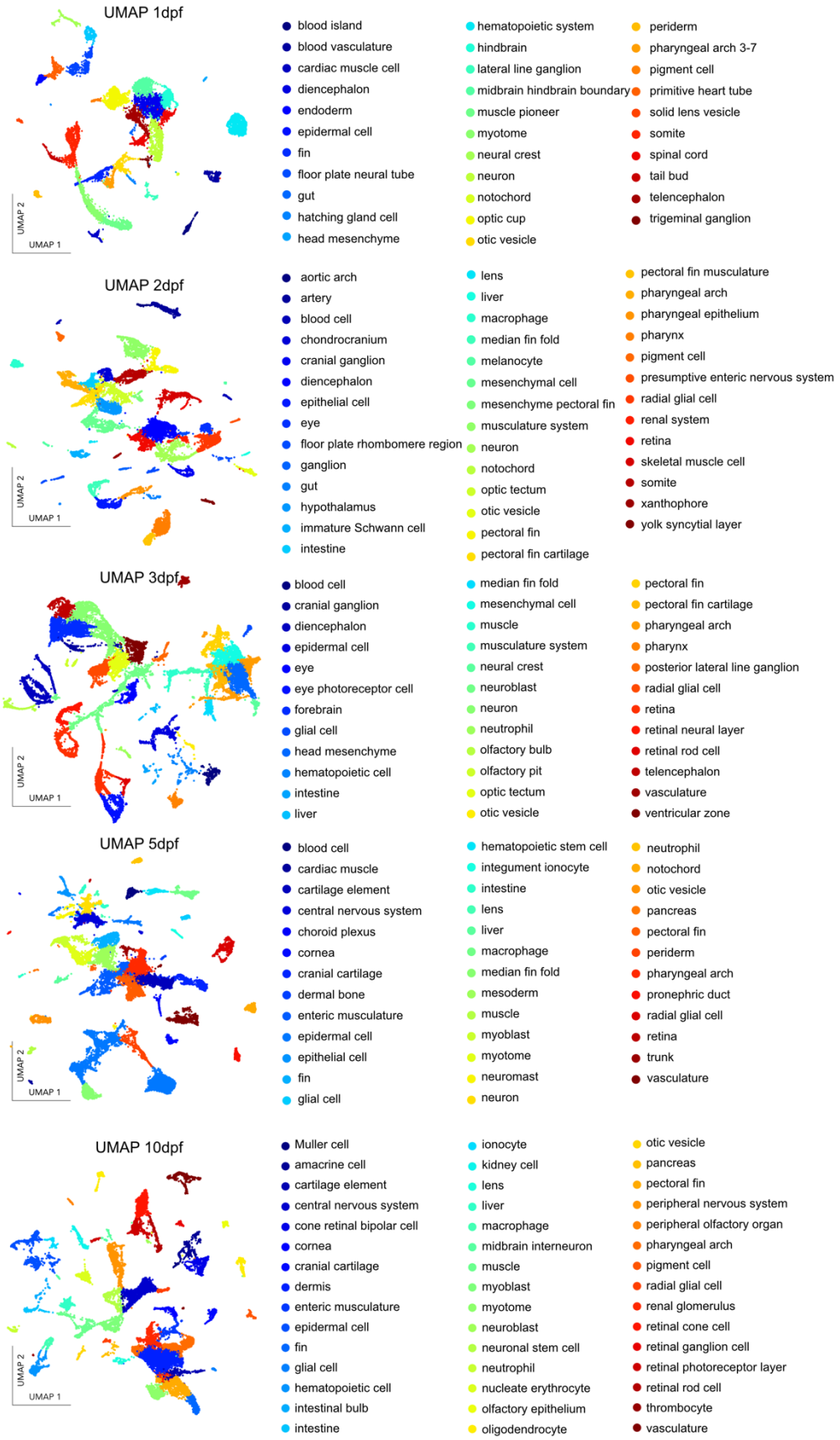


- adaxial cell
- brain
- common myeloid progenitor
- ectodermal cell
- floor plate
- hatching gland cell
- head mesenchyme
- heart
- hematopoietic system
- lateral mesoderm
- lens placode
- midbrain hindbrain boundary
- myotome
- neural crest
- neural tube
- notochord
- optic vesicle
- otic placode
- paraxial mesoderm
- periderm
- pronephros
- somite
- spinal cord neural tube
- telencephalon
- trigeminal placode

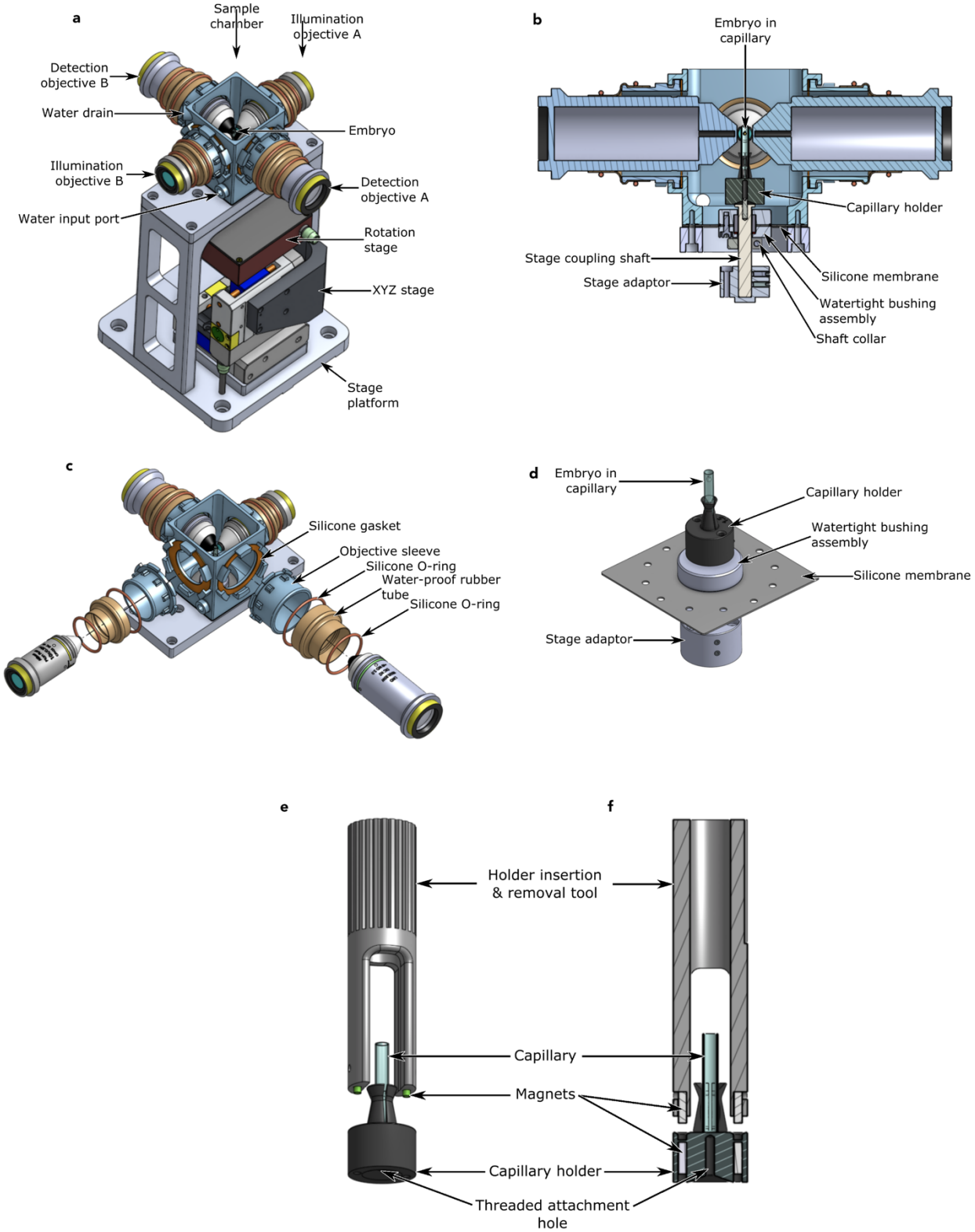
UMAP 19hpf - 20 somites



- Kupffer's vesicle
- cranial ganglion
- diencephalon
- epidermal cell
- floor plate rhombomere region
- head mesenchyme
- heart
- hematopoietic system
- hindbrain neural tube
- immune system
- lateral mesoderm
- lateral plate mesoderm
- midbrain
- midbrain hindbrain boundary
- myotome
- neural crest
- notochord
- optic vesicle
- otic placode
- paraxial mesoderm
- periderm
- pharyngeal arch
- pronephros
- solid lens vesicle
- spinal cord
- tail bud
- telencephalon



Supplementary Figure 4. Summary for the cell type annotations represented on each timepoint UMAPs. We annotated 529 clusters based on gene enrichment cross-checked with published articles (see Methods).



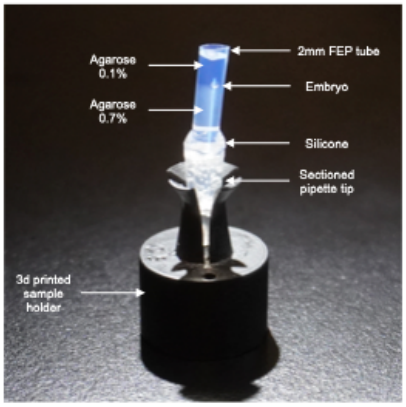
Supplementary Figure 5. Structure of the sample chamber (continued next page).

The sample chamber is designed to hold the sample (fish embryo) immersed in water and provide water-tight ports for two illumination and two detection objectives (allowing the objectives to move in 3 dimensions) aimed at the sample, and a water-tight connection to the combined XYZ and rotation (XYZR) stage that moves the sample with those four degrees of freedom.

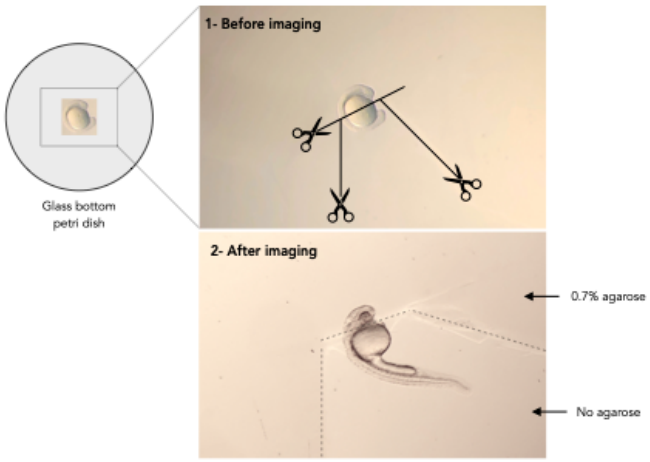
- **a.** The sample chamber is mounted on a custom-machined aluminum holder that places it above the XYZR stage, which is in turn mounted onto a custom-machined aluminum platform. The sample (fish embryo) sits in the middle of the chamber, surrounded by the four objectives.
- **b.** The sample chamber itself is composed of the following parts: (1) A main body (3D printed) with openings for the four objectives and input and output ports for water. (2) Four objective sleeves (3D printed) that connect to the chamber body using a twist-lock bayonet mechanism. (3) Laser-cut silicone gaskets that provide a water-tight seal between the objective sleeves and the chamber body. (4) Four rubber tubes that create a flexible and water-tight coupling between the objective sleeves and the objectives, while allowing movement of the objectives in all 3 dimensions relative to the sample. These tubes are made by cutting sections of the fingers from nitrile laboratory gloves. (5) Silicone O-rings that secure the rubber tubes to the objective sleeves and objectives, to reduce the chance of water leakages. (6) A custom-machined aluminum base to which the chamber body is bolted
- **c.** Cross section of the sample chamber, along the main axis of the detection objectives, showing the fish embryo in the capillary. The capillary is mounted onto a 3D printed capillary holder that thread onto a custom-machined stainless-steel shaft that couples to the XYZR stage. This shaft passes through a laser-cut silicone membrane clamped between the chamber body and the chamber base, via a water-tight bushing composed of two 3D printed halves that clamp onto the silicone membrane and hold in place a tubular silicone gasket (section of silicone tubing) that seals around the shaft, with silicone grease applied as lubricant between the tube and the shaft. The bushing assembly is designed such that the tubular silicone gasket is slightly compressed axially, to increase the sealing pressure against the shaft to avoid water leaks. The shaft can freely rotate inside the bushing (driven by the rotation stage), while the silicone membrane allows the shaft to move in 3 dimensions (driven by the XYZ stage). A 3D printed collar around the shaft limits the vertical sliding of the shaft in the bushing to about 2mm. The shaft is connected to the rotation stage via a custom-machined aluminum adaptor.
- **d.** Detail of the capillary holder showing the top half of the bushing assembly and the silicone membrane. The circular pattern of holes on the membrane are for the screws that attach the chamber body to the chamber base.
- **e.** Angle view. **F.** Cross section view of a 3D printed tool is used to easily insert the capillary holder into, and remove it from, the sample chamber. The tool has two 1/16" (1.587mm) diameter neodymium magnets (axially magnetized) that mate into corresponding cavities in the capillary holder. Identical magnets are embedded into the capillary holder (right underneath the tool-mating cavities) to provide a mating force between the tool and the holder. All magnets are glued in place using epoxy adhesive. With this design the tool can transfer the capillary holder to the chamber and screw it onto the stage-coupling shaft (or remove it from the shaft). The magnetic force produced by the magnets is strong enough to hold the weight of the capillary holder, and the magnet-to-cavity coupling allows the transfer of torque for screwing the holder onto the shaft.

All 3D printed chamber parts and holder were printed on Formlabs Form 2 and Form 3 3D printers using Formlabs Tough v5 resin.

a - MultiView light-sheet microscope zebrafish embryo mounting

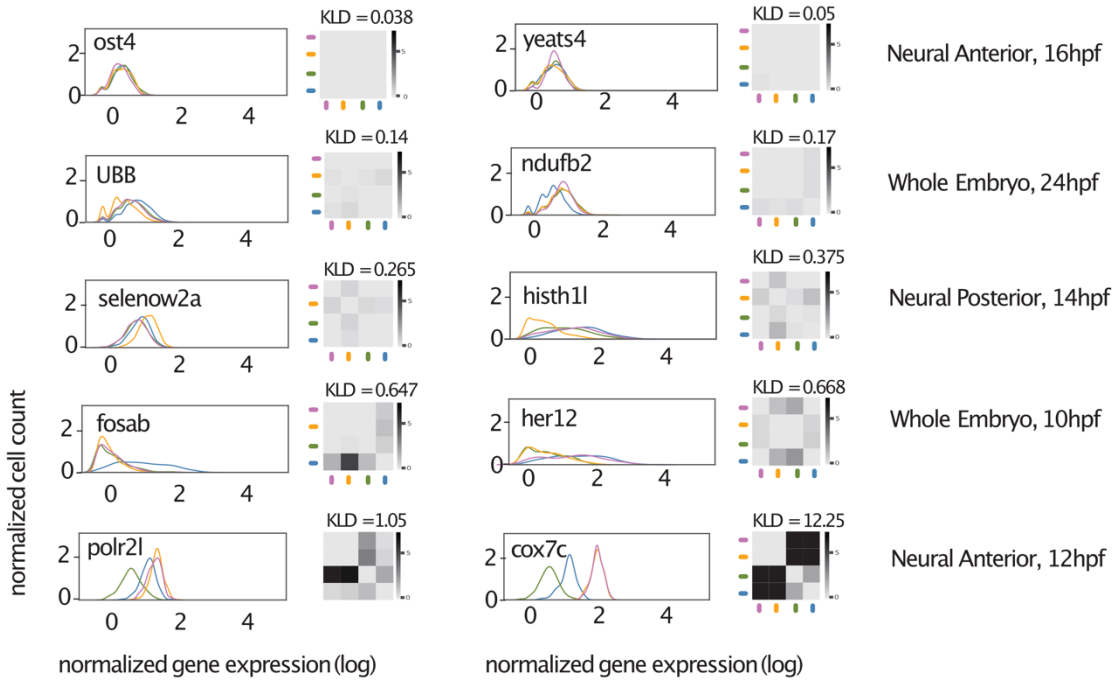


b - Single objective light-sheet microscope zebrafish embryo mounting

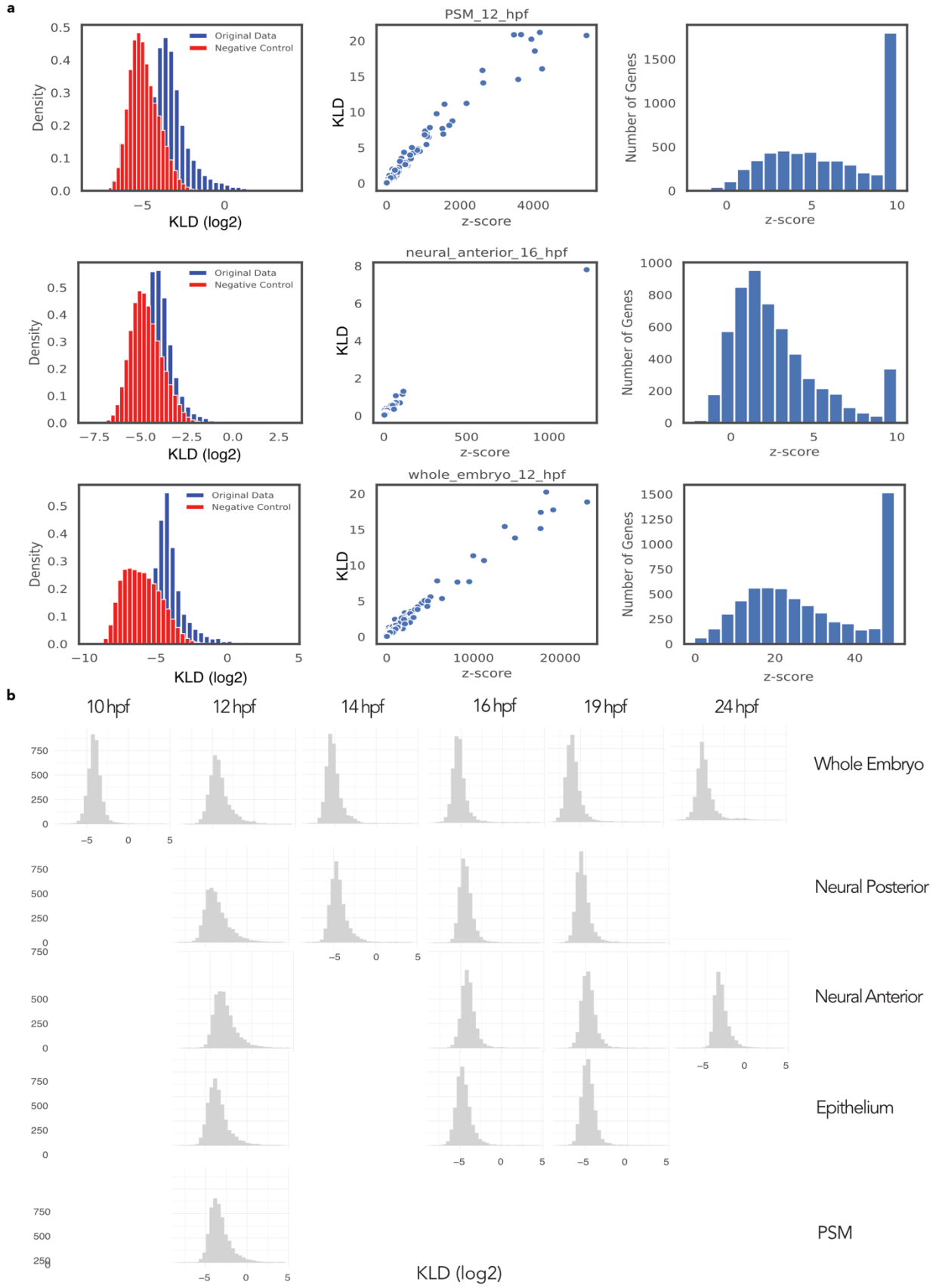


Supplementary Figure 6. Mounting strategies for long-term zebrafish imaging in the light-sheet microscope.

a. For the multiview light-sheet microscope the zebrafish are mounted in FEP tubes in 0.1% low melting point agarose in embryo medium supported by a 0.7% % low melting point agarose (Sigma, A0701). The low concentration agarose allows the full embryo development with tail elongation. The FEP is sealed in a pipette tip sealed with low-toxicity silicone (Kwik-Sil^{tm} from World Precision Instrument) and then placed in the capillary holder described Fig. S5 and placed in the chamber using the capillary holder insertion & removal tool.
b. Embryo is placed in a 35mm cell culture dish with a 20mm glass bottom (Stellar Scientific) in 0.7% low melting point agarose (Sigma, A0701). After correct positioning the agarose surrounding the posterior part is removed, to not prevent tail extension.

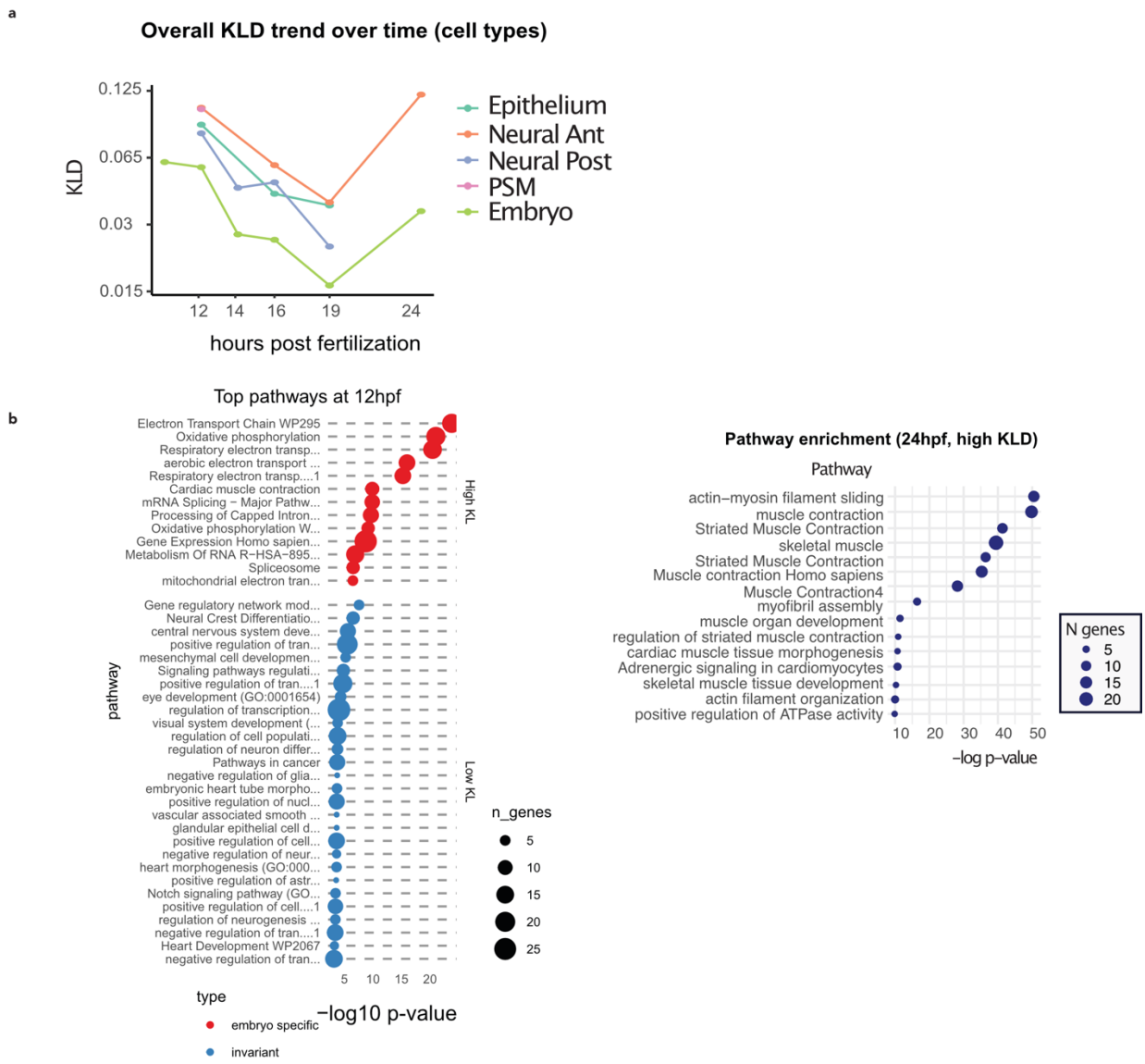


Supplementary Figure 7. Examples of genes with different KLD scores. Distributions of log normalized gene expression grouped by embryo for genes with increasing values of KLD (0.038:12.25). The heatmaps show KL(x,y) across all pairs of embryos (embryos compared to themselves along the diagonal have KL = 0). Each gene's KLD score (the median of the pairwise KL values, excluding the diagonal) is shown above the heatmap.



Supplementary Figure 8. KLD score control and distribution during development (continued next page).

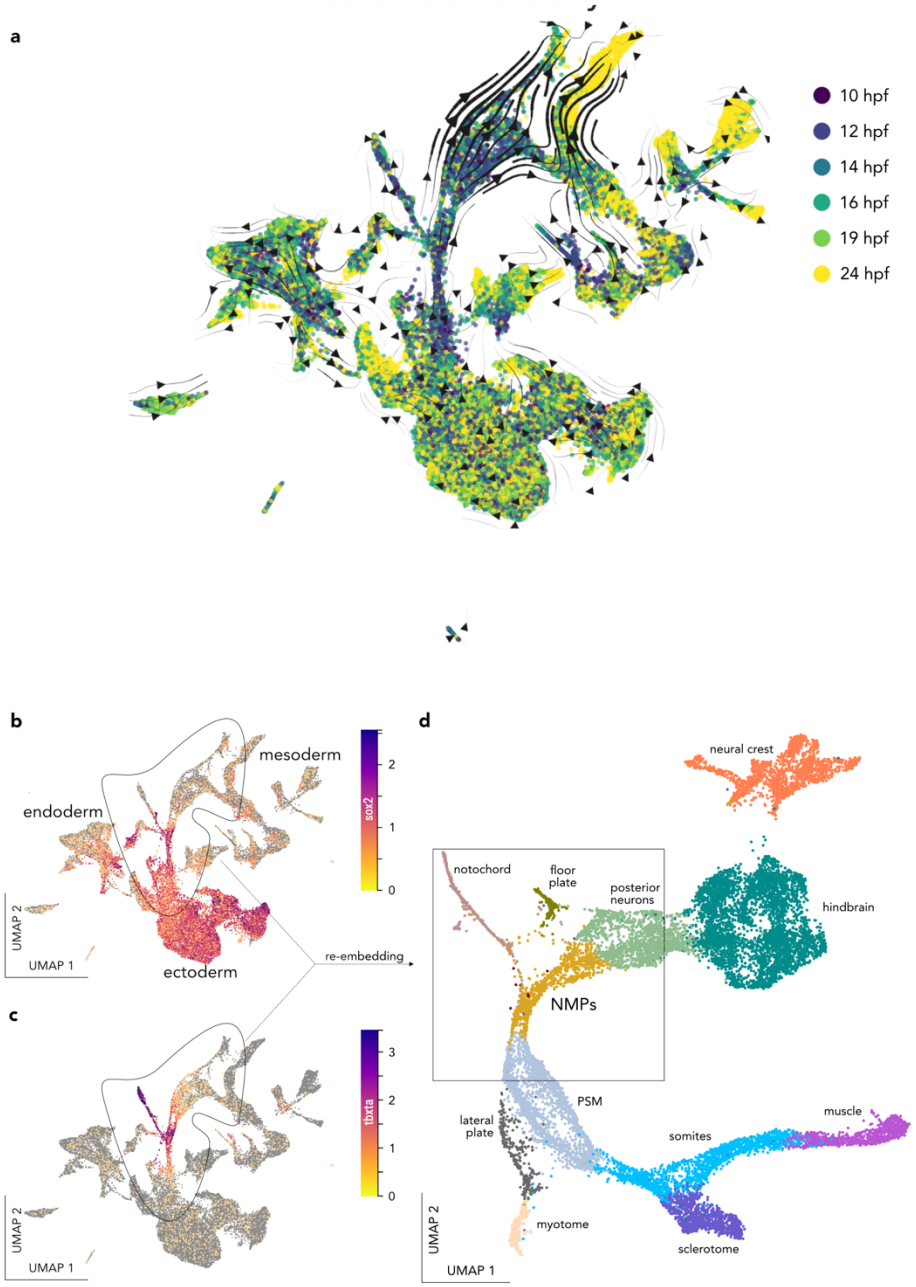
a. Negative control for the KLD scores. We generated a null distribution for the KLD scores by randomly reshuffling the embryo ID of individual cells. The null distribution allows testing the hypothesis that the differences observed between the four embryos can be attributed to random samples from the same distribution. For each gene, we randomized the embryo IDs of individual cells, re-calculated the KLD scores (red histograms), and compared them to the original distributions (blue histograms), repeating the randomization and calculation process 20 times. We calculated the null distributions for different cell types, including the whole embryo. The scatterplots show the z-scores, computed for each gene as the difference between the null distribution (20 null KLD scores for each gene) and their true KLD score. We found a strong positive correlation between the z-score and the KLD score and decided to use the latter. The histograms of z-scores show the fraction of genes with significant KLD (z-score >10). We show the null distributions and z-scores for the whole embryo (12hpf), the presomitic mesoderm (PSM, 12hpf), and the Neural anterior cells (16hpf) as representative examples.
b. KLD distributions across all genes for different cell types and time points. The histogram shows the KLD scores for all genes that passed the quality filters (fraction of cells with zero counts < 0.8, mean log-normalized expression > 0.2, at least 80 cells per fish for each cell type/timepoint) for that subset of cells. Missing histograms correspond to cell types that don't have enough cells per embryo (minimum of 80).



Supplementary Figure 9. KLD and enrichment pathway details overtime.

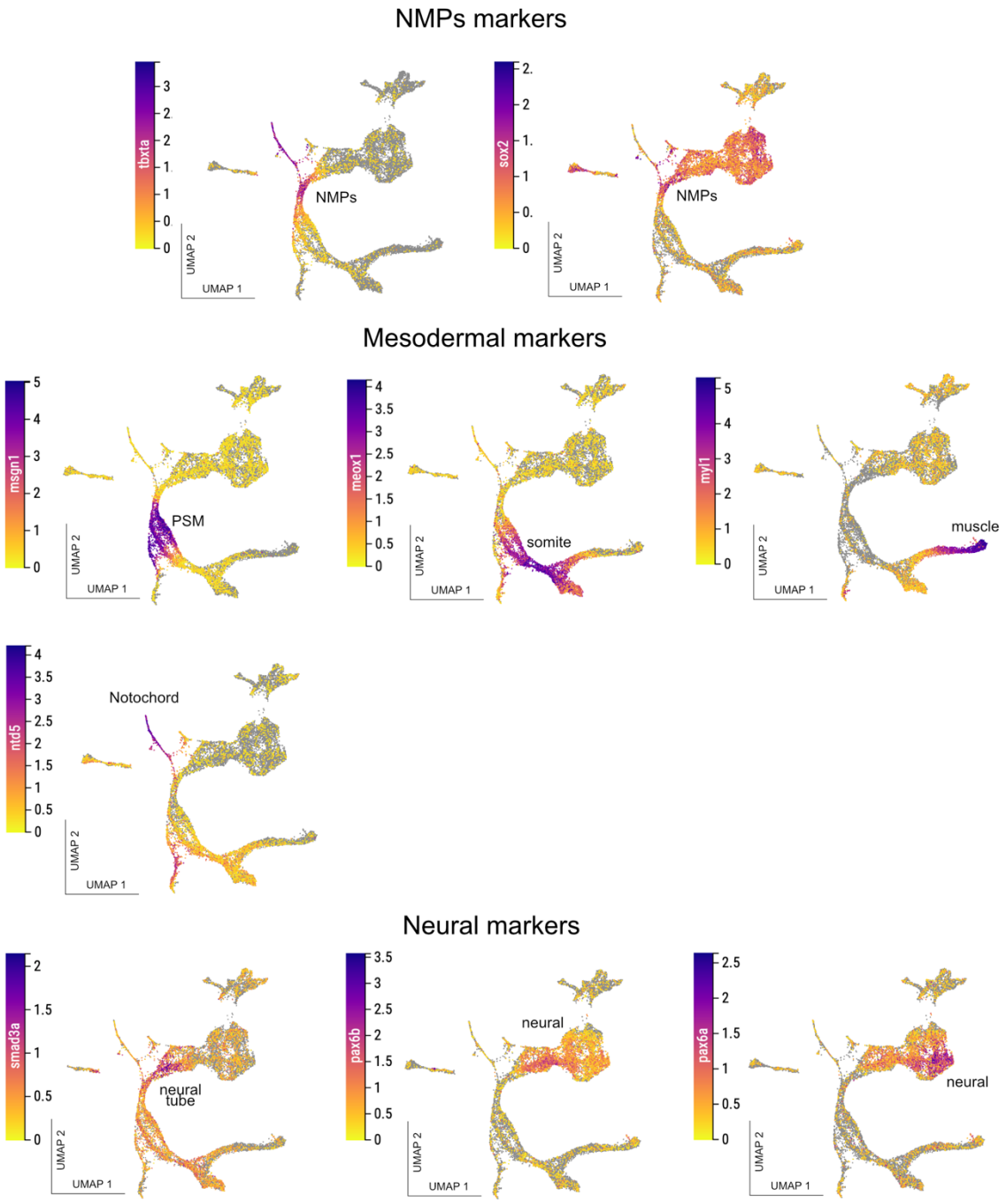
a. KLD trends for different cell types. The KLD median value across all genes for different cell types across time points. Each dot represents the median across all genes that passed the quality filters for that cell type. Some cell types do not have enough cells at some time points and therefore do not appear in the plot.

b. Top pathways at 12 and top pathways at 24. An extended list of enriched pathways in high and low KLD genes for the whole embryo at 12hpf. For 24hpf, we show the top enriched pathways in high KLD genes.

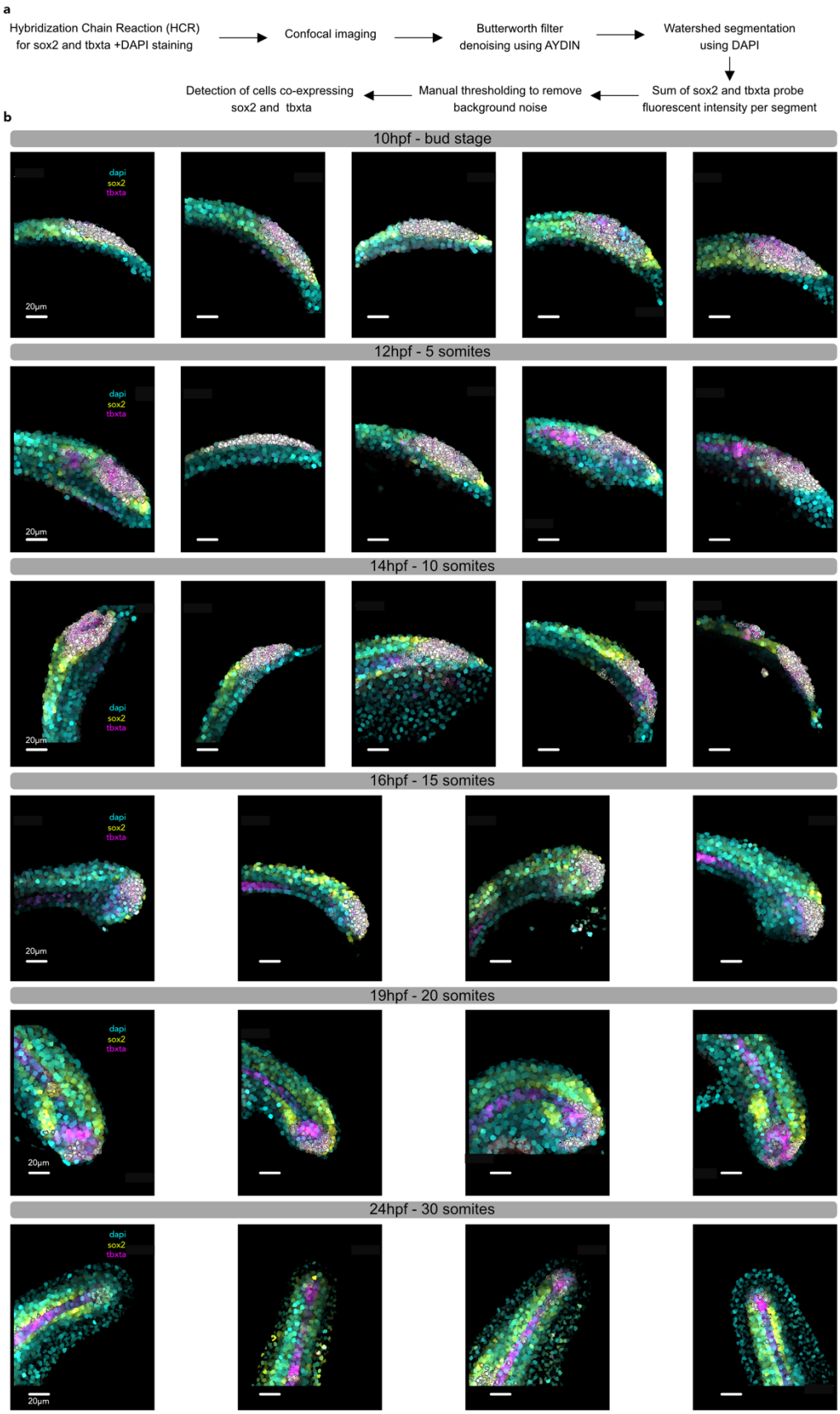


Supplementary Figure 10. Early time points UMAPs (10 hpf, 12 hpf, 14 hpf, 16 hpf, 19 hpf and 24hpf) and the NMP cluster.

a. UMAP of early zebrafish development (integration of 10 to 24 hpf time points in a single object), color codes by time points, the arrow streams correspond to the RNA velocities.
b. Early time points UMAP embedding color coded for sox2.
c. Early time points UMAP embedding color coded for tbxta.
d. The cluster co-expressing sox2 and tbxta (molecular definitions of the NMP cells) and there neighboring clusters are re embedded in a new UMAP.



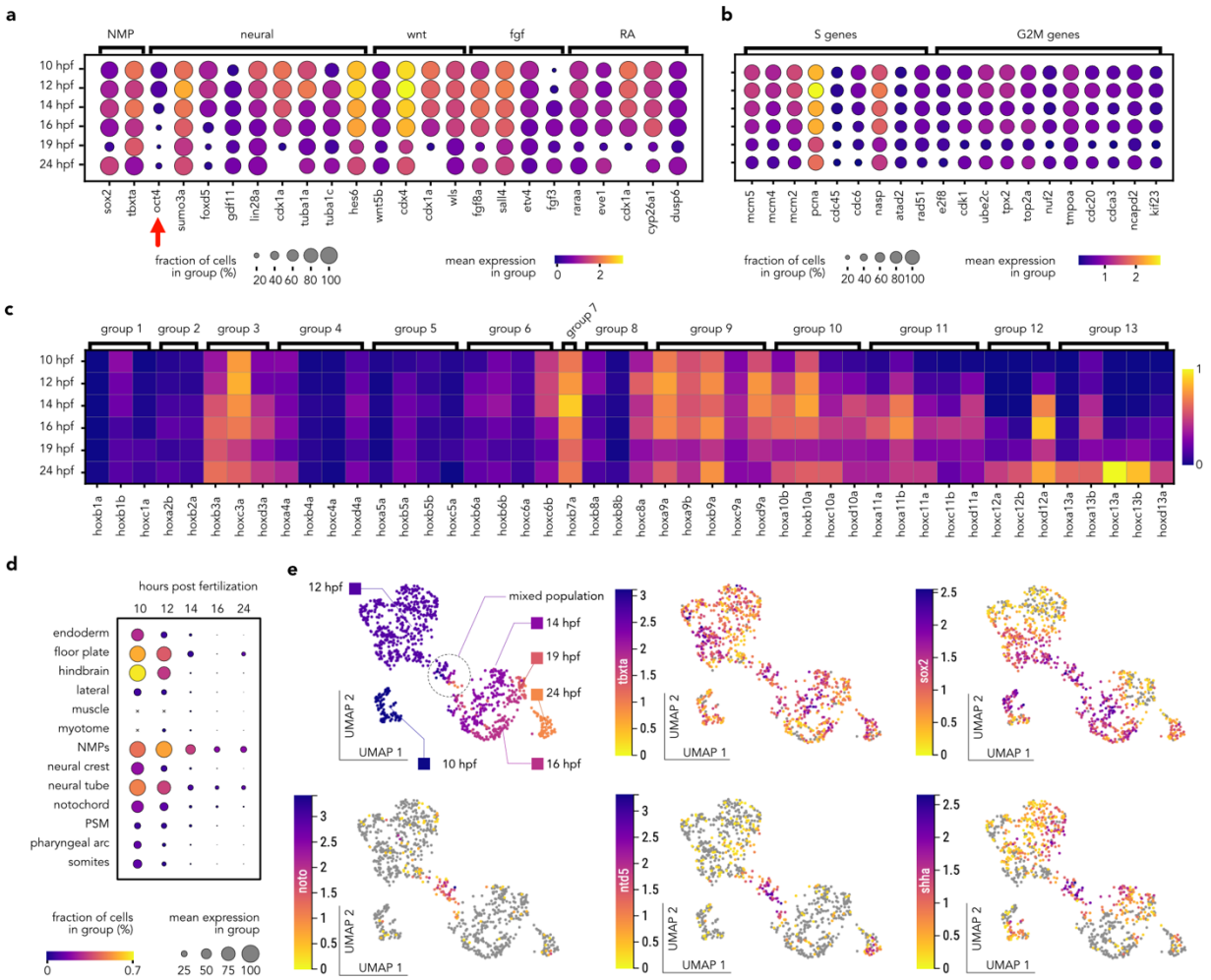
Supplementary Figure 11. NMP cells UMAP embedding color-coded for top genes important for the NMPs (top), mesodermal differentiation (middle) and neural differentiation markers (bottom).



Supplementary Figure 12. Spatial identification of NMPs presumptive territory using Hybridization Chain Reaction (HCR) during somitogenesis (continued next page).

a. The pipeline used to identify the NMP, image the zebrafish tail, segment the DAPI channel, and finally identified the NMP (co-expressing *sox2* and *tbxta*).

b. Processed images showing the evolution over time of the cells co-expressing *sox2* and *tbxta* - NMP presumptive territory.



Supplementary Figure 13. NMP cells molecular fingerprints extracted from the scRNAseq.

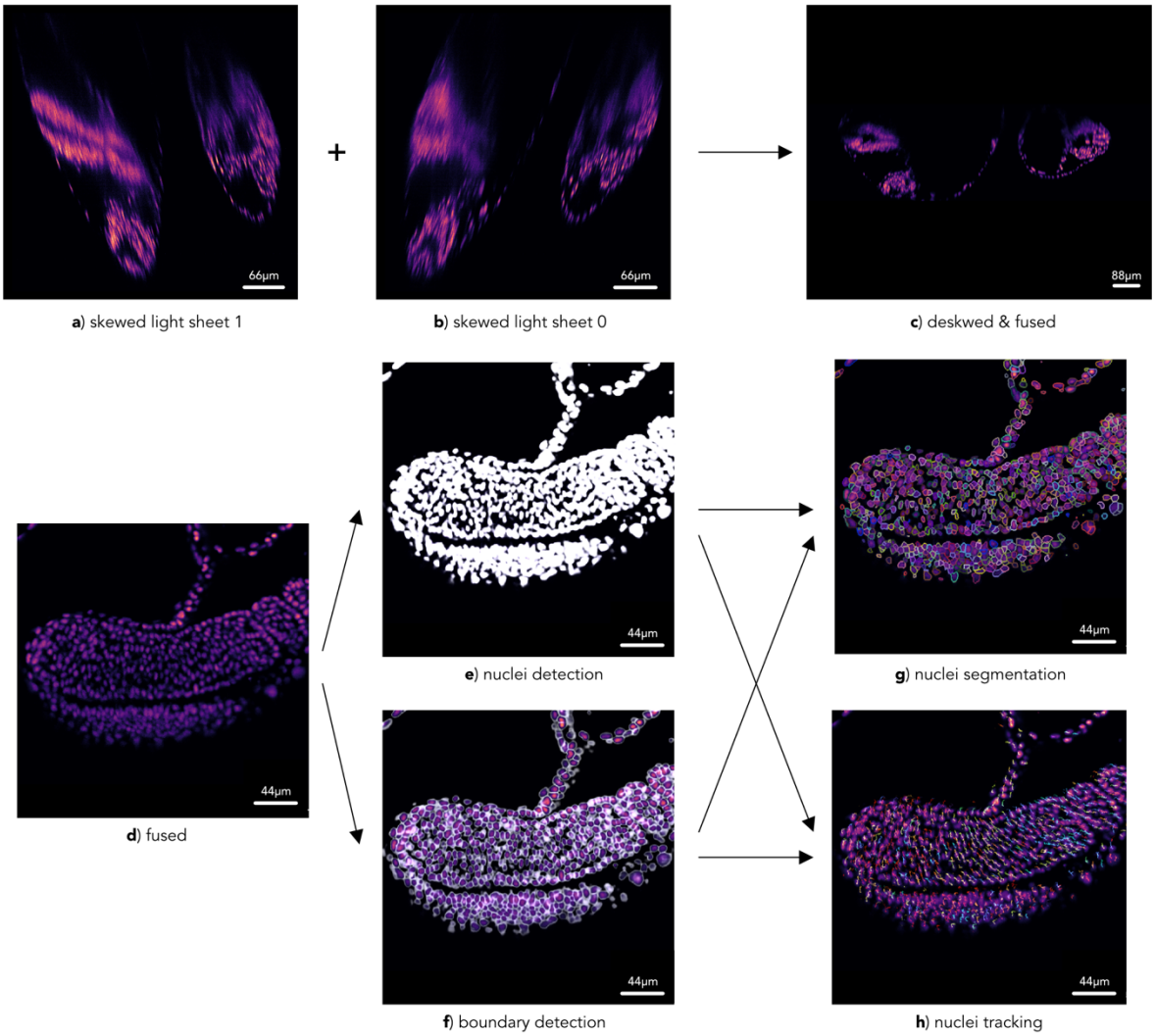
a. Selection of NMPs enriched genes for signaling pathway important for vertebrates' axial elongation showing a peak of expression early during development.

b. Selection of important genes for cell cycle enriched in the early NMPs.

c. Dotplot of the zebrafish hox genes expression overtime in the NMP clusters, showing a collinear activation of hox pathway.

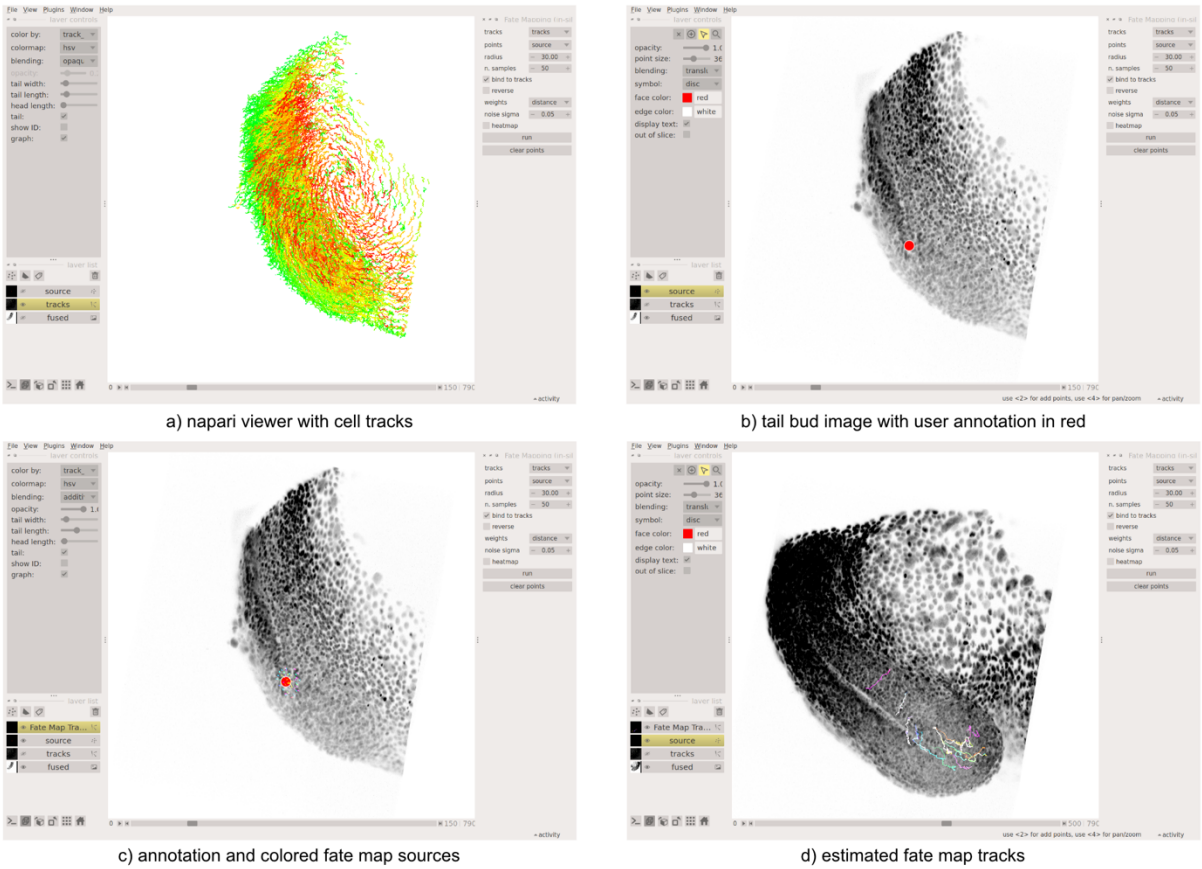
d. Zebrafish *oct4* (*pou5f3*) timeline of spatial expression in the posterior body clusters, showing late maintenance of *oct4* specifically in the NMP cells.

e. UMAP embedding of the NMPs cluster in the early integrated UMAP (10hpf, 12hpf, 14hpf, 16hpf, 19hpf, and 24hpf) color-coded by time point (top left). The same embedding is then color coded for the NMP molecular signature genes (*sox2* and tbxta) and then for three genes very important for notochord development.



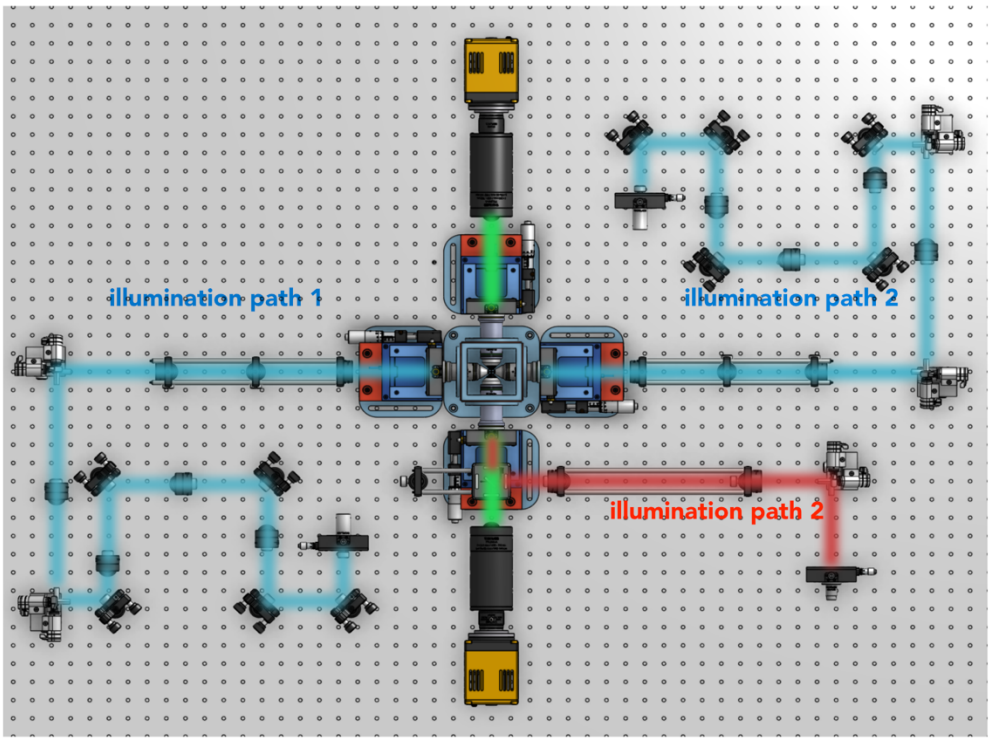
Supplementary Figure 14. Single-objective light-sheet microscope image processing pipeline.

a. Maximum intensity projection skewed views from light sheets 1 and 2;
b. Same as a.
c. Maximum intensity projection of deskewed and fused volume.
d. xy-slice of embryo's tail.
e. Neural network cell detection of region.
f. (Neural network cell boundary prediction of region (d). (g, h) Results from joint segmentation and tracking.
g. Segmentation masks.
h. Respective tracks.

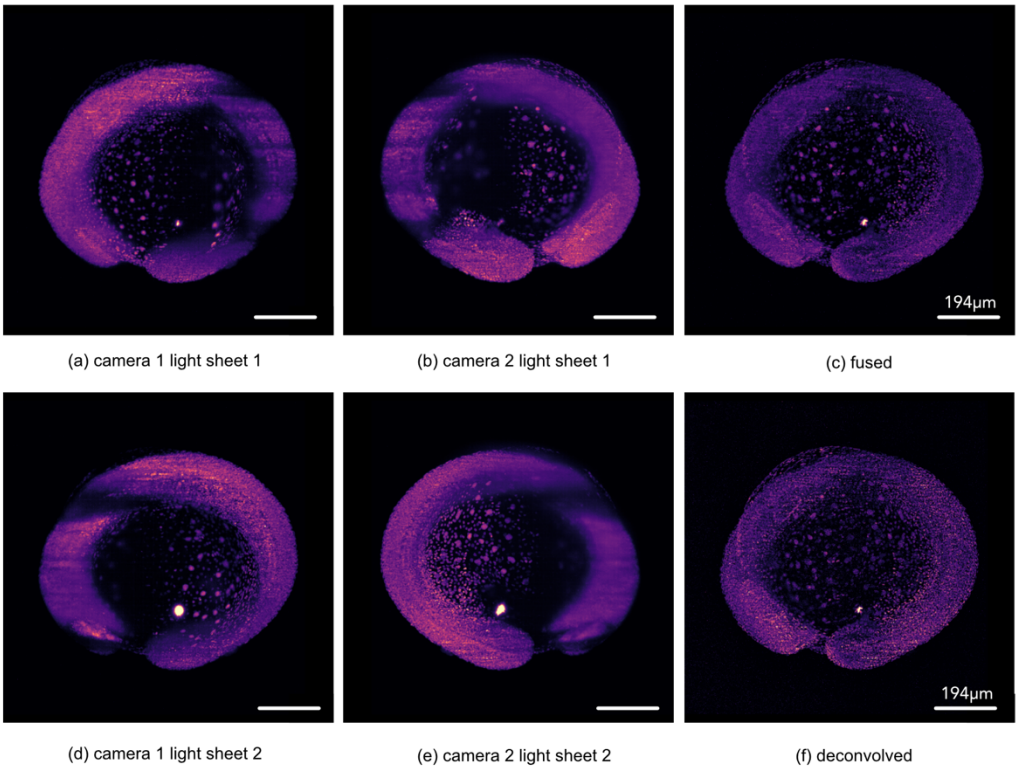


Supplementary Figure 15. The *in silico* tracking plugin for Napari.

a. Visualization of cell tracks in Napari.
b. Selection of region of interest.
c. Propagation of stepwise local movement estimation.
d. Display of the estimated fate map tracks.

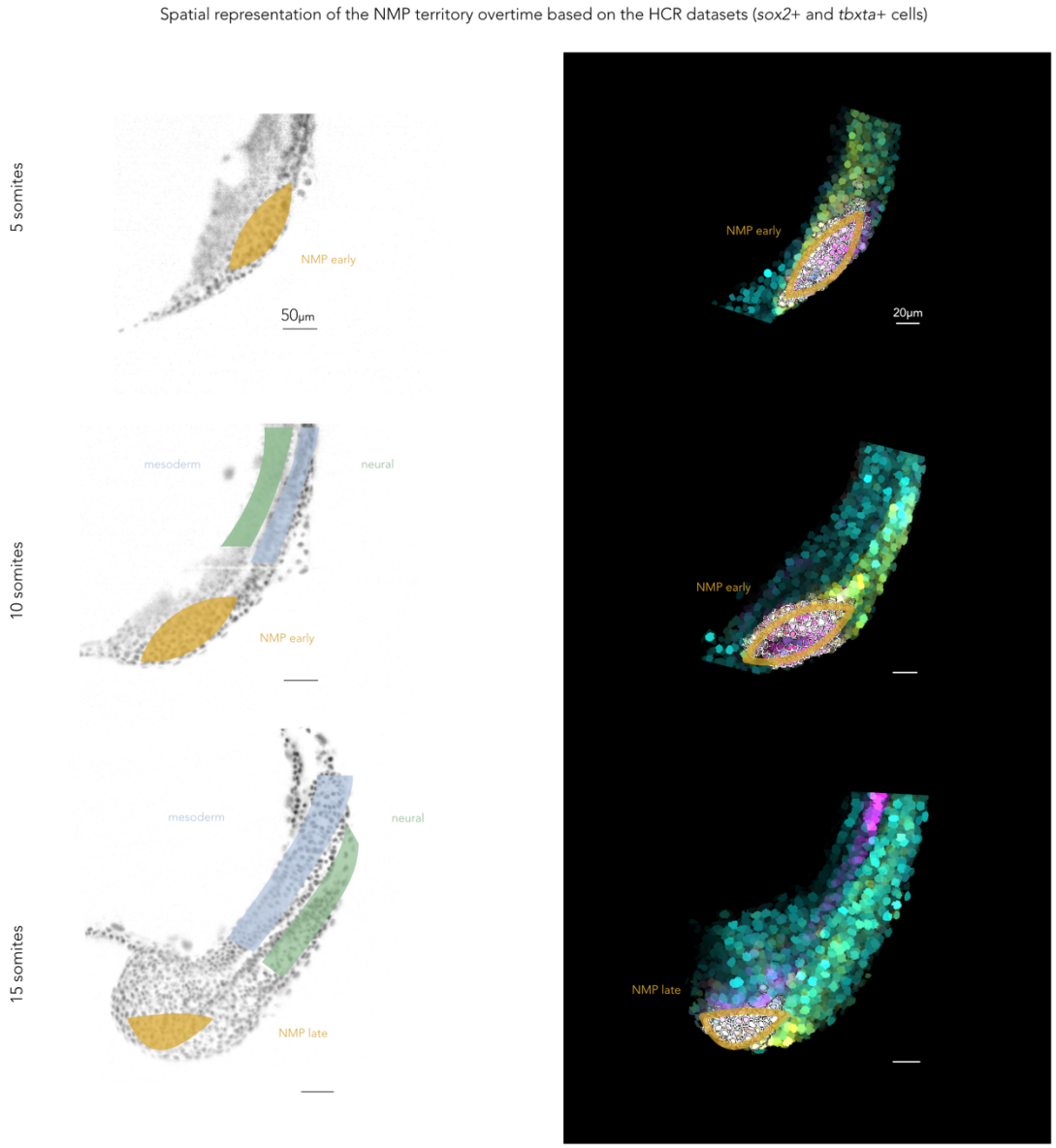


Supplementary Figure 16. Illustration of the illumination paths (in blue) and photoconversion arm (in red), in our multiview light-sheet microscope.

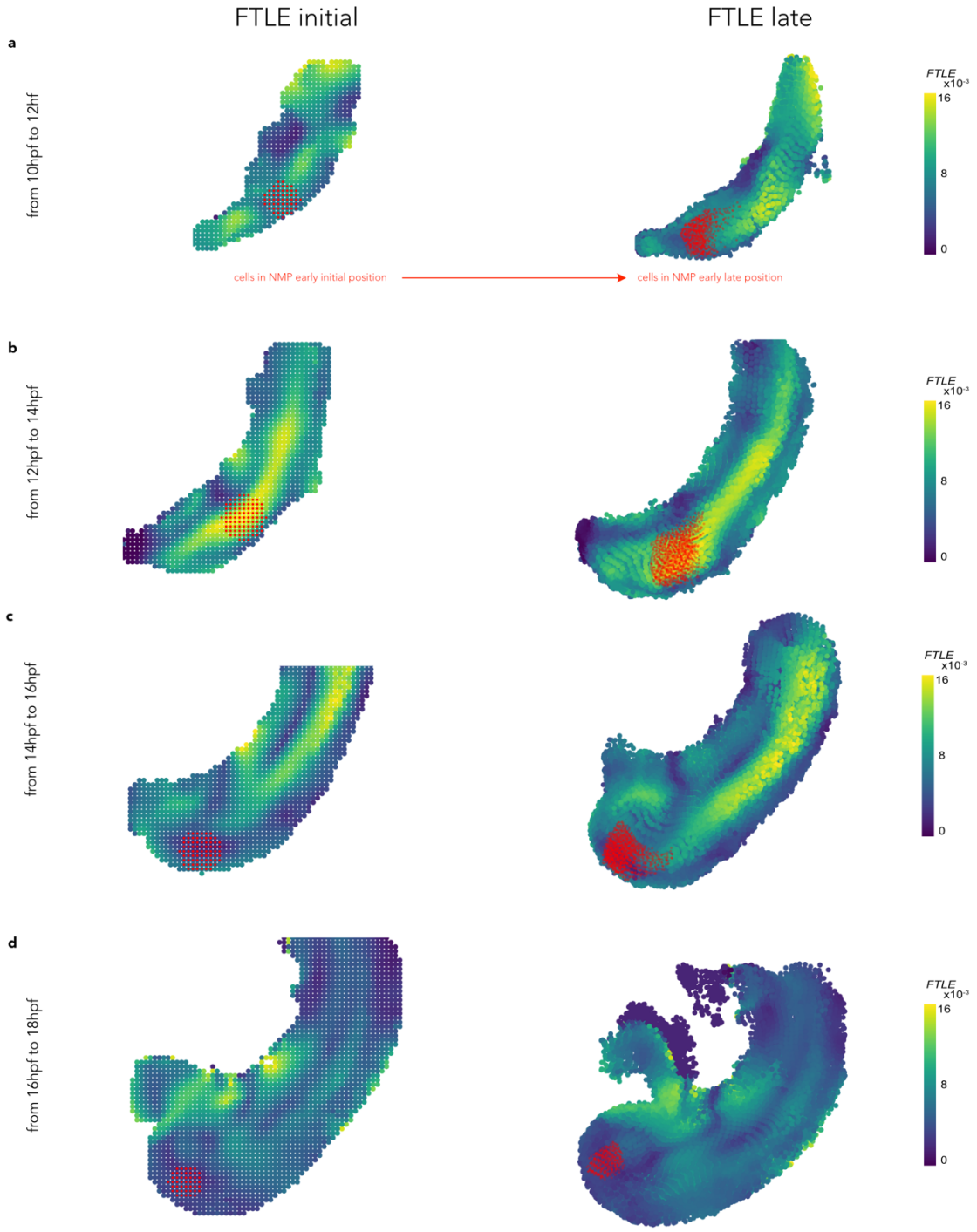


Supplementary Figure 17. Intermediate results of multi-view light-sheet microscope image processing pipeline. Volumes shown as maximum intensity projection on the z-axis of:

a. Camera 1 light sheet 1;
b. Camera 2 light-sheet 1;
c. Camera 1 light sheet 2;
d. Camera 2 light sheet 2;
e. Fused, camera 2 volumes are flipped before fusion.
f. Deconvolved volume after fusion.



Supplementary Figure 18. Spatial representation of the NMP presumptive territory used for the FTLE analysis is based on the HCR staining (cell double positive for *sox2* and *tbxta*), at 5somites or 12hpf, 10somites or 14hpf and 15somites or 16hpf.



Supplementary Figure 19. Temporal dispersion of putative NMP cells when computing the forward FTLE in Lagrangian Coherent Structures for different temporal windows. Labelled putative cells are represented in red in their initial location (FTLE initial) and their destination (FTLE late). In NMP early progeny is distributed in both mesodermal and neural tissue. In late NMP, the progeny only integrates the mesoderm.

SUPPLEMENTARY THEORY NOTE

Any framework to analyze spatio-temporal trajectories in morphogenesis requires a self-consistent description of cell motion that is independent of the choice of reference frame. This frame-invariant property, called objectivity[1], is a fundamental axiom of mechanics and ensures that the description of deforming biological tissue is independent of the coordinate frames we choose to describe its motion. Non-objective metrics, for instance, will yield different (inconsistent) results if cell velocities are described from a frame moving with a drifting embryo or a fixed lab frame (see e.g., Figure 1 in the reference[2]). The velocity field and the streamlines, which are trajectories of the frozen velocity field, are common nonobjective metrics[1].

To analyze cellular flows generated in the zebrafish tailbud, we use the recently developed Dynamic morphoskeletons (DM)[2], which is based on the notion of Lagrangian Coherent Structures (LCS)[3]. The DM is objective and founded on a Lagrangian description of tissue deformation captured by Finite Time Lyapunov Exponents (FTLE), which naturally combine local and global mechanisms along cell trajectories. Given a time interval $T = t_f - t_0$ of interest, the DM locates repellers from which cells diverge, attractors towards which cells converge, and their domain of attraction identifying the initial positions of the cells that will end up in the attractor. The DM reveals the intrinsic geometric feature of spatiotemporal trajectories and is robust to noise[2], hence it is ideal for quantifying morphogenesis.

Since morphogenetic flows do not undergo exponential stretching, here we identify the DM by computing the Lagrangian stretching rate, a simpler but equivalent version of the FTLE. Because we are specifically interested in the compartmentalization of the NMP region, we compute the forward Lagrangian stretching rate $\Lambda_{t_0}^{t_f}(\mathbf{x}_0)$ to identify the repelling spatiotemporal regions (repelling LCSs) in the embryo. Repellers are marked by high values of $\Lambda_{t_0}^{t_f}(\mathbf{x}_0)$. The Lagrangian stretching rate field $\Lambda_{t_0}^{t_f}(\mathbf{x}_0)$ computed for the time interval $[t_0, t_f]$ and initial point $\boldsymbol{x}_0$ is given by,

$$\Lambda_{t_0}^{t_f} = \frac{{}_{t_0}^{t_f}\lambda_2(\boldsymbol{x}_0)}{t_f - t_0},$$

where ${}_{t_0}^{t_f}\lambda_2(\boldsymbol{x}_0)$ denotes the largest singular value of the Jacobian of the flow map $\nabla \boldsymbol{F}_{t_0}^{t_f}(\boldsymbol{x}_0)$ and $\boldsymbol{F}_{t_0}^{t_f}(\boldsymbol{x}_0)$ is the flow map describing the cell trajectories from their initial position $\boldsymbol{x}_0$ to the final position $\boldsymbol{x}_f = \boldsymbol{F}_{t_0}^{t_f}(\boldsymbol{x}_0)$. We compute $\nabla \boldsymbol{F}_{t_0}^{t_f}(\boldsymbol{x}_0)$ from cell trajectories using a recently developed technique to identify LCSs from sparse and noisy trajectories[4]. This method relies on calculating the separation between neighboring trajectories over time. For an initial cell position located at $\boldsymbol{x}_0$, neighboring trajectories are those starting within a sphere of radius δ from the point $\boldsymbol{x}_0$. For all the results shown in this paper, $\delta = 100$. We use the publicly available code in the reference[4] and their strategy to select the optimal δ from data. The Lagrangian stretching rate $\Lambda_{t_0}^{t_f}(\mathbf{x}_0)$ describes a deformation rate, it has units of 1/time, and it quantifies the maximum Lagrangian deformation rate over the time interval $[t_0, t_f]$ of a region whose initial position is centered at $\boldsymbol{x}_0$ (Figure 1a). Nearby cells on the opposite sides of a repelling LCS are maximally separated (Figure 1b), facilitating cell populations to achieve differential mechanical environments for cell fate specification.

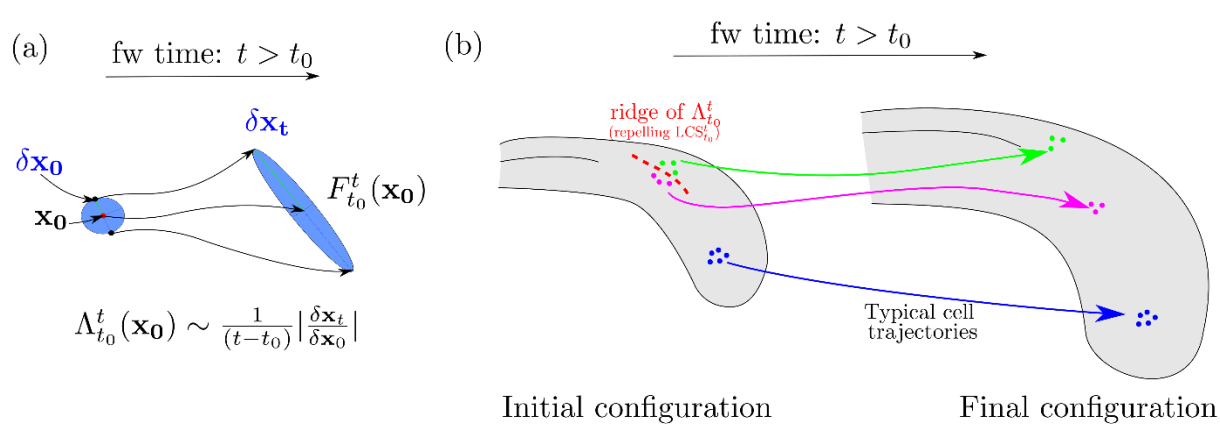


Fig 1: (a) The Lagrangian Stretching rate $\Lambda_{t_0}^{t}(\mathbf{x}_0)$ is a measure of the maximum separation generated by the flow over the time interval $[t_0, t]$ between two nearby points in the neighborhood of $\boldsymbol{x}_0$. (b) Ridges (ie sets of high values) of $\Lambda_{t_0}^{t}(\mathbf{x}_0)$ mark repelling LCS over $[t_0, t]$, which are the set of initial cell locations $\boldsymbol{x}_0$ where the Lagrangian stretching rate is maximum.

References

1. Truesdell, C., Noll, W., Truesdell, C., & Noll, W. (2004). *The non-linear field theories of mechanics* (pp. 1-579). Springer Berlin Heidelberg.
2. Serra, M., Streichan, S., Chuai, M., Weijer, C. J., & Mahadevan, L. (2020). Dynamic morphoskeletons in development. *Proceedings of the National Academy of Sciences*, *117*(21), 11444-11449.
3. Haller, G. (2015). Lagrangian coherent structures. *Annual review of fluid mechanics*, *47*, 137-162.
4. Mowlavi, S., Serra, M., Maiorino, E., & Mahadevan, L. (2022). Detecting Lagrangian coherent structures from sparse and noisy trajectory data. *Journal of Fluid Mechanics*, *948*, A4.