Joint profiling of gene expression and chromatin accessibility of amphioxus development at single cell resolution

Pengcheng Ma
Kunming Institute of Zoology

Xingyan Liu
Academy of Mathematics and Systems Science, CAS

Huimin Liu
School of Life Sciences, Xiamen University

Zaoxu Xu
Beijing Genomics Institute

Xiangning Ding
BGI-Shenzhen, Shenzhen 518083

Zhen Huang
Fujian Normal University https://orcid.org/0000-0002-8908-0355

Chenggang Shi
School of Life Sciences, Xiamen University

Langchao Liang
BGI-Shenzhen, Shenzhen 518083

Luohao Xu
University of Vienna https://orcid.org/0000-0002-3714-8047

Xiaolu Li
Kunming Institute of Zoology, CAS

Guimei Li
Kunming Institute of Zoology, CAS

Yuqi He
Kunming Institute of Zoology, CAS

Zhaoli Ding
Kunming Institute of Zoology, CAS

Chaochao Chai
BGI-Shenzhen, Shenzhen 518083

Haoyu Wang
BGI-Shenzhen, Shenzhen 518083

Jiaying Qiu

BGI-Shenzhen, Shenzhen 518083

Jiacheng Zhu
BGI-Shenzhen, Shenzhen 518083

Xiaoling Wang
BGI-Shenzhen

Peiwen Ding
BGI-Shenzhen, Shenzhen 518083

Si Zhou
BGI-Shenzhen

Yuting Yuan
BGI-Shenzhen

Wendi Wu
BGI-Shenzhen

Yanan Yan
Life Science College, Fujian Normal University

Yitao Zhou
Fujian Key Laboratory of Developmental and Neuro Biology, College of Life Sciences, Fujian Normal University, Fuzhou, 350117

Qi-Jun Zhou
Kunming Institute of Zoology, CAS

Guo-Dong Wang
Chinese Academy of Sciences https://orcid.org/0000-0002-9407-4363

Qiujin Zhang
College of Life Sciences, Fujian Normal University

Xun Xu
BGI-Shenzhen https://orcid.org/0000-0002-5338-5173

Guang Li
Xiamen University https://orcid.org/0000-0002-5543-5349

Shihua Zhang
zsh@amss.ac.cn https://orcid.org/0000-0003-0192-7118

Bingyu Mao (✉ mao@mail.kiz.ac.cn)
Kunming Institute of Zoology

Dongsheng Chen
BGI-Shenzhen

Article

Keywords: Amphioxus, snRNA-seq, scATAC-seq, developmental embryos, adult tissues

Posted Date: May 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-504113/v1

Abstract

Vertebrate evolution was accompanied with two rounds of whole genome duplication followed by functional divergence in terms of regulatory circuits and gene expression patterns. As a basal and slow-evolving chordate species, amphioxus is an ideal paradigm for exploring the origin and evolution of vertebrates. Single cell sequencing has been widely employed to construct the developmental cell atlas of several key species of vertebrates (human, mouse, zebrafish and frog) and tunicate (sea squirts). Here, we performed single-nucleus RNA sequencing (snRNA-seq) and single-cell assay for transposase accessible chromatin sequencing (scATAC-seq) for different stages of amphioxus (covering embryogenesis and adult tissues). With the datasets generated we constructed the developmental tree for amphioxus cell fate commitment and lineage specification, and revealed the underlying key regulators and genetic regulatory networks. The generated data were integrated into an online platform, AmphioxusAtlas, for public access at http://120.79.46.200:81/AmphioxusAtlas.

Introduction

The origin of vertebrates has been a central topic for centuries in the field of evolutionary and developmental biology. Cephalochordates (commonly called amphioxus or lancelets) and urochordates (also known as tunicates) are sister groups of vertebrates, and thus are key animals for understanding the evolution of vertebrates. However, during the past 500 million years, while the tunicates have evolved very derived morphology and genome through rapid evolution[1,2], amphioxus has evolved at very slow rate and retained most morphological and genomic features of ancestral chordates[3,4]. Moreover, compared to tunicate embryogenesis, which is largely determinant with early decision of cell fates, amphioxus development is highly regulative as that of vertebrates[2]. In line with these, developmental and genomic studies have suggested that the gene regulatory network governing amphioxus early development is highly homologous to those in vertebrates[4].

Transcriptomic and chromatin accessibility profiling at single cell level have been increasingly used to dissect the early developmental program in model and non-model organisms[5-8]. To dissect the mechanisms of cell lineage specification and the logic of the regulatory network of amphioxus embryos and shed insights into the origin of vertebrates, we performed snRNA-seq and scATAC-seq of amphioxus *Branchiostoma floridae* embryos from late blastula to early larval stages. From these datasets, the developmental trajectory of the different cell lineages and the underlying regulatory modules were resolved based on expression of canonical marker genes and corresponding chromatin states. Single-nucleus expression profiling was also performed for different amphioxus adult tissues. A comprehensive online platform was developed for storage and public access of these datasets.

Results

A single-cell transcriptional atlas of the amphioxus embryo

In amphioxus, the fertilized egg cleaves rapidly to form the blastula as a single-layered ball of cells[9], which then gastrulates by simple invagination of the vegetal pole inward the animal pole. Using the SPLiT-seq protocol[10], we performed snRNA-seq sequencing of amphioxus (*Branchiostoma floridae*) embryos across nine stages from late blastula to early larva (Figure 1A, B), profiling a total of 148, 875 single cells. After quality control, data from 7508 to 24383 individual cells for each stage were used for further analysis (Figure 1B), with median genes ranging from 254 to 764 *per* cell (Supplementary Table S1).

Projection of these transcriptomes using the Uniform Manifold Approximation and Projection for dimensionality reduction (UMAP)[11] identified coherent clusters of cells at different stages or when pooled together (Figure 1B and Supplementary Figure 1). Differentially expressed cluster-specific genes were identified using Model-based Analysis of Single-cell Transcriptomics (MAST)[12] and the identities of each cluster were annotated based on the expression of key marker genes (Figure 1B, C).

To remove the batch-effects from different stages during constructing the UMAP atlas, we develop a simple algorithm to construct the single-cell network as a replacement of the *k*-nearest-neighbor (KNN) searching step of UMAP and to facilitate lineage-tree construction (Materials and Methods). The two-dimensional (2D) projection of the single-cell network shows that the cells were clearly clustered into different lineages with preservation of the order of developmental stages (Figure 1B), as suggested by the expression of the lineage-specific markers (Figure 1C, *Wnt8* for mesoderm, *Sox17* for endoderm, *Hu-Elav* for neural ectoderm, *Chrd* for notochord, *FoxJ1* and *Keratin* for epithelial ectoderm[13-15]. A virtual lineage tree was constructed by taking the pre-defined clusters in each stage as nodes and connecting nodes across time points by “ancestor voting” based on the single-cell network (Materials and Methods, Figure 1D, E). At late blastula stage, the primordial germ cells (PGCs), and endodermal trajectories were already distinguished, while the ectodermal and mesodermal lineages were clustered together, which then broke into the epithelial ectodermal and neural ectodermal, and notochord and somite mesodermal lineages, respectively, at early gastrula stage (G3). On the lineage tree, the notochord lineage was well established early at the mid-gastrula stage (G3), which then broke into two lineages at early neurula stage (N1). However, plotting of the cell clusters suggested that the two clusters more likely represent different phases of the same cell population rather than two different sub-lineages (Supplementary Figure 2). The somite mesodermal lineage (G3.2) broke into 3 lineages, with the G5.7 lineage representing the anterior pharyngeal mesoderm (expressing *Eya*, *Six4/5*, and *Pax3/7*)[16,17], the G5.3 lineage representing the posterior pharynx mesoderm (expressing *Hox11*)[18] and the G5.5 lineage likely representing the posterior tailbud mesoderm, which expresses the posterior marker *Cdx* (Figure 1E and Supplementary Figure 3)[19]. The endodermal lineage (G3.1) differentiated into the anterior pharynx lineage (expressing *Six3/6* and *Nkx2.1*)[17,20] and posterior gut lineage (expressing *Pax1/9*, *Ilp* and *Msxlx*)[21-23], which further differentiated into fore-, mid-, and hind-gut lineages at L0 stage (Figure 1E and Supplementary Figure 4). The neural ectodermal lineage differentiated into three populations at N3 stage, with the N3.4 and N3.7 representing the anterior (expressing higher *Otx, Fezf, Pax4/6, Six3/6, Lhx2/9, OligA, OligB/C* and *Brn1/2/4*)[17,24-27] and posterior (expressing higher *Gbx, Wnt7b, Pax2/5/8, Cdx, Hox1, Hox3* and *Msx*) cell populations[18,19,25,28-

[31]. The N3.12 lineage represents differentiated neurons, expressing highly the neuronal markers and transporters, including *Pouf1, Hu_Elav, Tlx, VaChT, VGLUT, VGAT, ChAT*, *etc* (Figure 1E and Supplementary Figure 5)[14,27,32-34]. The identity of the G3.5 lineage was not assigned, which expresses highly cell cycle related genes with few tissue specific markers (Supplementary Table 2).

Single-cell chromatin accessibility profiling of amphioxus development

We performed single-cell ATAC-seq from samples ranging from gastrula to larva, and obtained data for around 37,000 cells passing quality control measures (Figure 2A and Supplementary Table 3). Specifically, the chromatin accessibility status of 1535, 10091, 6398, 9703, 5225 and 4880 cells were analyzed from blastula, early gastrula, late gastrula, early neurula, middle neurula and larva, respectively (Figure 2A and Supplementary Table 3). Genomic feature analysis showed more than 25% of peaks fell within the promoter regions of genes (Supplementary Figure 6). Low-dimensional visualization of the chromosome accessibility atlas at six stages revealed increasing complexity during embryogenesis (Figure 2B).

To integrate the cells from snRNA-seq and scATAC-seq, the latter was first transformed into gene activities using Cicero[35] and combined with the transcriptomic data of the same stage. Then data integration and label transfer were conducted using Harmony followed by a *k*-nearest-neighbor classifier[36]. UMAP visualizations showed that cells of the same type were projected proximately in both independent and integrated analysis, confirming the reliability of our procedure (Supplementary Figure 7). Next, we extracted the ATAC-seq data for different clusters to reveal the cell type-specific chromatin accessibility. As expected, the chromosome opening status for the selected marker genes were well correlated with their lineage-specificity (Figure 2B).

Verification of cell lineages by *in situ* hybridization

Next, we selected the intersection of differentially expressed genes (DEGs) and genes associated with differentially accessible peaks of the same cell lineage at gastrula and neurula stage as the lineage-specific genes, resulting in 19 potential marker genes for endoderm lineage, 57 potential marker genes for epithelial ectoderm lineage, 9 potential marker genes for neural ectoderm lineage, 68 potential marker genes for mesoderm lineage and 25 potential marker genes for notochord mesoderm lineage (Supplementary Table 4). To verify the marker genes identified in above analysis, we performed *in situ* hybridization with probes for 14 genes which have not been studied before (Figure 3A). Among them, *BfAlcam* is enriched in neural ectoderm cell lineage, *BfTsta3*, *BfMlp*, *BfGrb1l*, *BfLac* and *BfRbr* in epithelial ectoderm cell lineage, *BfCreb3l2*, *BfCol11a1*, *BfRbms1* and *BfSlit2* in notochord mesoderm cell lineage, and *BfSlc6a15*, *BfC6*, *BfMim_l256* and *BfHmcn* in endoderm cell lineage (Figure 3A and Supplementary Figure 8). The results show that all the 14 genes analyzed show consistent and specific expression in the corresponding cell lineages as deduced by the transcriptomic and ATAC-seq data (Figure 3， Supplementary Figure 7 and Supplementary Table 4, 5). For example, according to our *in silico* analysis, *BfAlcam* is predominantly presented in the neural ectoderm cell lineage from G6 stage, and weakly in the

the epithelial ectoderm cell lineage from N1 stage (Figure 3A). Indeed, the *in situ* hybridization results of *BfAlcam* showed an expression pattern exactly fitting with our prediction at these stages (Figure 3B), supporting our cell lineage assignment from the transcriptomic and ATAC-seq data.

The timing of zygotic genome activation in amphioxus embryos

During early embryonic development, the zygotic genome is gradually activated and the maternal factors are cleared during the process known as maternal-to-zygotic transition. In many species, the zygotic genome activation (ZGA) is characterized by a minor wave during the early cleavage stage followed by a major wave when the cell division slows down[37]. The timing of ZGA varies widely across animals. Here, we attempted to infer the timing of ZGA in amphioxus based on our snRNA-seq data. We plotted the number of detected genes at each stage. There is a slight increase at 32-cell stage and a sharp increase was observed at 32-cell to blastula stage (Figure 4A), suggesting that ZGA occurred most likely during this period. We then examined the expression of known zygotic genes during early development and found that their expressions seemed to initiate at about 32-cell to blastula stage and climaxed at gastrula stage (Figure 4B). To assess the embryonic genome activation (EGA), the stage-specific highly expressed gene transcripts were extracted by comparing their transcriptome to those of previous neighboring stages respectively (Supplementary Table 6). GO enrichment analysis showed that the RNA metabolism related genes, which are known to be crucial for early embryonic development[38,39], were strongly activated at 32-cell stage (Figure 4C). Overall, these observations suggest that ZGA occurs at 32-cell to blastula stage in amphioxus. Interestingly, a group of neural related genes became activated at blastula stage (Figure 4C). This phenomenon most likely reflect a local planar neural induction event at the border of the notochordal mesoderm and the ectoderm, as the notochordal mesoderm has not invaginated to underline the ectoderm for the major neural induction event to occur at this stage

A single-nucleus transcriptomic atlas of amphioxus adult tissues

We also performed snRNA-seq of different amphioxus adult tissues, including neural tube, epidermis, notochord, endostyle, *etc,* to explore their cell populations (Figure 5A, B and Supplementary Table 7). Specifically, 16 and 7 cell clusters were identified in the neural tube and epidermis respectively (Figure 5B). Interestingly, the cluster 2 of epidermis showed more neural characters when compared with the neural clusters, with high expression of *Coe*, *MAP18* and *Hu-Elav* (Figure 5C, D). This group of cells most likely represents a group of epidermis–derived neural cells in the epidermis. During amphioxus embryonic development, the epidermal sensory neurons were among the earliest differentiated neurons, which appear at late neurula stage[40]. However, this group of cells was not detected in our developmental lineage analysis (Figure 1E). This could be due to either insufficient sequencing depth of our data, the cells could be clustered into the neural groups, or that there is no distinct difference between epidermal neurons and other cells in epidermis. A cell lineage tracing experiment should help to clarify this issue.

A multi-omics resource center for amphioxus

To fully exploit the valuable resources generated here, we developed a dedicated and comprehensive online platform named AmphioxusAtlas (freely accessible at http://120.79.46.200:81/AmphioxusAtlas), enabling the browsing and querying of processed single-cell data from blastula to larvae and bulk data from gametes to adults. AmphioxusAtlas is mainly composed of five functional modules: scTranscriptomic, scEpigenetic, Trajectory, BulkRNAseq and TissueAtlas. In the scTranscriptomic module, we presented the UMAP plot of scRNAseq data for all the stages, along with differentially expressed genes (DEGs) in each cluster, and GO/KEGG terms enriched in corresponding DEGs. Additionally, a query box was provided for users to search the featureplot and boxplot for the expression pattern of any gene of interest. In the scEpigenetic module, users could browse the basic chromatin accessibility information (including the genomic coordinates for ATAC peaks, putative target genes, and transcription factor motifs enriched in the peaks), and the ATAC signatures for specific genes in each cluster could be queried. In the trajectory module, users could search for the chromatin accessibility and expression profile of a given gene simultaneously to trace their dynamic changes during cell fate commitment and lineage specification. In the bulkRNAseq module, users could obtain the longitudinal characteristics of genes during distinct developmental time points. In TissueAtlas module, the gene expression patterns in different tissues of adult amphioxus were demonstrated.

Discussion

Overall, we generated the first single cell atlas for amphioxus, a basal chordate lineage of particular value for studying the evolution and development of vertebrate. These results could be valuable for understanding the origin and evolution of chordate and throw light upon the genetic and epigenetic mechanism underling phenotype novelty of chordate during evolution. Embryo development relies on precise temporal regulation of various transcription factors, which subsequently activate the expression of hundreds of downstream targets. Here, we profiled 148, 875 cells from nine developmental stages and constructed the first amphioxus developmental trajectory at single cell resolution, systematically depicting the molecular dynamics underlying developmental amphioxus. On top of that, we identified a variety of key regulators and developmental genes during amphioxus embryogenesis, including both well-known transcription factors essential for embryonic development in vertebrates and tunicates, as well as novel genes with previously unrecognized functions.

Gene expression regulation is a fundamental question in developmental biology, as developmental patterns are largely determined by the restricted spatial-temporal expression domains of developmental genes, orchestrated by the dynamic interactions between a variety of transcription factors and corresponding *cis* regulatory elements (CREs). Traditionally, CREs have been detected using time-consuming experimental methods such as enhancer trapping, systematic sequence deletions and mutations. Later on, ATAC-seq opens the gate to screen for putative CREs at genome-wide scale and scATAC-seq enables researchers to explore the heterogeneity of CREs among distinct lineages and stages. Our study painted a systemic epigenetic portrait for the sequential linage specification of endoderm, mesoderm, notochord, neural ectoderm, epithelial ectoderm, and primordial germ cells.

As another proto-vertebrate group, the early development of ascidians is largely determinant and the embryonic cell lineages can be clearly traced back to individual blastomere at the initial gastrula stage[8]. Compared with ascidians, which shows increasing neural lineages at neurula and tailbud stages, amphioxus embryos showed much less complexity in the early neural development. As a basal chordate, amphioxus holds the key to understand the origin and evolution of vertebrate cell types. Although awaits further annotation, our single-cell transcriptomic and epigenomic analysis of amphioxus embryos and adult tissues lay the basis for future related studies.

Materials And Methods

1. Animal husbandry

Amphioxus *Branchiostoma floridae* were obtained from Dr. Jr-Kai Yu's laboratory at the Institute of Cellular and Organismic Biology, Academia Sinica, Taiwan. They were maintained and induced to spawn following the protocol as we described and used in B. belcheri[41,42]. Eggs fertilization and embryos culture were carried out according to our previous report[43]. Embryos are staged according to a recent study[9].

2. Single cell RNA sequencing library preparation

We prepared snRNA-seq libraries on the Split-seq platform[10]. Embryos at indicated stages were harvested and stored in RNAlater solution (AM7020, Ambion). Nuclei extraction was performed as described[10]. Briefly, indicated embryos were transferred into a 1 mL Dounce homogenizer containing 1mL homogenizing buffer (250 mM sucrose, 25 mM KCl, 5 mM $MgCl_2$, and 10 mM Tris [pH=8.0]; 1 μM DTT, RNase Inibitor and 0.1% Triton-X100). 5-10 strokes of loose pestle followed by 10-20 strokes of tight pestle were performed. The homogenates were filtered with a 40 μm strainer into 5 mL Eppendorf tubes and then spun down for 4 minutes at 600 g at 4^oC. The pellet was re-suspended and washed in 1 mL of PBS containing RNase inhibitors and 0.1% BSA. At last the nuclei were passed through a 40 μm strainer again before being counted. The nuclei were split into 48 wells, each containing barcoded well-specific reverse transcription primer, for in-cell reverse transcription. The second and third barcoding consist of a ligation reaction. After the third round of barcoding, the nuclei were divided into 16 aliquots and then lysed before cDNA purification. Purified cDNA was subjected to template switch and a round of real-time PCR amplification. PCR reactions were stopped at the beginning of plateau stage. At last, 600 pg purified PCR products each were used to generate Illumina compatible sequencing libraries. Each library was labeled by a distinct, indexed PCR primer pair, which is served as the fourth barcode.

3. Single-nucleus RNA sequencing and data processing

Libraries were sequenced on NextSeq systems (Illumina) using 150 nucleotide kits and paired-end sequencing. Read 1 covered the transcript sequences and read 2 covered the UMI and UBC barcode combinations. Firstly, we added the fourth barcode (sequencing index, 6 nt) at read 2 ends, then discarded reads which had more than one mismatched base with the third barcode. Thirdly, any reads in

UMI region had more than one low quality base (phred <=10) were also discarded. The sequencing results were aligned to exons and introns in the reference genome (https://www.ncbi.nlm.nih.gov/assembly/GCA_015852565.1/) and aggregated intron and exon counts at the gene level were calculated by kallisto and bustools software as described (https://bustools.github.io/BUS_notebooks_R).

4. Single-cell ATAC-seq library preparation

The embryos harvested at indicated stages were lysed in cold lysis buffer (10 mM Tris-HCl [pH 7.4], 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20, 0.1% NP40, 0.01% Digitonin and 1% BSA). The nuclei were extracted by gentle pipetting for 10 times. Larvae were homogenized in 2 mL Dounce homogenizer (SIGMA) containing 2 mL lysis buffer. Dounce homogenization and filtration were performed as described above[10]. Nuclei were pelleted by spinning at 500 g for 5 min at 4^{o}C. Then nuclei were washed twice by suspending pellet in chilled PBS (Gibco) with 0.04% BSA (BBI). The nuclei were re-suspended in diluted nuclei buffer (10x Genomics). The 10x ATAC libraries were constructed according to the Single Cell ATAC v1 workflow (https://www.10xgenomics.com/products/single-cell-atac), and proceeded with the MGI Easy Universal Library Conversion Kit (App-A, MGI) to convert the libraries' structure.

5. Single-cell ATAC sequencing and data processing

Libraries were sequenced on MGI2000 (MGI). Raw data were split into reads and cell barcode using custom scripts. Reads were mapped to reference genome (https://www.ncbi.nlm.nih.gov/assembly/GCA_015852565.1/) using Cell Ranger ATAC v1.2.0 with default parameters. Single-cell accessibility counts were generated after running 'cellranger-atac count'.

6. Analysis of snRNA-seq data

6.1 Data quality control and normalization for snRNA-seq data

After the matrix was exported, quality control was performed to remove low quality cells and potential doublets. Considering the range of cell library sizes (*i.e.* sequencing depth) varied among stages, we costumed different filtering thresholds (based on the number of detected genes) for each embryonic stage, which were: B (400-3,000 genes), G3 (350-3,000 genes), G4 (250-3,000 genes), G5 (150-3,000 genes), G6 (350-3,000 genes), N0, N1 and N3 stage (150-3,000 genes), L0 stage (100-3,000 genes). The threshold for filtering the adult tissue cells was 150-3,000 genes. As a result, a total of 148,875 embryonic cells and 235,170 adult tissue cells were remained for subsequent analysis. Furthermore, genes expressing in less than 3 cells were filtered out. This step was implemented using the build-in functions 'scanpy.pp.filter_cells' and 'scanpy.pp.filter_genes' from ScanPy[44].

The resulting gene-by-cell matrices were then log-transformed (with a pseudo-count added) followed by the library-size normalization, with the median of library-size as the size-factor. This was implemented using the two build-in functions ‘scanpy.pp.normalize_total’ and ‘scanpy.pp.log1p’ from ScanPy, for total-count normalization and log-transformation respectively.

6.2 Cell clustering and visualization within each stage

Selection of highly variable genes

We observed that cell groups were dominated by two types of RNA capturing primers (random and polyT) (Supplementary Figure 1). Unless special circumstances, we will refer to these two groups caused by technical noise as “polyT group” and “random group”, respectively. To avoid their effects, highly variable genes (HVGs) were first identified separately within each group, and then merged together. Cells in B stage were exceptional, which were clustered into three populations that were driven by different states of cell division, so the HVGs were selected separately in each division states.

We used the approach in Seurat v2 to identify HVGs within each group. Specifically, it calculated the average expression and dispersion (variance or mean) for each gene and placed these genes into several bins (30 bins in our case) based on (logarithmized) average expression. The normalized dispersions were then obtained by scaling with the mean and standard deviation of the dispersions within each bin. Genes with a (log-normalized) mean expressions above a certain value (costumed for each stage) and a normalized dispersion higher than 0.25 were identified as highly variable ones. This was implemented using the build-in function ‘scanpy.pp.highly_variable_genes’ from ScanPy. At last, we took those genes that were highly variable in both groups and those with top dispersion as HVGs for downstream analysis (Supplementary Table 1 and 2).

Preprocessing for dimensionality reduction

We restricted the expression matrix to those genes found highly variable. Before performing dimensionality reduction, the gene-by-cell expression matrix was centralized and scaled. This was done for each group of the same RNA capturing primer as described above. For B stage, the expression matrix was centered and scaled within each cell cycle state. Besides, cell groups in this stage were much more confounded by cell library size than other stages so we treated the logarithmized library size (*i.e.*, counts-per-cell) as the latent factor and regressed it out.

Visualization of stage-wise snRNA-seq data using UMAP

We first performed principle component analysis (PCA) and selected the number of top principal components (PCs) with highest explained variances based on the “elbow” of the scree plot of the principle components. In practice, we found the final results are quite robust to the number of selected PCs. UMAP was performed to further embed each cell from the reduced PC space onto a 2D map. UMAP is a graph-base manifold learning method that can provide a good visualization with the intrinsic structure of the original data preserved. Meanwhile it is also a computationally efficient tool for large-scale datasets. It first computes the approximate k nearest neighbors (KNNs) for each data point, building a weighted mutual-KNN graph with each node representing each data point (cell in our case), and embeds each node of the graph into the lower dimensional space. Note that the Euclidean distance metric is not scale-invariant so that it is quite sensitive to batch effects. Instead, we searched the approximate KNNs for each single cell based on cosine distance in the PC space, with the number of neighbors setting as 20. This was implemented using the build-in function ‘scanpy.tl.umap’ from ScanPy, which is a convenient wrapper of the original function ‘UMAP’ from the Python package ‘umap-learn’.

Clustering of cell populations and identification of differentially expressed genes

To cluster single cells into distinct populations, we used a graph-based clustering approach, which applies Leiden[45] community detection on the weighted KNN graph build by UMAP. The Leidenalgorithm is very similar to the Louvain[46] community detection algorithm that is wildly used for single cell clustering. This clustering method was achieved by the build-in interface ‘scanpy.tl.leiden’ from ScanPy. Cluster-pecific genes were found using “model-based analysis of single-celltranscriptomics” (MAST)[12] comparing each cluster versus the others, achieved by the function ‘FindAllMarkers’ in Seurat. After comparing the differentially expressed genes for each cluster, we manually merged those groups with no significant difference from each other.

6.3 Visualization of the embryonic cells from all the developmental stages

Construction of the single-cell graph for the merged snRNA-seq data

In order to construct a 2D atlas of single-cells from all the developmental stages that reveals both the stage order and the intrinsic lineage structures, we first built a single-cell graph, with nodes as cells and edges connecting cells with similar expression profiles. A simple and direct approach is to apply a global k-nearest-neighbor search for each cell among all the stages. However, due to the batch effects caused by technical and biological noises, the difference of the same cell type (or lineage) across stages might be greater than that of different cell types (or lineage) in the same stage, which can lead to the k-nearest neighbors of each cell more likely to be in the same stage rather than its progenitor cells or daughter cells from the adjacent stages. Therefore, a global k-nearest neighbor search without any constraints would result in a biased single-cell network. To remove these batch effects and keep the order of stages, we borrowed the idea from a study by Wagner *et.al.*[7] and designed a novel approach, stage-wise-KNN, for our scRNA-seq dataset.

For genes used to calculate the single-cell network, we merged HVGs from each stage, and kept those appeared in more than three stages. We also included some canonical marker genes collected from published literatures. Before inputting to the next step, the expression matrices from each stage were first z-scored (mean-shifted and scaled to unit variance) for each of the used genes. Let’s denote the resulting matrix $X^{(t)}$ at time point t, with each element representing the z-score of gene i in cell j.

The single-cell network was constructed by connecting only the adjacent stage-pairs, with no edges connecting cells in the same stage, except for the first stage (B stage). Specifically, for each pair of adjacent stages between time point t and t+1, we first applied PCA on the concatenated matrix to get the reduced dimensions, and searched the KNNs of each cell in stage *t*+1 from its parent stage t, based on the PC space. A binary edge would be connecting two cells if one is in the parent stage (the earlier timepoint) and is detected as a member of the KNNs of the other. As the start timepoint, the B stage was the only exception that cells in which would also connect to their KNNs in the same stage.

This single-cell network preserved the order of stages and revealed the structural relationships between different lineages. We reasoned that if two cells in the same stage had similar expression profiles, they would have a majority of KNNs in common from the previous stage, which would “attract” them to be embedded onto a nearby place in the consequent 2D map.

Parameter setting

Considering the different number of cells and the varied biological complexity in different stages, we adapted different number of PCs and nearest-neighbors for each pair of stages. We used the top 30 PCs for pairs from B to G6 stage, and the top 50 PCs for that from B6 to L0 stage. The size of neighborhood for cells from B to N1 stage was defined as *k*=10, while they were set as *k*=5 and *k*=3 for N3 and L0 stage, respectively.

Embedding the single-cell network using a force-directed graph layout algorithm

To apply a force-directed graph layout algorithm to embed the single-cell network, we utilized the build-in function ‘scanpy.tl.umap’ of ScanPy after setting the pre-built network in the slot “connectivities”. The parameter “min_dist” of this function was set as 0.1.

6.4 Construction of the developmental tree

The coarse-grained developmental tree was constructed by taking previously defined clusters in each stage as nodes and connecting nodes across timepoints by “ancestor voting”. A vote from a cell for its ancestor cell was defined as the nearest neighbor in the previous stage, which was determined when building the stage-wise-KNN based single-cell network. For each cluster in stage *t*+1, its ancestor node in stage *t* would be the cluster winning the largest number of votes from cells in this cluster.

A step of “group refinement” would come after the “ancestor voting” between each two adjacent stages. In case that some clusters in stage *t* had no descendent nodes in stage *t*+1, probably because of different cluster resolutions, they would take all the single-cells that had voted for them from the other clusters and group a new descendent node.

We also made some manually adjustment to the developmental tree using expert knowledge, for example, merging those branches with no significant gene expression difference. Finally, the constructed developmental tree was visualized using the R package ‘ggtree’ [47].

6.5 Annotation of embryonic cell populations and lineages

Considering the unbalanced sequencing depths of different stages, library-size normalization was applied to the raw expression counts of each cell with 1000 as the uniform size-factor, followed by log-transformation with a pseudo-count added. Lineage-specific genes were then found using MAST by comparing the expression profiles of cells in each lineage with those of the others, implemented by the function 'FindAllMarkers' in Seurat.

GO enrichment analysis of DEGs was performed by an R package "clusterProfiler" (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) with the customized reference database "org.bf.eg.db" (*Branchiostoma floridae*). The statistical significance was adjusted and the "pvalueCutoff" parameter was set as 0.05 for both GO and KEGG analysis. Top 20 significantly enriched GO terms were selected to show by bubble plot.

6.6 Notochord lineage analysis

The notochord lineage cells were extracted out for further lineage analysis. HVGs were selected and dimensionality reduction was performed using UMAP. MAST was used to find out the differential expressed genes for each cluster that formed the notochord lineage. Diffusion pseudo-time (DPT)[48] was applied to infer the pseudo-time of each cell.

The single-cell graph with re-clustering labels were further abstracted into a more concise and interpretable graph by partition-based graph abstraction (PAGA)[44]. In brief, each cluster was regarded as a node, and the weight of a connection between each pair of nodes was estimated based on a hypothesis test against a null model. This gave us a coarse-resolution of the partitioning and allowed a more global perspective of the intrinsic structure of the data.

7. Analysis of single-cell ATAC-seq data

7.1 Quality control and preprocessing of scATAC-seq data

Peak-cell matrics and fragment files were analyzed with Signac (v 0.2.5). For the consequent peak-cell matrix, we filtered cells that were not adequately sequenced compared to the main populations in the same stage. As done for snRNA-seq data, we customized different filtering thresholds (based on the number of detected peaks) for each embryonic stage, which were: B (600-3,500 peaks), G3 (1,500-10,000 peaks), G6 (1,000-15,000 peaks), N1 and N3 (1,500-10,000 peaks), L0 (1,500-15,000 peaks). The resulting number of cells for each stage was listed in the Supplemetary Table 2.

7.2 Independent visualization of scATAC-seq data for each stage

For independent visualization of the epigenomic data, TF-IDF transformation was applied to the peak-cell matrix of each stage and partial singular value decomposition (SVD) and UMAP were performed for dimensionality reduction. These steps were implemented using the built-in functions of Signac (RunTFIDF with scale.factor=10000, RunSVD with n=50 and reduction.key = 'LSI_') and Seurat (RunUMAP with n.neighbors = 30).

7.3 Construction of gene activity matrix from scATAC-seq data using Cicero

Before integrating the scATAC-seq data with the snRNA-seq data, we first used Cicero[35] to transform the peak-by-cell matrices of different developmental stages into gene-by-cell matrices, with the values representing gene activity scores. These gene activity matrices were then normalized to the uniform column-sums, the median of the column-sums in that stage, followed by log transformation with a pseudo-count added.

8. Integration of snRNA-seq and scATAC-seq data and label transfer

For each developmental stage with both transcriptomic and epigenetic data, we began with the normalized gene activity scores and the normalized gene expression values calculated from the previous analysis, z-scored (centered and scaled to unit variances) separately based on the HVGs selected from snRNA-seq data, which was used for clustering analysis before. The resulting z-scored matrix-pair in that stage was concatenated and used to calculating the reduced dimensions using PCA. We then corrected the coordinates in PC space by performing Harmony, a method for removing batch effects. After that, a KNN-classifier was built using the corrected coordinates and the cluster labels of snRNA-seq data as the “training samples” and that from scATAC-seq data as the “testing samples”, with the predicted cluster labels as the final transferred labels for those scATAC-seq cells.

Note that we also applied the Seurat build-in method (CCA-anchor) for integration and label-transfer. The most of transferred labels was in consistent with that from Harmony-KNN, while with less confidences, so we adopted the results of Harmony-KNN for the downstream analysis.

9. Combined analysis to identify new lineage specific genes

To identify new lineage marker genes, we carried out combined analysis using our epigenetic and transcriptomic data from each stage. Briefly, we take the intersections of the DEGs ($p < 0.01$) and differentially accessible peaks ($p < 0.001$). The candidate new lineage marker genes were selected with high expression in two successive stages.

10. Whole mount *in situ* hybridization assays

Fourteen genes showing specific expression in different germ layers in both single-nuclear RNA-seq and single-cell ATAC-seq data were chosen for this analysis. Their mRNA sequences were amplified from a mixed cDNA library using primers listed in (Supplementary Table 5), cloned to pGEM-T-Easy vector (PR-A1360, Promega), and verified by DNA sequencing. Digoxigenin (DIG)-labelled antisense riboprobes (11093088910, Roche) of all these genes were synthesized with Sp6 (P1081, Promega) or T7 (P2075, Promega) RNA polymerase. Embryos at desired stages were fixed overnight with 4% (wt/vol) PFA-MOPS-EGTA (pH 7.5) at 4°C, and then stored at -20°C in 70% ethanol (vol/vol) for use. Whole mount *in situ* hybridization (WISH) was performed as previously described (Yu and Holland 2009 Cold Spring Harb Protoc). Stained embryos were photographed using an inverted microscope (Olympus, IX71).

11. Construction of website and database

HTML and PHP were used for webpage construction, and MySQL was used for the storing and query functions. All the code was installed on a Linux Host with Apache webserver.

12. Data availability

All the snRNA-seq and scATAC-seq data have been deposited in the CNSA (https://db.cngb.org/cnsa/) of CNGBdb with accession number CNP0000891.

Declarations

Acknowledgements

We are thankful to the production team of China National Gene Bank, Shenzhen, China. This study was supported by the “Strategic Priority Research Program” of the Chinese Academy of Sciences (XDB13000000 to B.M. and S.Z.); the National Key Research and Development Program of China (2019YFA0709501 to S.Z.); the National Natural Science Foundation of China (61621003 to S.Z., 31872186 to G.L.); National Ten Thousand Talent Program for Young Top-notch Talents (to S.Z.); CAS Frontier Science Research Key Project for Top Young Scientist (QYZDB-SSW-SYS008 to S.Z.); Natural Science Foundation of Fujian Province of China (2019J01022 to G.L.); the open project of State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences (GREKF19-08 to G.L.); Open Research Fund of State Key Laboratory of Cellular Stress Biology, Xiamen University (SKLCSB2020KF005 to P.M.); and P.M. was supported by the Youth Innovation Promotion Association of Chinese Academy of Sciences.

Author Contributions

Project design and supervision: S.Z, G.L., X.X., Q.Z., D.C., and B.M.; sample collection: Z.H., C.S., Q.Z., Y.Y., Y.Z., Z.H. and G.L.; snRNA-seq library preparation: P.M., Q-J.Z., G.L., X.L., Y.H. and Z.D.; 10x Genomics scATAC-seq library preparation: Z.X. , L.L. and C.C.; snRNA-seq data analysis: X.L., Q-J. Z., G-D.W. and P.M.; scATAC-seq data analysis: D.C., Z.X., X.D., C.C., H.W., J.Q., J.Z., X.W., P.D., S.Z., Y.Z. and W.W.; data interpretation: B.M., D.C., P.M., G.L. H.X., G-D. W., L.X. and Z.H.; draft writing: D.C., P.M., X.L. and Z.X.; manuscript revision and approval: B.M., S.Z. and G.L.

Competing interests

The authors declare no competing interests.

References

1. Berna, L. & Alvarez-Valin, F. Evolutionary genomics of fast evolving tunicates. *Genome Biol Evol* 6, 1724-1738 (2014).

2. Holland, L.Z. Genomics, evolution and development of amphioxus and tunicates: The Goldilocks principle. *J Exp Zool B Mol Dev Evol* 324, 342-352 (2015).
3. Bertrand, S. & Escriva, H. Evolutionary crossroads in developmental biology: amphioxus. *Development* 138, 4819-4830 (2011).
4. Marletaz, F. *et al.* Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* 564, 64-70 (2018).
5. Briggs, J.A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* 360 (2018).
6. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* 360 (2018).
7. Wagner, D.E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* 360, 981-987 (2018).
8. Cao, C. *et al.* Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature* 571, 349-354 (2019).
9. Carvalho, J.E. *et al.* An updated staging system for cephalochordate development: one table suits them all. *BioRxiv* 112193 (2020).
10. Rosenberg, A.B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* 360, 176-182 (2018).
11. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* (2018).
12. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* 16, 278 (2015).
13. Schubert, M., Holland, L.Z., Panopoulou, G.D., Lehrach, H. & Holland, N.D. Characterization of amphioxus AmphiWnt8: insights into the evolution of patterning of the embryonic dorsoventral axis. *Evol Dev* 2, 85-92 (2000).
14. Satoh, G., Wang, Y., Zhang, P. & Satoh, N. Early development of amphioxus nervous system with special reference to segmental cell organization and putative sensory cell precursors: a study based on the expression of pan-neuronal marker gene Hu/elav. *J Exp Zool* 291, 354-364 (2001).
15. Le Petillon, Y. *et al.* Nodal/Activin Pathway is a Conserved Neural Induction Signal in Chordates. *Nat Ecol Evol* 1, 1192-1200 (2017).
16. Holland, L.Z., Schubert, M., Kozmik, Z. & Holland, N.D. AmphiPax3/7, an amphioxus paired box gene: insights into chordate myogenesis, neurogenesis, and the possible evolutionary precursor of definitive vertebrate neural crest. *Evol Dev* 1, 153-165 (1999).
17. Kozmik, Z. *et al.* Pax-Six-Eya-Dach network during amphioxus development: conservation in vitro but context specificity in vivo. *Dev Biol* 306, 143-159 (2007).
18. Pascual-Anaya, J. *et al.* Broken colinearity of the amphioxus Hox cluster. *Evodevo* 3, 28 (2012).

19. Zhong, Y., Herrera-Ubeda, C., Garcia-Fernandez, J., Li, G. & Holland, P.W.H. Mutation of amphioxus Pdx and Cdx demonstrates conserved roles for ParaHox genes in gut, anus and tail patterning. *BMC Biol* 18, 68 (2020).
20. Venkatesh, T.V., Holland, N.D., Holland, L.Z., Su, M.T. & Bodmer, R. Sequence and developmental expression of amphioxus AmphiNk2-1: insights into the evolutionary origin of the vertebrate thyroid gland and forebrain. *Dev Genes Evol* 209, 254-259 (1999).
21. Butts, T., Holland, P.W. & Ferrier, D.E. Ancient homeobox gene loss and the evolution of chordate brain and pharynx development: deductions from amphioxus gene expression. *Proc Biol Sci* 277, 3381-3389 (2010).
22. Lecroisey, C., Le Petillon, Y., Escriva, H., Lammert, E. & Laudet, V. Identification, evolution and expression of an insulin-like peptide in the cephalochordate Branchiostoma lanceolatum. *PLoS One* 10, e0119461 (2015).
23. Liu, X., Li, G. & Wang, Y.Q. The role of the Pax1/9 gene in the early development of amphioxus pharyngeal gill slits. *J Exp Zool B Mol Dev Evol* 324, 30-40 (2015).
24. Candiani, S., Castagnola, P., Oliveri, D. & Pestarino, M. Cloning and developmental expression of AmphiBrn1/2/4, a POU III gene in amphioxus. *Mech Dev* 116, 231-234 (2002).
25. Castro, L.F., Rasmussen, S.L., Holland, P.W., Holland, N.D. & Holland, L.Z. A Gbx homeobox gene in amphioxus: insights into ancestry of the ANTP class and evolution of the midbrain/hindbrain boundary. *Dev Biol* 295, 40-51 (2006).
26. Albuixech-Crespo, B. *et al.* Molecular regionalization of the developing amphioxus neural tube challenges major partitions of the vertebrate brain. *PLoS Biol* 15, e2001573 (2017).
27. Ren, Q. *et al.* Step-wise evolution of neural patterning by Hedgehog signalling in chordates. *Nat Ecol Evol* 4, 1247-1255 (2020).
28. Sharman, A.C., Shimeld, S.M. & Holland, P.W. An amphioxus Msx gene expressed predominantly in the dorsal neural tube. *Dev Genes Evol* 209, 260-263 (1999).
29. Schubert, M., Holland, L.Z. & Holland, N.D. Characterization of two amphioxus Wnt genes (AmphiWnt4 and AmphiWnt7b) with early expression in the developing central nervous system. *Dev Dyn* 217, 205-215 (2000).
30. Schubert, M. *et al.* Retinoic acid signaling acts via Hox1 to establish the posterior limit of the pharynx in the chordate amphioxus. *Development* 132, 61-73 (2005).
31. Short, S., Kozmik, Z. & Holland, L.Z. The function and developmental expression of alternatively spliced isoforms of amphioxus and Xenopus laevis Pax2/5/8 genes: revealing divergence at the invertebrate to vertebrate transition. *J Exp Zool B Mol Dev Evol* 318, 555-571 (2012).
32. Candiani, S., Holland, N.D., Oliveri, D., Parodi, M. & Pestarino, M. Expression of the amphioxus Pit-1 gene (AmphiPOU1F1/Pit-1) exclusively in the developing preoral organ, a putative homolog of the vertebrate adenohypophysis. *Brain Res Bull* 75, 324-330 (2008).
33. Kaltenbach, S.L., Yu, J.K. & Holland, N.D. The origin and migration of the earliest-developing sensory neurons in the peripheral nervous system of amphioxus. *Evol Dev* 11, 142-151 (2009).

34. Candiani, S., Moronti, L., Ramoino, P., Schubert, M. & Pestarino, M. A neurochemical map of the developing amphioxus nervous system. *BMC Neurosci* 13, 59 (2012).

35. Pliner, H.A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* 71, 858-871 e858 (2018).

36. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 16, 1289-1296 (2019).

37. Jukam, D., Shariati, S.A.M. & Skotheim, J.M. Zygotic Genome Activation in Vertebrates. *Dev Cell* 42, 316-332 (2017).

38. Aanes, H. *et al.* Differential transcript isoform usage pre- and post-zygotic genome activation in zebrafish. *BMC Genomics* 14, 331 (2013).

39. Despic, V. *et al.* Dynamic RNA-protein interactions underlie the zebrafish maternal-to-zygotic transition. *Genome Res* 27, 1184-1194 (2017).

40. Mazet, F., Masood, S., Luke, G.N., Holland, N.D. & Shimeld, S.M. Expression of AmphiCoe, an amphioxus COE/EBF gene, in the developing central nervous system and epidermal sensory neurons. *Genesis* 38, 58-65 (2004).

41. Li, G., Yang, X., Shu, Z., Chen, X. & Wang, Y. Consecutive spawnings of Chinese amphioxus, Branchiostoma belcheri, in captivity. *PLoS One* 7, e50838 (2012).

42. Li, G., Shu, Z. & Wang, Y. Year-round reproduction and induced spawning of Chinese amphioxus, Branchiostoma belcheri, in laboratory. *PLoS One* 8, e75461 (2013).

43. Liu, X., Li, G., Feng, J., Yang, X. & Wang, Y.Q. An efficient microinjection method for unfertilized eggs of Asian amphioxus Branchiostoma belcheri. *Dev Genes Evol* 223, 269-278 (2013).

44. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19, 15 (2018).

45. Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 9, 5233 (2019).

46. Blondel, V.D., Guillaume, J.L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics-Theory and Experiment* (2008).

47. Yu, G.C., Smith, D.K., Zhu, H.C., Guan, Y. & Lam, T.T.Y. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8, 28-36 (2017).

48. Haghverdi, L., Buettner, F. & Theis, F.J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 31, 2989-2998 (2015).

Figures

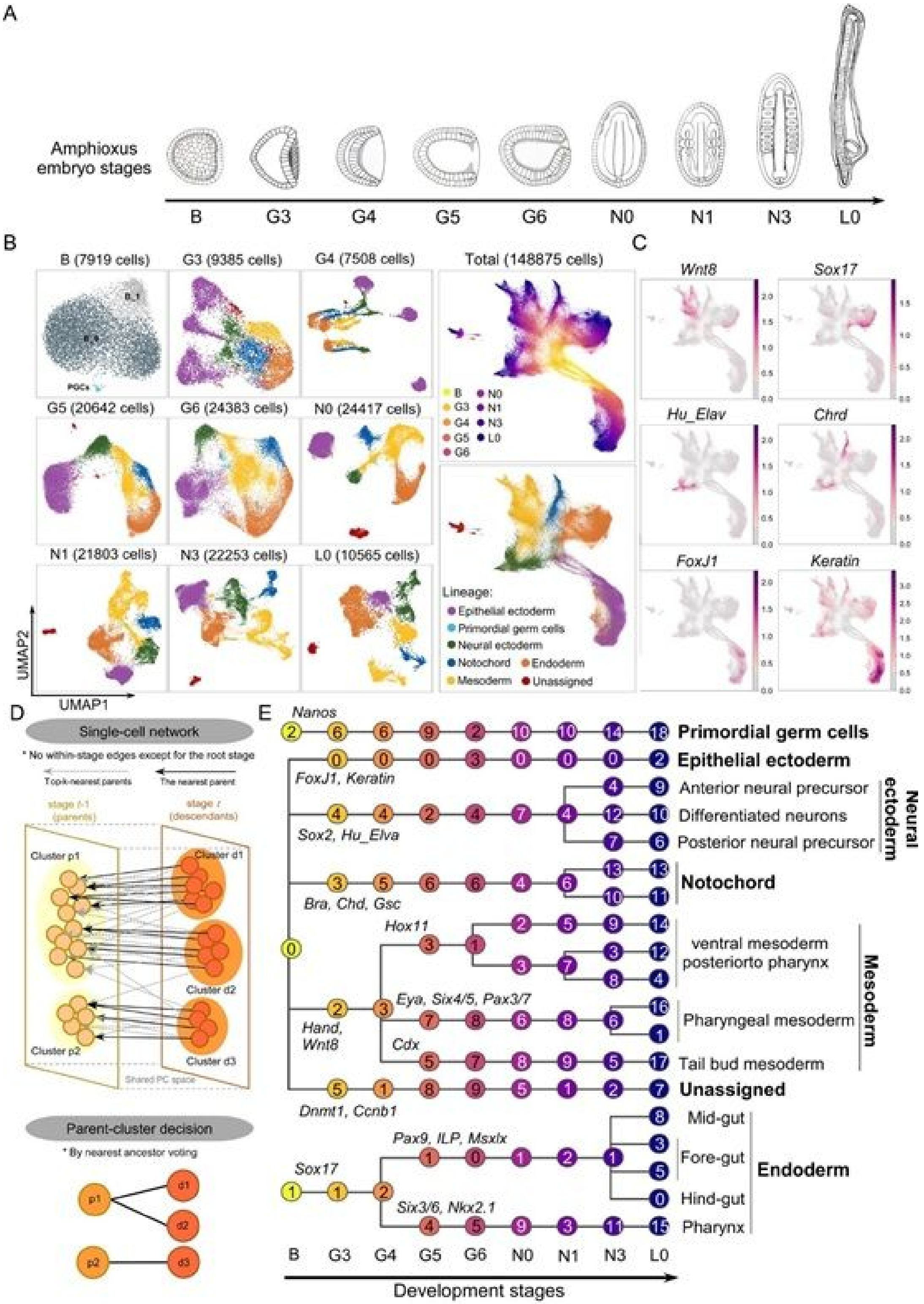

Figure 1

Single cell transcriptional atlas and cell lineage specification during amphioxus embryo development. (A) Diagram of amphioxus embryos from blastula to early larva used for snRNA-Seq and scATAC-seq analysis in this study. (B) UMAP plots for each developmental stage or the total 148,875 single cell transcriptomes, constructed in dimensionality-reduced principal component analysis subspace defined by highly covariable genes. Cells are colored by germ layer identities inferred from expressed marker genes or their developmental stage origins, respectively. (C) Plots of the expression of representative lineage-specific genes in different cell populations. (D) Schematic of the mapping algorithm used to make similarity connections between clusters across developmental stages (Materials and Methods). (E) Virtual cell lineage tree were constructed using transcriptome profiles from sequential developmental stages. Generated using the mapping algorithm in (D).

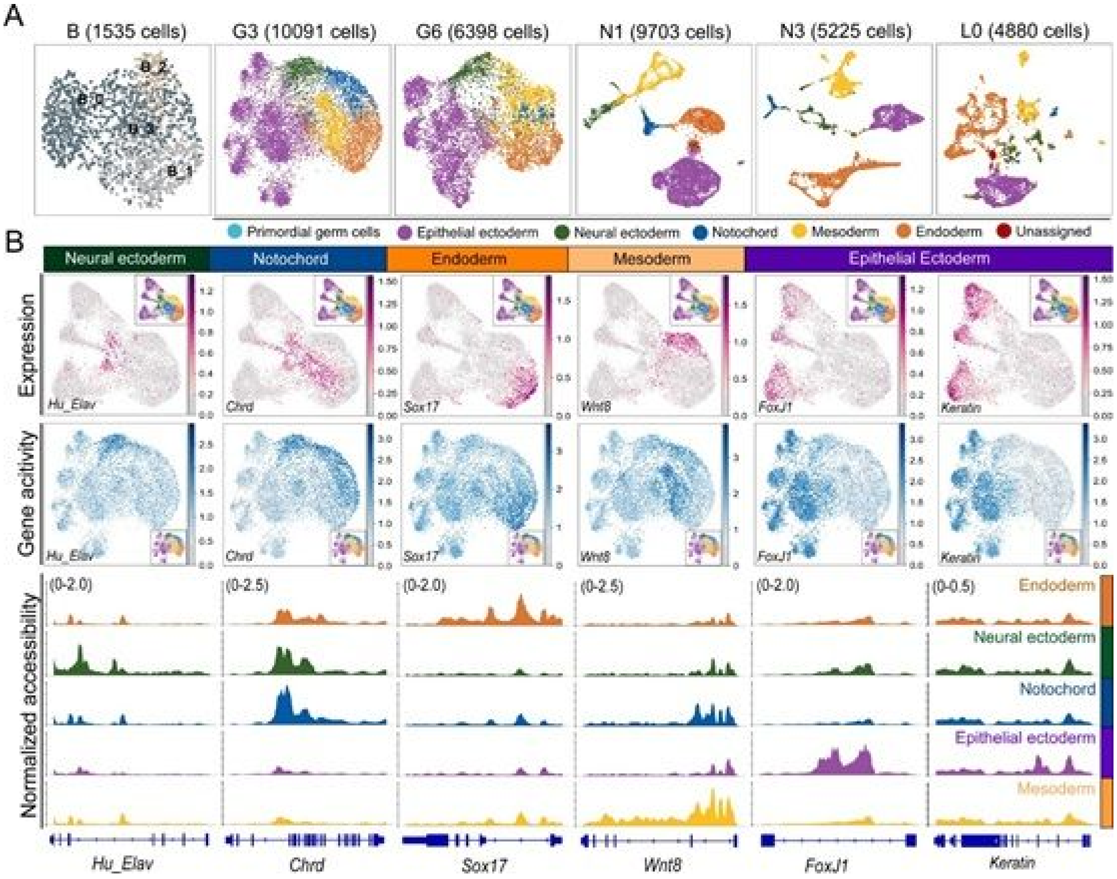

Figure 2

Overview of combined analysis of single nuclear RNA-seq and ATAC-seq data for amphioxus embryos at the indicated developmental stages. (A) UMAP plots of single cell epigenome from scATAC-seq data for amphioxus embryos at the indicated developmental stages, which is colored by the lineage labels transferred from the corresponding snRNA-seq data. (B) The expressions of the lineage-specific markers from the snRNA-seq analysis correlated well with their gene activities deduced from the scATAC-seq data. The expressions (snRNA-seq), gene activities (predicted by Cicero) and chromatin accessibilities (visualized by IGV) for the selected genes were shown.

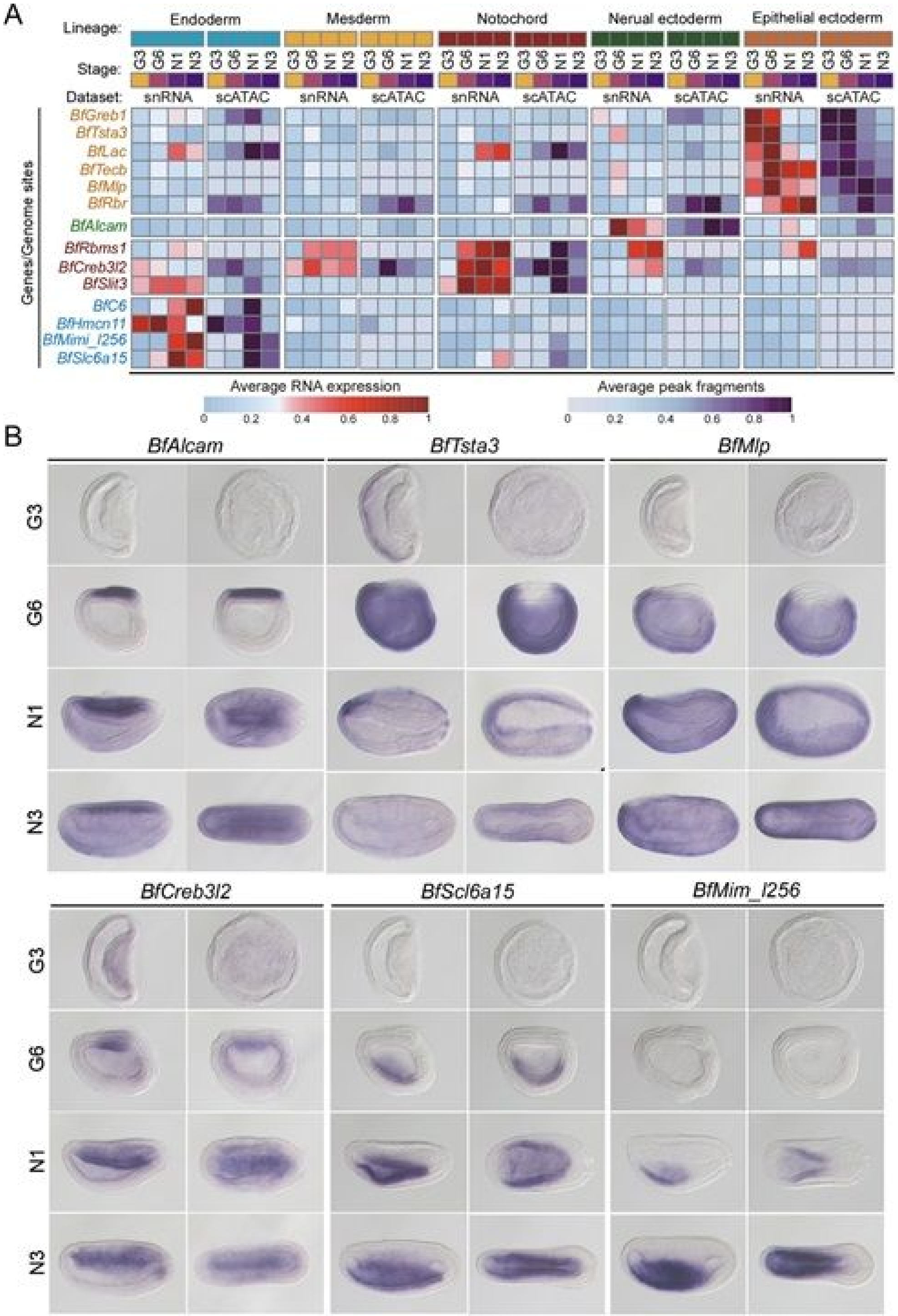

Figure 3

Identification of new lineage-specific genes through combined analysis of scRNA-seq and scATAC-seq data and validation by whole mount in situ hybridization. (A) Hierarchical clustering of expression heatmaps showing differentially expressed new marker genes for each cell lineage. Source data are provided in Supplementary Table 5. (B) Whole mount in situ hybridization results showing the expression patterns of the selected genes.

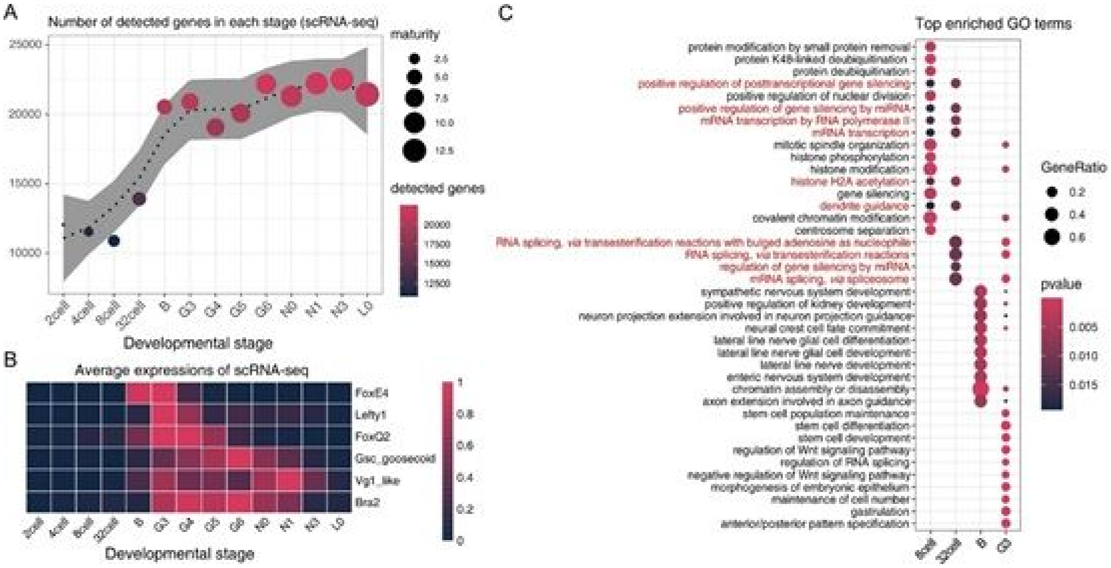

Figure 4

Zygotic genome activation occurs at the period from 32-cell to blastula stage during amphioxus embryo development. (A) Number of detected genes from the scRNA-seq data for each amphioxus developmental stage. (B) Expression heatmaps of known ZGA genes during amphioxus embryo development. Values are averaged and max-normalized for each gene, respectively. (C) Enriched GO terms of the activated genes for each stage compared with its parent stage. Source data are provided in Supplementary Table 6.

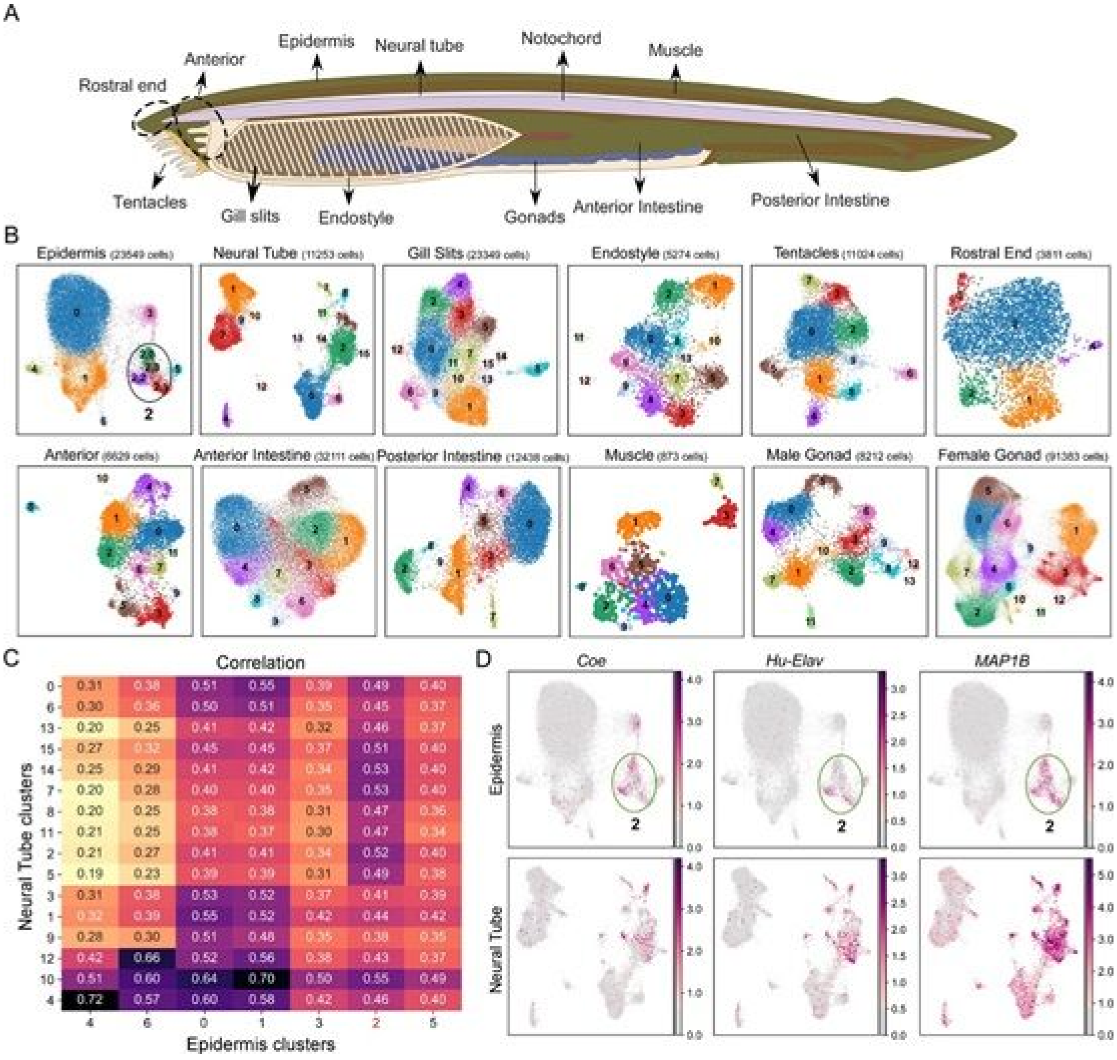

Figure 5

Overview of the single-nucleus RNA sequencing analysis for adult amphioxus tissues. (A) Schematic overview of the adult amphioxus tissues used in our snRNA-Seq analysis. (B) UMAP plots of the snRNA-seq data from 12 adult amphioxus tissues, colored by clusters. (C) Spearman correlation between each cluster in epidermis and neural tube, based on the expressions of top 20 DEGs. (D) UMAP plots showing the expression levels of three neural markers in epidermis and neural tube cell clusters.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable4.xlsx
- SupplementaryTable5.xlsx

- SupplementaryTable6.xlsx
- SupplementaryTable7.xlsx
- Maetal.SupplInformation.docx