Computational and Structural Biotechnology Journal 20 (2022) 3653–3666

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Method article

Computational modeling and analysis of the morphogenetic domain signaling networks regulating *C. elegans* embryogenesis

Ben Niu [a,1], Thao Nguyen Bach [a,1], Xingyu Chen [a], Khyati Raghunath Chandratre [a], John Isaac Murray [c], Zhongying Zhao [b], Michael Zhang [a,*]

[a] *Center for Systems Biology, The University of Texas at Dallas, 75080, USA*
[b] *Department of Biology, Hong Kong Baptist University, Hong Kong*
[c] *Department of Genetics, The University of Pennsylvania, USA*

ARTICLE INFO

Article history:
Received 23 November 2021
Received in revised form 29 May 2022
Accepted 30 May 2022
Available online 8 June 2022

Keywords:
Computational image analysis
Morphogenetic domain cell signaling network
Single cell gene expression modeling
Collective cell behavior
Machine learning

ABSTRACT

Caenorhabditis elegans, often referred to as the 'roundworm', provides a powerful model for studying cell autonomous and cell–cell interactions through the direct observation of embryonic development *in vivo*. By leveraging the precisely mapped cell lineage at single cell resolution, we are able to study at a systems level how early embryonic cells communicate across morphogenetic domains for the coordinated processes of gene expressions and collective cellular behaviors that regulate tissue morphogenesis. In this study, we developed a computational framework for the exploration of the morphogenetic domain cell signaling networks that may regulate *C. elegans* gastrulation and embryonic organogenesis. We demonstrated its utility by producing the following results, i) established a virtual reference model of developing *C. elegans* embryos through the spatiotemporal alignment of individual embryo cell nuclear imaging samples; ii) integrated the single cell spatiotemporal gene expression profile with the established virtual embryo model by data pooling; iii) trained a Machine Learning model (Random Forest Regression), which predicts accurately the spatial positions of the cells given their gene expression profiles for a given developmental time (*e.g.* total cell number of the embryo); iv) enabled virtual 4-dimensional tomographic graphical modeling of single cell data; v) inferred the biology signaling pathways that act in each of morphogenetic domains by meta-data analysis. It is intriguing that the morphogenetic domain cell signaling network seems to involve some crosstalk of multiple biology signaling pathways during the formation of tissue boundary pattern. Lastly, we developed the Software tool 'Embryo aligner version 1.0' and provided it as an Open Source program to the research community for virtual embryo modeling, and phenotype perturbation analyses (https://github.com/csniuben/embryo_aligner/wiki and https://bioinfo89.github.io/C.elegansEmbryonicOrganogenesisweb/).

1. Introduction

Embryonic development is precisely controlled through the communication of cell populations that define morphogenetic domains. Turning's kinetics equations that model the signaling interactions of morphogen gradients at tissue scale predict accurately the processes of pattern formation and morphogenesis in embryogenesis [1–4]. The quantitative and precise analysis of the cell signaling networks across morphogenetic domains is fundamental to the understanding of the collective cellular behaviors underlying embryogenesis and human diseases. Cell signaling networks can be used to explain development in specific cell types however a whole-organism level understanding is lacking.

Caenorhabditis elegans is well-suited to study the signaling networks, owing to its simplicity for observations and measurements. *C. elegans* is the first multicellular organism with the whole genome sequenced [5]. Intriguingly, the worm's genes turned out to be quite similar to that of the human given their common ancestor lived in the pre-Cambrian era (500–600 million years ago). Approximately 60% of the *Homo sapiens* proteins in SwissProt matched with the *C. elegans* proteins in the Wormpep database [6]. Among the available *C. elegans* protein sequences, at least 83% of *C. elegans* proteome has human homologous genes [7]. It is also the first

* Corresponding author.
E-mail addresses: ben.niu@utdallas.edu (B. Niu), michael.zhang@utdallas.edu (M. Zhang).
[1] Equal contribution; joint first authors.

https://doi.org/10.1016/j.csbj.2022.05.058

model organism with the cell lineage identity traced and fully annotated at single cell resolution in the developing embryos [8–10]. In addition, it is the first organism with the neural connectome fully mapped [11]. Studies with *C. elegans* contributed significantly to the research fields of cell biology, gene regulation and imaging [12–14].

Experimental biologists have been investigating the cell signaling pathways that regulate *C. elegans* development for decades. The protein kinases of the *RTK-Ras-ERK*, Eph receptor, Neurotransmitter, Hedgehog, Insulin like growth factor, *TGF-β* and Wnt signaling pathways play critical roles in regulating mammalian embryogenesis, development, aging and diseases and all of their orthologous counterparts have been identified in *C. elegans* [15]. Recently, Packer et al. and Tintori et al. studied *C. elegans* embryo development by performing single cell RNA-sequencing [16,17] experiments. Ma et al. profiled the expression atlas of 266 transcription factors by conducting protein-fusion fluorescent reporter assays [18]. High quality transcriptome and protein expression data were generated at single cell resolution from these experiments, which provides a great resource for understanding how the gene expression pattern changed at the different stages and locations of *C. elegans* development. In light of Ma et al.'s work that uses lineage-based method, we developed the correlative image-based method to interrogate quantitatively and precisely the relationship between embryonic geometry and gene expression regulation.

In addition to the studies of molecular cues in cell signaling, it is an emerging research area to understand how the biomechanical stimuli have the influence on cell fate decision and collective behaviors during *C. elegans* embryogenesis [19]. Mechanistically, it remains poorly understood, for instance, on how the mechanical tension that originates from muscles could stimulate the remodeling of adherent junction required for embryo elongation. Developmental biologists figured out that the next challenge will be to measure the properties of mechanical forces (magnitude and direction) at the cellular scale, and to understand how different proteins as well as feedback mechanisms regulate those properties. Biophysical methods and computer modeling will also be instrumental to clarify how the cooperation of signaling and mechanical forces may shape the organism [20]. To meet this challenge, it is of fundamental importance for computational biologists to establish the *in silico* model of the developing *C. elegans* embryo characterizing the geometric, the gene expression and the behavioral features of the worm, not only at the individual cell [21–23] but also collectively at the tissue sub-population levels.

In the area of Computational Biology, by far, most of the studies with *C. elegans* were conducted at single cell level or at the level of whole adult animal [15]. Relatively few reports in Computational modeling have been made at the level of morphogenetic domains on early embryogenesis. Still, in *C. elegans* development, it remains not quite clear at system level on how the signaling networks communicate across the morphogenetic domains to coordinate the behaviors of composite cell populations in gastrulation and early organogenesis.

Different from the analysis of skeletal or hair follicle development, which involves the study of only a small number of cell populations spatially arranged in a relatively simple and linear fashion in the mouse growth plates and embryo skins [2,3], the analysis on *C. elegans* embryogenesis requires the precise measurement, simultaneous tracing, and mathematical modeling of at least eight cell sub-populations, which can hardly be achieved without using powerful computing systems.

At single cell level, outstanding studies on cellular behaviors have been carried out through the joint efforts of experimental and computational scientists. A number of powerful computer programs for automated cell localization, tracing, lineage annotation and gene expression quantification, have been created [8,24–27]. They were applied to the quantitative analysis of *C. elegans* embryo morphogenesis and cell behaviors without primarily focusing on the spatial temporal gene expression profiles [27–29]. To complement to these previous studies, we proceeded further to make full use of the computationally integrated spatiotemporal gene expression information for interrogating the relationships of gene regulation, collective cellular behaviors and dynamic morphogenetic pattern formation. The proposed method can be utilized to identify at system level the biology signaling pathways and the morphogenetic domain signaling networks that regulate tissue morphogenesis. The analysis on the collective cell behaviors during the process of coordinated migration, the self-assembly of the cells into polarized tissues, and the establishment of tissue continuity through long range cell communication for example requires the computer algorithm be capable of analyzing the data not only at single cell level, but at the level of cell sub-populations and morphogenetic domains defined by differential gene expression patterns.

At sub-cellular level, correlative image analysis of the Confocal Fluorescence and the Transmission Electron Microscopic (TEM) sample slices has long been utilized by biologists to investigate the relationships of differential gene expression patterns, cell anatomies and altered cellular behaviors. At the morphogenetic domain level that involves the signaling interactions of multiple cell populations, however, an effective computational method is to be developed to further improve our understanding about the relationships of gene regulation, long range morphogenetic domain signaling and the tissue dynamics of *C. elegans* embryogenesis.

Recently, the integration of single cell expression and image data by Machine Learning techniques has made it technically possible to study the spatiotemporal relationship of gene regulation and tissue histology or organ morphogenesis that involve the interactions of multiple cell populations [21–23]. In a number of real-world pattern recognition and image understanding tasks, Machine Learning algorithms have been proved extremely useful [30,31]. Once the training data with positive/negative labels become available, the models trained outperformed human experts not only in efficiency but accuracy on very large scale image databases. In our study, we performed 4-dimensional tomographic section analysis of the virtual embryo model. A large set of computational images that combine the gene expression and the morphological patterns in the developing virtual embryo were generated in high spatial and temporal resolution for hundreds of genes. Owning to the large size and the dynamic nature of such dataset, it is a great challenge for human experts to conduct analysis of morphogenetic domains to obtain statistically significant results without using Machine Learning methods. The development of a computational pipeline for the domain level embryonic pattern analysis and bioinformatics pathway mapping is thus highly desirable.

To study the spatiotemporal gene expression regulation in *C. elegans* embryo, from fertilization to bean stage (~350 cells) [15], we developed Machine learning method to predict cell positions from gene expression profiles. During early embryogenesis, cell fate decision depends on the specific geometric setting of the neighboring cells and the genes that were differentially expressed. Analyzing the relationship between gene expression and cell position helps to illuminate at population level how the cells adjust their molecular profiles to remodel the niche micro-environment for patterning and organogenesis. In feedback, it is also important to understand how the remodeling of the niche geometry may contribute to further gene expression changes and altered cellular fates.

To better understand the processes of gastrulation and early organogenesis in the context of gene expression regulation, Packer

et al. [17] and Tintori et al. [16] have performed extensive and comprehensive amount of work by conducting single cell RNA-sequencing experiments, computational analysis and modeling with the *C. elegans* embryos across developmental stages. Cell lineage identities were computationally identified with the mRNA expression profiles of the signature genes, which has generated for the research community, and for the first time, a complete view about the full spectrum of the expression levels of many thousands of the genes in the developing embryos at single cell resolution. This has greatly increased understanding about the molecular pathways that are enriched for the different cell populations at the different stages of development. Recently, the integration of the 10X genomics droplet single cell RNA-sequencing technology with the spatial microarray [32], presented another great effort toward better understanding the relationship between gene expression regulation, cell spatial arrangement and tissue pattern formation. The current high-throughput methods by single cell RNA-sequencing are extremely powerful but not without a limitation in the sense of technical feasibility, since the spatial and the temporal information that is critical to our understanding about the cellular life and function cannot always be fully and simultaneously captured in vivo in a high-throughput manner. By leveraging the power of computer algorithms for data pooling, normalization and optimal spatiotemporal alignment, however, it is feasible to integrate and perform meta-data analysis on the high-quality datasets generated across experiments for studying *C. elegans* embryogenesis, in the context of spatiotemporal gene expression regulation.

Recently, in an initial effort to reconstruct cell position information from the gene expression pattern, Karaiskos et al. [33] developed the Computer program, DVEX-'Drosophila-Virtual-Expression-eXplorer', that can be utilized to predict cell positions from their gene expression profiles in Drosophila embryo. The authors established the virtual drosophila embryo model and performed digital in situ hybridization, 'vISH', at the embryonic stage of ~6000-cell for over 3100 genes. In doing so, they analyzed the distinct combinatorial gene expression patterns that regulate complex tissue morphogenesis in Drosophila development.

Different from the previous studies, our primary goal is to understand at the morphogenetic domain level the cell signaling networks that regulate, i) the collective cellular behaviors, such as the coordinated cell proliferations, migrations, and the self-assembly of cells into polarized tissues; ii) the distinct combinatorial gene expression patterns that define morphogenetic domains; iii) the spatiotemporal signaling dynamics that control complex tissue boundary pattern formation. Recently, Packer et al. [17] has profiled the transcriptomes of 86,024 single embryonic cells. They identified 502 terminal and preterminal cell types, mapping most single-cell transcriptomes to their exact position in *C. elegans*' invariant lineage. They found that multilineage priming contributes to the differentiation of sister cells at dozens of lineage branches and most distinct lineages that produce the same anatomical cell type converge to a homogenous transcriptomic state. By combining cryo-sectioning with and mRNA sequencing, Ebbing and Vertesy et al. [34] have generated gene expression maps of young adult *C. elegans* males and hermaphrodites, which provide a powerful resource for identifying tissue and sex-specific genes. As a first step, we asked whether cell position can be predicted solely on the basis of gene expression at a given developmental stage. In light of the previous work by Packer et al.[17], Ebbing and Vertesy et al. [34], Tintori et al. [16] and Karaiskos et al.[33], we computationally transformed the spatiotemporal gene expression profile of developing *C. elegans* embryo with the data from the EPIC database of gene reporter assays (https://waterston.gs.washington.edu/epic/) to study the quantitative relationship between gene expression and cell position. The EPIC database includes the cell lineage expression tree information at single cell resolution for over 250 transcription factor reporter gene assays during *C. elegans* embryogenesis. It provides with spatiotemporal resolution the *in vivo* information about the expression levels of the transcription factor encoding genes that play critical roles in regulating *C. elegans* gastrulation and early organogenesis. The expression patterns in EPIC were generated by live imaging and cell tracking of fluorescent expression reporters, so the physical position and expression level of each cell and at each point in time are known. We applied the transformed data to train a Machine Learning model that predicted accurately the cell positions from the gene expression profiles without the spatial labels in the cross-validation test.

Many signaling pathways require direct cell–cell contact. To identify cell contacts in the *C. elegans* embryo, we established a virtual reference embryo model from the 4-cell to 345-cell stage (Fig. 1A). We identified the direct cell–cell interaction relationships by generating Voronoi diagrams with the cell nuclear positions in 2- and 3-dimensional space. Cells in the Voronoi diagram were represented as polygon shapes and color coded for visual distinction (Fig. 1B-C, Supplementary Fig. 3, movie 1). The workflow of the analysis involves two major steps (Fig. 1D), i) the transformation of the spatiotemporal gene expression profile raw data; and ii) the correlative image and regression analysis of gene expression and embryo anatomical patterns. In the first step we performed gene expression data compiling, data pooling, and data normalization, as indicated in the flowchart (Fig. 1D). In the second step, we generated the anatomical profile of the developing virtual embryo through the spatiotemporal alignments of all the embryo samples (Supplementary Fig. 2). The gene expression and the anatomical profiles of the embryo were then integrated for training Machine Learning model and correlative image analysis.

We made predictions on cell positions by training a Random Forest Regression model with the transformed spatiotemporal gene expression data. The gene expression data in each embryo sample were log2-scale transformed, then normalized by using the Z-score method [35] (Fig. 1E). The heatmaps showed the estimated kernel density functions of the gene expression levels in all the embryo samples before and after the Z-score normalization. The embryo samples in the group one, including 138 samples, showed relatively stronger and broader gene expression patterns than those of 115 samples in the group two that displayed relatively weaker signal intensity and narrower spatial range of expression. (Supplementary Fig. 2).

Given the gene expression profiles of the cells in the virtual embryo model at each developmental time point, the Random Forest Regression model predicted accurately the cell positions along the Ventral-Dorsal (Pearson correlation score = 0.875), the Left-Right (Pearson correlation score = 0.878), and the Anterior-Posterior body axis (Pearson correlation score = 0.900) by 2-fold cross validations at the 345-cell stage of the embryo. On the negative control sample with randomly generated mock gene expression data, the correlation score is only 0.13. (Fig. 1F, Supplementary movies 2–4).

To better understand how gene expression regulates *C. elegans* embryogenesis at single cell and tissue scale, we developed a computational framework to study the dynamic relationships of gene expression regulation, morphogenetic domain cell signaling network, and the collective cell behaviors in the developmental process of *C. elegans* gastrulation and early organogenesis.[1].

2. Methods and datasets

2.1. Graphics data and gene expression information

The integration of gene expression and cell spatiotemporal data was achieved by leveraging the cell identity information annotated

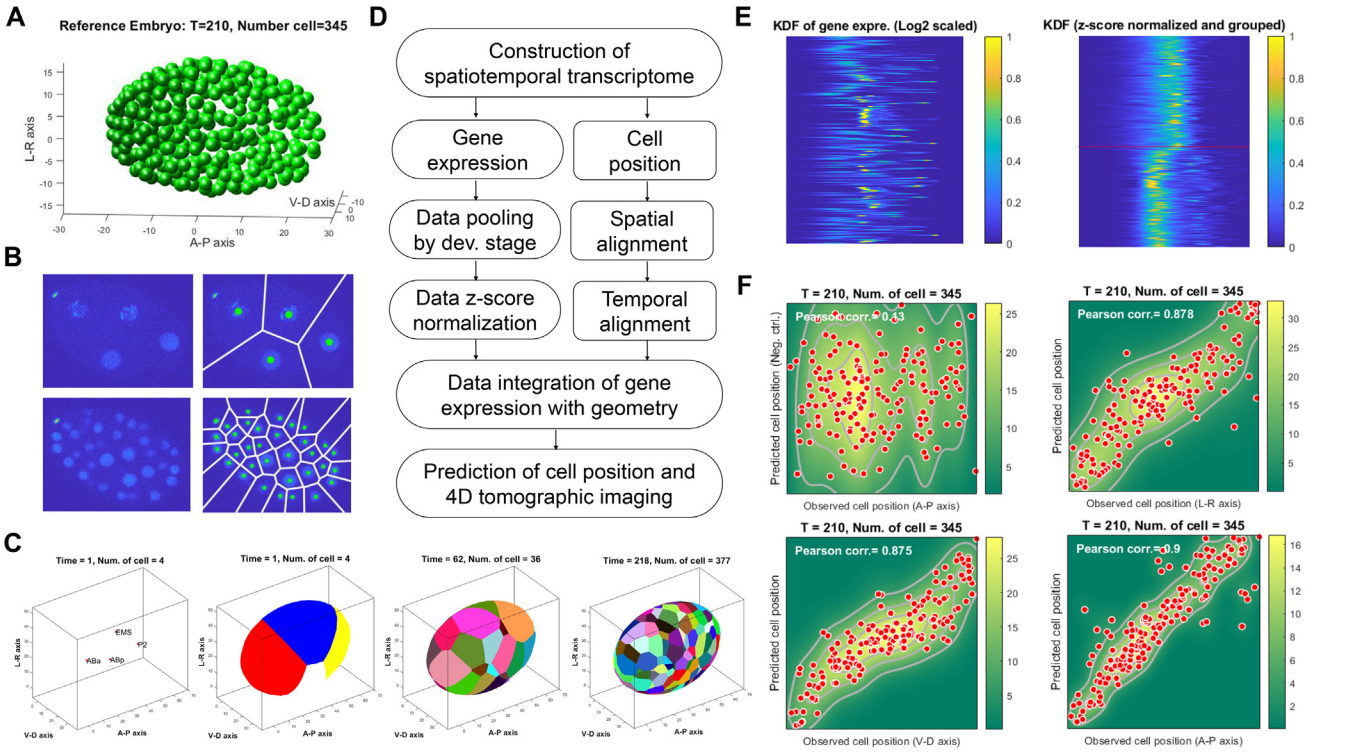


Fig. 1. Construction of spatiotemporal gene expression profile and prediction of cell position. (A) A virtual reference embryo model of C. elegans was constructed at the developmental stage of 4 to 345 cells. (B-C) The direct cell–cell interaction relationships were identified by generating Voronoi diagrams with the cell coordinates in 2- and 3-dimensional space, at the early 4-, 36- and the later 345 cell stages. (D) The workflow of the analysis. Involves two major steps, i) the construction of the spatiotemporal gene expression profile; and ii) correlative image analysis by spatiotemporal alignment of embryo cells. (E) The gene expression data in each embryo sample were log2-scaled, then normalized by using the Z-score method. (F) Given the gene expression profiles of the cells, the Random Forest Regression model was trained which predicted accurately the cell positions along the Ventral-Dorsal (Pearson correlation score = 0.874), the Left-Right / Medial-Lateral (Pearson correlation score = 0.877), and the Anterior-Posterior body axis (Pearson correlation score = 0.9) by 2-fold cross validation respectively.

by man experts [36]. Similar to somite number, widely used for staging the intrinsic developmental time of vertebrate embryos, the cell number of *C. elegans* embryos has also been widely adopted as a measure of development time point. The known corresponding relationship in cell identity, position and developmental stage allows the pooling and the normalization step to be carried out across embryos to generate the reference embryo model.

Data pooling. The cell position and gene expression information were obtained from the EPIC database [36]. The 4-dimensional graphics data were obtained from Dr. Du Zhuo's lab [37]. For each of the 254 reporter experiments on 120 genes, we downloaded the cell coordinates and the reporter gene expression data files from the EPIC Web database. For each embryo sample, the cellular age was converted from the format of video recording time to the total count of embryo cell number. The reporter gene expression data of all the embryos were then pooled into a file containing 267,418 rows and 254 columns, with each row representing the expression profile of a cell measured at one time point, and each column representing the expression levels of a reporter gene measured over all the cell types at different time points determined by cell numbers.

Data normalization. The gene expression data obtained from EPIC database were log2 scale transformed, then normalized within each embryo sample by employing the Z-score method [35]. Two subgroups of the samples were identified by K-means clustering [38,39] of the estimated kernel density function (KDF) [40] of the reporter gene expression levels (Supplementary Fig. 1). The embryo samples in group one, including 138 samples, showed relatively stronger and broader gene expression patterns than those of the 115 samples in group two.

For a total number of 254 embryo samples, the data pooling and normalization steps were performed on the System Biology Computer Cluster with 16-CPUs (Intel® Xeon® processor E5-2600, 8 cores per socket) and 95 GB memory. The data pooling computer code was implemented with Python 3.6.8, Matlab 2019b and SQLite 3.7.17, and run in the distributed computing environment with Slurm manager software version 18.08.9. It took approximately 5 h to finish processing all the samples on the Computer Cluster.

2.2. Establishment of the developing virtual reference embryo model

Four-dimensional alignment of embryonic cells. The cell position alignment is achieved by leveraging the known cell identity information of the embryo pairs and solving a mathematical optimization problem with the Procrustes method [41]. The control embryo sample in the database was utilized as the template sample. All the other embryo samples were aligned to it for each computer image time frame (Supplementary Fig. 3, movie 5). For each set of pair-wise alignment, the cell coordinates of the two embryo samples were first centralized and the distances between the corresponding cell types were measured to calculate the Mean Squared Error (MSE: green bars). The MSE was minimized by optimally aligning the two samples through the rotation, the scaling and the translation transformations identified with the Procrustes method [41]. In addition to MSE, we also measured the similarity between the two embryo samples by checking the total number of cell types that they have in common. We calculated the ratio of the number of common cell types over the number of all the cell types in the pair of samples and used it as the second criteria of

optimality. After performing alignments over all the samples in the database, we established the developing virtual reference embryo model by taking the average of all the aligned embryo samples at each developmental stage.

Identification of directly interacting cells. We generated the Voronoi diagram in 3-dimensional space with the cell nucleus coordinates. For each cell in the Voronoi diagram, we identified its nearest neighboring cells in the Voronoi cell (Supplementary Fig. 3, movies 8–9). They directly interact with each other and have equal distance to the common Voronoi cell boundary [3]. Analysis at single cell level can be achieved through the integration of gene expression profiles with cell spatiotemporal information in 3-dimensional Voronoi diagram.

2.3. Prediction of cell locations with spatiotemporal gene expression profile

For the virtual reference embryo at each developmental stage, we trained the Random Forest Regression model [42]. We took the gene expression profiles of the cells as the training input variable, the cell positions as the output variable. We generated random mock gene expression data as the negative control for model training and testing. Given the gene expression profile of a cell, the model predicts its position along the three body axes of Anterior-Posterior (A-P), Left-Right (L-R) and Ventral-Dorsal (V-D). We performed 2-fold cross-validation by using 50% of the samples for model training and the remainder 50% for testing. The Pearson's correlation scores were calculated between the predicted and observed cell positions for performance evaluation. We performed cross-validation with the lineage-informed and the randomly shuffled sampling strategy respectively. In the lineage-informed cross-validation, the pairs of daughter cells were selected for training and testing, while in the randomly shuffling mode, the samples were randomly selected for training and testing without utilizing the information about the ancestry relationship. The Random Forest Regression Tree models trained in both ways showed the similar level of performance in prediction accuracy. When the lineage information is further provided and utilized for training, the model demonstrated improved results in making accurate and stable predictions. It should also be noted that in the randomly shuffled training and testing mode, the model performed consistently and showed more accurate result of prediction when the embryo developed over ~100 cell stage as the Random Forest Tree Regression model, similar to other Machine Learning algorithms for Statistical Learning, requires sufficient amount of sample (>50 cells) be provided for training when cell lineage information is not utilized. We have made the MATLAB source code of the procedure publicly available.

2.4. Four-dimensional tomographic imaging of gene expression pattern with the virtual reference embryo

Gene expression pattern graphics were generated through the integrative analysis of the virtual reference embryo model and the spatiotemporal gene expression profile. In silico histology sectioning were performed by slicing the virtual embryo model generated from the Voronoi diagram along each of the three body axes. For all the genes at each of the 210 time steps of embryo development, we performed *in silico* histological sectioning of the virtual embryo along the Anterior-Posterior (A-P), the Left-Right / Medial-Lateral (L-R) and the Ventral-Dorsal (V-D) body axis, and generating a series of 46, 31 and 20 two dimensional images respectively. Therefore, a total number of 5,173,980 gene expression pattern graphics images ((46 + 31 + 20) sectioning positions × 210 time points × 254 reporter genes) with high spatiotemporal resolution could be generated. Linear interpolation and graphics image smoothing with Gaussian kernel were then performed in a 3-dimensional mesh grid to evaluate the gene expression signal intensity at each grid position and developmental stage (Fig. 2A-B, Supplementary Fig. 4.). The expression patterns of the 6 genes that were selected as the positive controls of our analysis, *pha-4, hlh-1, cnd-1, end-3, irx-1, and glp-1* were widely studied. They were compared with the computationally constructed tomographic images and validated the results.

2.5. Four-dimensional tomographic section of anatomical patterns in a developing virtual embryo

The *in silico* anatomical pattern graphics were generated by lineage tracing the progenitor cells in the developing virtual embryo during gastrulation and early organogenesis. After the lineage-specific Voronoi diagram was generated, a 3-dimensional mesh grid was created. The mesh grid data points that were located inside the Voronoi cells were identified. The whole embryo point cloud model was then established by assembling the lineage specific point cloud models and color coded according to the cell lineage identity. Two-dimensional anatomical graphics were then computationally generated by sectioning the virtual embryo model along each of the three body axes at each development stage and anatomical position (Fig. 3, Supplementary Fig. 1, 11-12). The Voronoi diagram has been proved an effective method in the analysis of Biology Pattern Formation [3]. In this study on *C. elegans* embryo development, our result is accurate and consistent with the previous findings, which applied the algorithm to determine the direct cell neighboring relationships and the embryonic anatomical patterns [43].

The 4-dimensionl point cloud model of the developing reference virtual embryo was generated on the System Biology Deep Learning Workstation with 32-CPUs (Intel Xeon Silver 4215 CPU, 2.50 GHz), 4 GPUs (NVIDIA GeForce RTX 2080 Ti), in Matlab version 2020a, under Linux CentOS stream 8 Operating System. The Voronoi method can be applied to study C. elegans embryogenesis effectively at early developmental stage. At later stage of development, cell morphology varies significantly in size and shape. More sophisticated version of the Voronoi algorithm can be developed to characterize the geometry features of embryos at single cell and tissue scale.

2.6. Principal component analysis and K-means clustering of 4D gene expression pattern

At each developmental stage and slicing position along the three body axes, the 2-dimensional gray scale gene expression pattern graphics of all the 254 samples (120genes) were reshaped into 1-dimensional vectors. Principal Component Analysis [44] was then performed in the vector space (Fig. 2C-D). The top 100 leading eigenvectors with the largest eigenvalues were utilized for feature extraction. We performed K-means clustering with the extracted graphics image features and identified 20 distinct gene expression patterns with the images (Fig. 5). The center graphics image of each image cluster was obtained by taking the average of all the images in the cluster. The expression pattern images of each cluster along the A-P axis in the central, the most anterior and the most posterior positions were shown in Supplementary movies 23–25. Here, the three body position refer to the slicing plane at the 1/4, the 1/2 and the 3/4 position along each of the body axes respectively (Supplementary Fig. 1). We identified 20 clusters representing spatiotemporal differential gene expression patterns. For each cluster, we searched for enriched signaling pathways by performing Gene Ontology analysis for signaling-related terms. In total, 14 signaling pathways were identified in the central, the most anterior and the most posterior positions of the embryo. We assigned a numerical

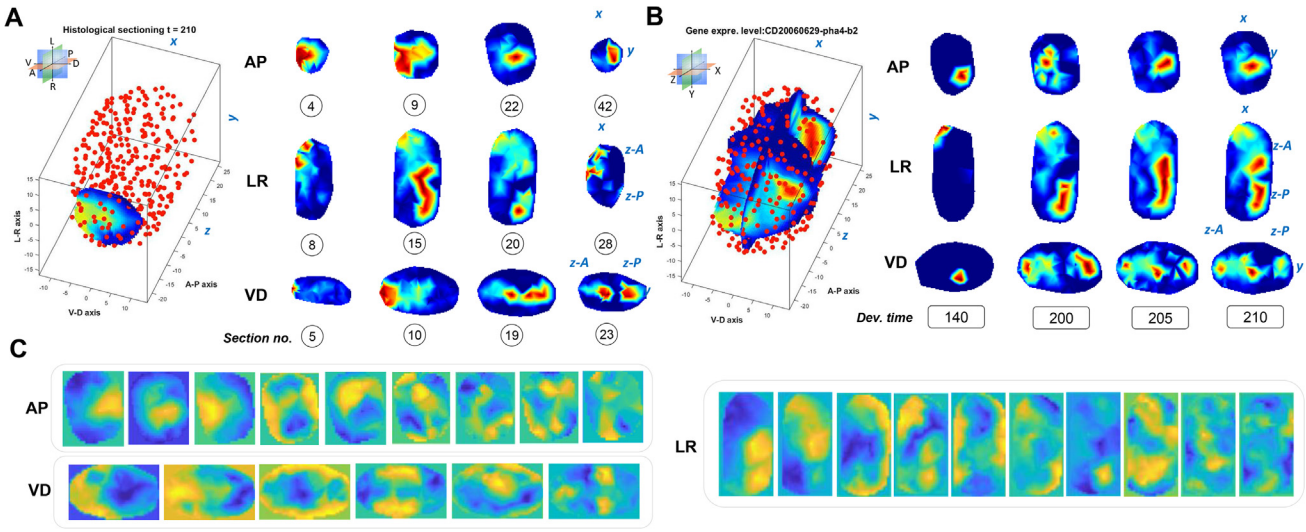


Fig. 2. Four dimensional tomographic imaging of gene expression patterns and Principal Component Analysis of images. (A) For each of the 254 embryo samples (120 genes) in the EPIC database and each of the 210 time steps, we conducted tomographic imaging with the virtual embryo along each of the 3 body axes. (B) The genes showed the dynamically changing expression patterns during embryogenesis. The central section images of the *pha-4* gene, for instance, was upregulated at the later stage of time step 137 to begin with the processes of pharynx and intestine organogenesis. (C) Principal Component Analysis of the images identified the tissue regions where gene expression levels varied most significantly, as highlighted in bright yellow colors. The leading Principal Component images were presented for each of the 3-body axis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

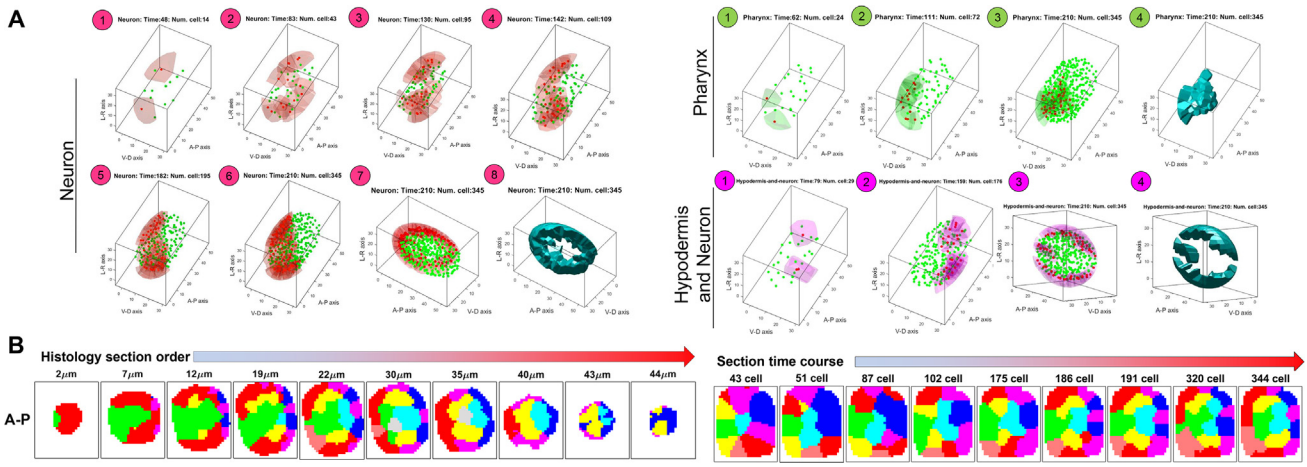


Fig. 3. Method of 4-dimensional tomographic imaging identifies collective cellular behaviors and complex tissue boundaries. (A) In silico cell lineage tracing with the virtual reference embryo identified the coordinated processes of cell proliferation, differentiation and migration. Neural progenitor cells (red color coded, Supplementary movie 10) started with proliferation at the anterior and the posterior regions of the embryo (Fig. 4A, Neuron ①-②). Pharynx progenitor cells (green color coded, Supplementary movie 12) were specified in the anterior body part of the embryo then began with proliferation (Pharynx ①-②). They became regionalized and formed a convex polygon shape in contacts with the neurons, the muscle and the intestine precursor cells (Pharynx ③-④). The hypodermis and neuron cells (purple color coded, Supplementary movie 13) were specified in the ventral embryo body part along the V-D axis (Hypodermis and neuron ①-②). They contributed to the development of the future body wall and the sensory organs of the animal (Hypodermis and neuron ③-④). (B) Serial section computer images were generated from the point cloud model color coded according to cell lineage type, which define the complex tissue boundaries. Representative slice computer images obtained at different sectioning positions and time points were presented. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ID to each of the 14 signaling pathways and generated a 14-bit binary barcode that indicates the "presence" – "absence" state of each pathway in each spatial expression cluster (white color: "presence;" black color: "absence"). We then performed hierarchical clustering and grouped the gene expression patterns with the binary pathway barcodes. We analyzed the geneset of each image cluster and identified the biology signaling pathways that were highly expressed by using the KEGG Gene Ontology database [45] (Fig. 5, Supplementary Fig. 7-9). The gene expression and the corresponding anatomical pattern images were retrieved from the database and superimposed for correlative image analysis. For each computer image cluster with tissue-specific gene expression patterns, we identified the putative biology signaling pathways that regulate tissue boundary formation and early organogenesis through Gene Ontology enrichment analysis [45] (Fig. 6, Supplementary Fig. 10.).

3. Result

3.1. Definition of biological and computational terms

Cell signaling events occur at multi-scales and at different levels. This work focuses mainly on the study of cell signaling at the morphogenetic domain level. To be accurate in discussion,

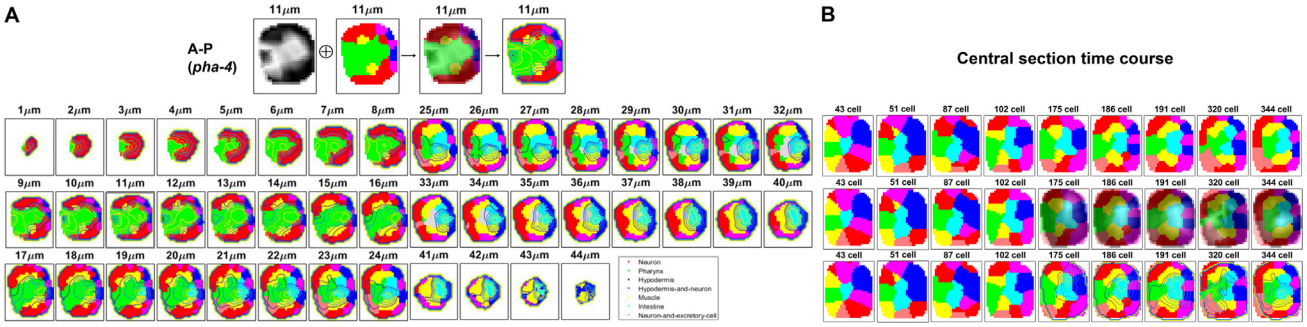
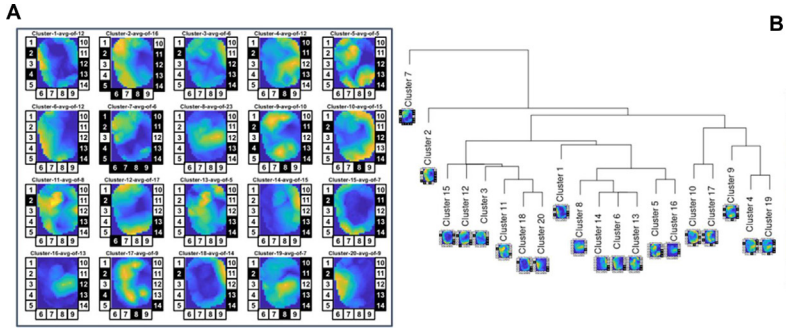


Fig. 4. Correlative image analysis of tissue boundary pattern formation along A-P axis during gastrulation. (A) The gene expression and the morphological pattern computer images generated from the same sections were superimposed for studying the relationship between gene regulation and morphogenesis. The gene expression pattern of *pha-4*, as an example, showed that at 345 embryonic cell stage, its expression domain was confined to the progenitor populations of pharynx (green color coded) and intestine (cyan color coded). (B) The time course image with central position section indicated that the pha-4 gene was upregulated at detectable level at around 175 cell stage. The spatiotemporal correlative image profile of another example gene, *hlh-1*, was presented in Supplementary Fig. 14. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



Code	Pathways	Code	Pathways
1	FoxO signaling pathway	8	Calcium signaling pathway
2	Phosphatidylinositol signaling system	9	Dorso-ventral axis formation
3	Wnt signaling pathway	10	TGF-beta signaling pathway
4	mTOR signaling pathway	11	Jak-STAT signaling pathway
5	MAPK signaling pathway	12	Hedgehog signaling pathway
6	Notch signaling pathway	13	ECM-receptor interaction,
7	ErbB signaling pathway	14	ABC transporters

Fig. 5. Correlative image analysis and the mining of KEGG database identified the biology signaling pathways in the morphogenetic domains (central body position). (A) We identified 20 distinct gene expression patterns by computer image clustering analysis in the projected feature space of Principal Component Analysis. (B) We performed GO term analysis and identified 14 highly expressed biology signaling pathways from the gene expression image clusters. We generated 14-bit binary barcode that indicates the "presence" – "absence" state of the pathways in detection (white color: "presence;" black color: "absence"). We then performed hierarchical clustering and grouped the gene expression patterns according to pathway barcode similarity. By correlative analysis of the gene expression and the anatomy patterns, we identified the cell populations that highly express the biology signaling pathways. We performed pairwise comparison with the barcodes and constructed the map of the morphogenetic domain signaling network. For instance, along the A-P axis, we detected the mTOR signaling pathway in cluster 12 (neuron) but not cluster 17 (muscle). The *mTOR* signaling pathway regulates a wide variety of cell Biology Processes, such as lipid metabolism, autophagy, protein synthesis, ribosome biogenesis, cytoskeletal organization and cell survival [26]. The results of analysis along the other 2 body axes of V-D and L-R were presented in the Supplementary Fig. 7.

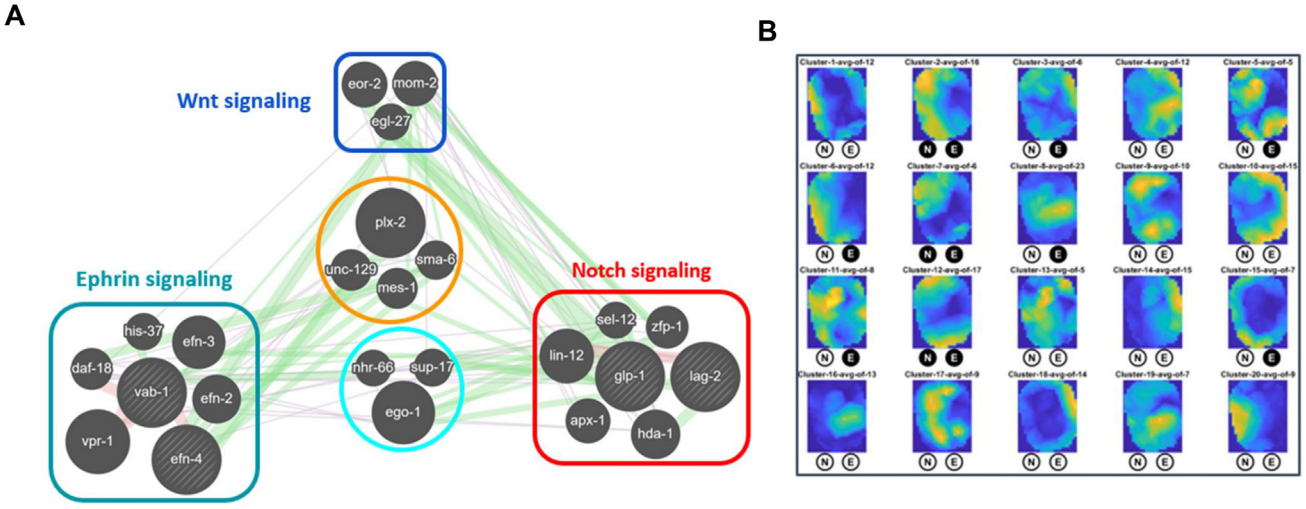


Fig. 6. The morphogenetic domain signaling network model suggested that the Notch and the Ephrin signaling pathways cross-talk and play important roles in regulating tissue morphogenesis. (A) We identified the differentially expressed genesets in morphogenetic domains by clustering and correlative image analysis. Results of GO term analysis indicate that Wnt, Notch and Ephrin pathway components were enriched in these genesets. The three pathways, Wnt (blue), Notch (red) and Ephrin (dark blue) crosstalk with each other through intermediate hub genes (orange circle). (B) The Notch and the Ephrin signaling pathways were highly expressed during embryo early organogenesis. The components of the pathways showed regional specificity of gene expression in multiple morphogenetic domains. (N and E stand for Notch and Ephrin signaling pathway components, respectively in the circle shape markers. The white color indicates "detected" and black color indicates "not detected."). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

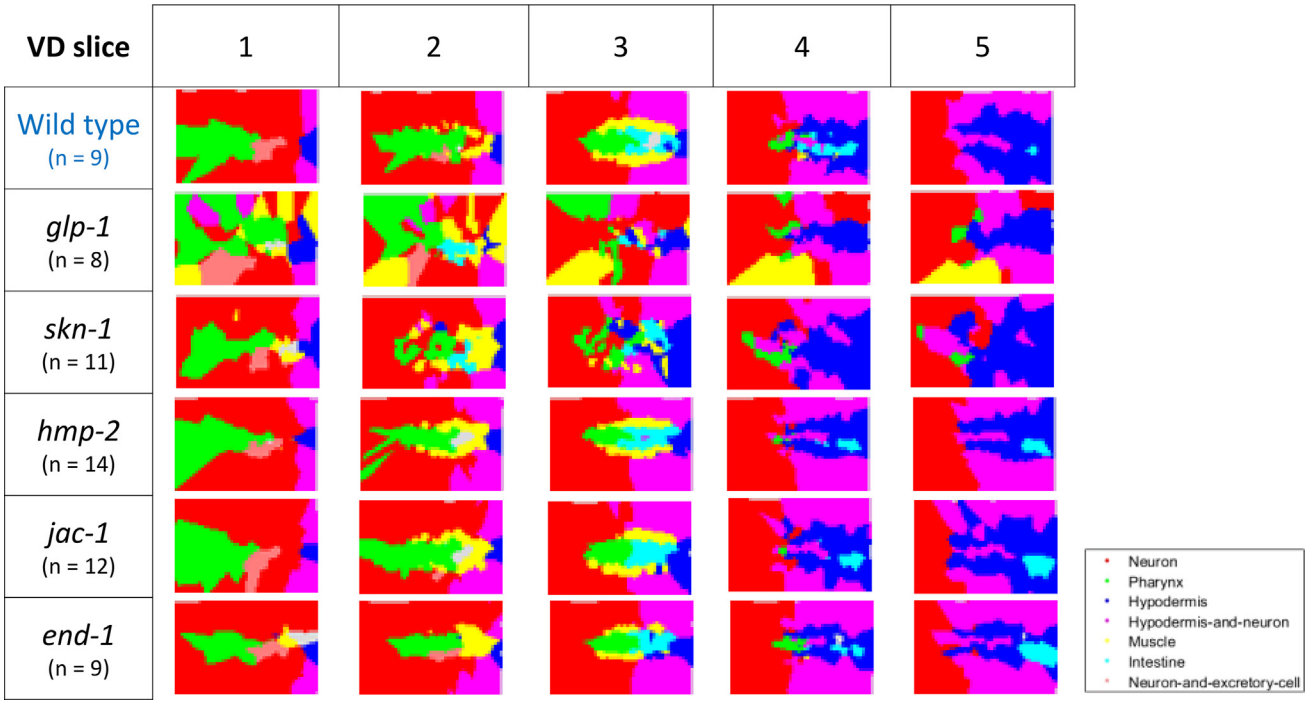


Fig. 7. Homozygotes mutants of the genes glp-1 and skn-1 at 345 embryonic stage showed the abnormally formed tissue boundaries and the loss of tissue continuity required for normal early organogenesis. The mutation of the gene glp-1, a central regulator of the Notch signaling pathway, as shown in the cell population signaling network in Fig. 6, caused severe abnormalities in tissue boundary pattern formation and cell fate specification. The mutation of the gene skn-1, a downstream effector of the Notch signaling pathway, showed the abnormal development of pharynx, muscle and intestine characterized by the loss of tissue continuity and the abruptly formed tissue boundaries. The gene hm-2 and jac-1 gave rise to no significant phenotypes compared to the WT. The end-1 homozygotes mutants displayed the abnormalities in intestine development (Supplementary Fig. 8).

we clarify the related biology and computational terms by definition and with examples.

Intracellular signaling network. Intracellular signaling is the mechanism in which signaling ligand molecules bind to the cell surface receptors to trigger the signaling cascades that activate or repress the cellular processes in proliferation, growth, differentiation, migration and programmed cell death. For instance, in *C. elegans* embryogenesis, the *WRM-1* protein initially accumulates in the nuclei of all cells, Wnt signaling promotes the asymmetric retention of *WRM-1* in the nuclei of the responding cells [46].

Cell-to-cell signaling network. Cells signal to each other by direct contacts [15] or by the release of a substance from one cell that is taken up by another cell [47]. The fate and behavior of the cell is determined through the communication with its neighboring cells. Cell-to-cell signaling plays the role in regulating cell fate decision for example at the 4-cell stage during *C. elegans* embryo development. Precisely, depending on the differential geometric setting of direct cell-to-cell interactions, Notch signaling is activated as a result of the ligand-receptor binding between the neighboring cells of P2 and ABp, but not P2 and ABa [48]. In another example, the cell-to-cell signaling networks through the gap-junctions also contributes to cell fate decision in the formation of the left–right asymmetric olfactory neuronal network [49].

Morphogenetic domain (Morphogenetic field). A morphogenetic domain is a group of cells able to respond to discrete, localized biochemical signals, which lead to the development of specific gene expression pattern [50,51]. The cellular fates within a morphogenetic domain are primarily specified prior to or during the process of organogenesis. However, the specific cellular programming of the individual cell in a domain is flexible. For instance, an individual cell in a cardiac field can be redirected via cell-to-cell signaling to replace the specific damaged or missing cells.

Morphogenetic domain signaling network. The morphogenetic domain signaling network is regulated by a population of cells that secret the molecules collectively to form the morphogen gradients that can diffuse over distance for directing cell communications and collective behaviors. The gradient can be interpreted by cells in a position dependent and level sensitive manner in regulating cell fate decision and behavior output [47,52,53]. The combinatorial code of the morphogen gradients played important roles in inducing complex tissue pattern formation in embryo development. For instance, *Wnt* act both as a long-range diffusible morphogen and through direct cell contact to regulate gene expression and to orient cell divisions, which plays critical role in regulating the patterning of anterior-posterior (A-P) body axes and stem cell self-renewal [52]. Recently, it was also reported that the *C. elegans* protein disulfide isomerases (PDIs) are the master regulator of *Wnt* secretion. The *PDI-1* is required for *Wnt* gradient formation and controls the *Wnt* gradient directed neuron migration over distance during *C. elegans* embryogenesis [47].

Computer image of spatial gene expression cluster. Each cluster image shows the spatiotemporally correlated expression pattern of a set of genes, for which the enriched biological signaling pathways were computationally identified. The combinatorial pattern of the gene expression cluster images predicts the progenitor cell populations and the spatiotemporal extent of the morphogenetic domains of the fast growing precursor organs.

3.2. PCA and K-means clustering analysis of 4-dimensional tomographic images identify distinct gene expression patterns that define morphogenetic domains

The key to better understanding the coordination of cellular differentiation in development is to look at the differential transcrip-

tion within and between cells [16,17,54]. We proposed the reference virtual embryo model integrated with the spatiotemporal gene expression profile. The model can be used to study the associations between gene expression regulation and morphological pattern formation. Specifically, we questioned, i) which set of genes showed similar expression patterns in space and time, since mechanistically embryogenesis is governed by the synergistic and competitive interactions of signaling molecules and transcription factors, that is the 'combinatorial code' of gene expression in space and time [2,55]. The specific combinatorial pattern of genes defines the cell lineage, the trajectory of the developmental cell, and the future cellular behavior outputs. ii) what is the quantitative relationship between the 'combinatorial codes' of gene expression patterns [55] and the dynamical morphological features of the fast-growing organs produced from germ layers.

To identify groups of spatially co-expressed genes, we performed four-dimensional tomographic image processing through dimensionality reduction (image feature extraction), followed by *K*-means clustering analysis. We first generated computer images of the gene expression pattern on the virtual reference embryo model. In the current study, we selected one time point (the last time step of 210 at 345 embryonic cell stage) and one sectioning position (the middle section) along each of the 3 body axes for the analysis on a subset of 762 images (3 sectioning positions $\times$ 1 time point $\times$ 254 reporter genes). The representative tomographic images for the gene *pha-4* were presented as an example, where the yellow and the red colors indicated the cell populations in the embryonic pharynx and the intestine regions where *pha-4* highly expressed (Fig. 2A, Supplementary movie 6). The gene also showed dynamically changing expression patterns over the time course of development. It was not up-regulated until the later stage of gastrulation (time step of 137) (Fig. 2B, Supplementary movie 7).

In order to identify the differential gene expression patterns in the tomographic projections, we performed Principal Component Analysis with the tomographic images generated along the 3 body axes. The leading PCs with the largest eigenvalues were presented in the order from left-to-right and top-to-bottom (Fig. 2C). The result of the eigenvalue plot (Supplementary Fig. 6) indicated that the top 50 PCs were sufficient in capturing over 99% of the variance in the graphics image dataset. A representative sample image (red box) can be accurately reconstructed by using up to 32 PCs, demonstrating the effectiveness of the simple method for image feature extraction. We also performed Principal Component Analysis on the graphics image dataset generated along L-R and V-D axis in the central, the most anterior and the most posterior positions (Supplementary Fig. 18–20). We used the first 100 PCs for dimensionality reduction and extracted the image features with full details for subsequent clustering analysis. We defined 20 distinct gene expression patterns in the central, the most anterior and the most posterior embryo body sections for the developmental stage by *K*-means clustering (Fig. 5, Supplementary Fig. 7–9). The genes that belong to each of the clusters were presented in Supplementary Tables 1–3. The gene expression image clusters identified the sets of transcription factor encoding genes that were spatiotemporally co-expressed, suggesting that they may interact in regulating the signaling and the transcriptional dynamics of embryogenesis.

Four dimensional tomographic imaging of virtual embryo anatomy identifies collective cellular behaviors.

Recently, Cao et al. [56] developed a deep learning computer algorithm and applied it to quantify the morphological parameters of the individual cells in *C. elegans* embryos. They generated a time-lapse 3D atlas of cell morphology for the *C. elegans* embryo from the 4- to 350-cell stages, including cell shape, volume, surface area, migration, nucleus position and cell–cell contact with resolved cell identities. The work arose widely interests as morphological changes in cell size and shape indicate the state transition of cells in gene regulation and biology function. We sought to extend this work by investigating the morphological features of the embryo at the tissue and organ level. Also, we established the method to interrogate the relationship between gene expression regulation and embryonic pattern formation. Following the cell lineage definition by Chen et al. [43], we developed a computer point-cloud algorithm that can be utilized to study the morphological features of the fast growing organ precursors that are derived from the collective activities of multiple cell populations. The proposed method will also help to understand at tissue and organ level the process of embryogenesis in the context of gene expression regulation for coordinated cell proliferation, differentiation and migration.

Specifically, at tissue and organ level, the anatomical structure of the *C. elegans* embryo is the result of natural evolution with the optimization goal of the worm to pass its genetic information to the next generation by survival, feeding and reproduction. Under this constraint, the worms evolve with its genetic program, giving rise to the formation of the three germ layers of ectoderm, endoderm and mesoderm through the process of gastrulation that lay out for early organogenesis. Already at the early stage of 26-cell when gastrulation just starts, for instance, the embryo begins to prioritize on the different populations of the progenitor cells according to their biological functions and importance for future nutrition acquisition and reproduction. The gastrulation program, through a series of coordinated cellular behaviors, setups the body plan to place the germ and the endoderm cells to the central compartment of the embryo body, so that these cells would be in a well-protected environment for maximizing the chance of survival and reproduction.

To explain the collective behaviors of the cells in position placement that happened in such a highly deterministic and orderly fashion, Meinhardt and Schnabel et al. [29] hypothesized a model of *C. elegans* embryo pattern formation in which cells autonomously generate a positional value on their surface depending on their fate. The cells then read the gradient positional information of their neighbors and try to match their own positional value with that of their neighbors. When there exists position discrepancy, the cell migrates actively relative to their neighbors for match and correction through the process of cell sorting. Torpe et al. [47] recently found that the function of *PDI-1*, a protein chaperone, is required to correctly form an anterior-posterior *EGL-20/Wnt* gradient during *C. elegans* embryonic development, which controls *EGL-20/Wnt*-dependent neuronal migration. Labouesse et al. [57] combine the lineage and organ/tissue specification models, and suggested that the existence of a cell fate map and the discovery of the domains of the organ/tissue identity genes imply another level of organization beyond cell lineage for patterning the *C. elegans* embryo.

In light of with these previous works, to better understand the dynamic process of pattern formation in gastrulation and early organogenesis in association with gene regulation, we performed *in silico* cell lineage tracing and computationally identified at population/domain level a variety of coordinated cellular processes in proliferation, differentiation, aggregative migration and the self-assembly of the cells into polarized tissues (Fig. 3A, Supplementary movie 10–16).

Specifically, as illustrated in Fig. 3A on the biology process of neurogenesis, two neuron progenitor cells (red color coded, Supplementary movie 10) initially were specified in the anterior and the posterior regions of the embryo body (Fig. 3A, Neuron ①-②). A 'hand-shaking' like process then occurred when the two populations of neural progenitor cells kept proliferating locally, then migrated collectively toward each other (Neuron ③-④), ultimately forming the continuous morphological structure that gives rise to the development of the precursor nervous system (Neuron

⑤-⑧). It has been recently demonstrated that the Wnt signaling gradient produced by the epidermal cells regulate the migration of neural progenitor cells from the posterior to the anterior region of the embryo [47]. It however remains still poorly understood at the earlier stage before 350-cells which exact molecular cue and signaling pathway contributed to the directed cell proliferation, placement and ultimately the formation of the continuum domain of the nervous system in the embryo.

In the development of the organ of pharynx, the pharynx progenitor cells (green color coded, Supplementary movie 12) were specified in the anterior body part of the embryo then began with proliferation (Pharynx ①-②). They became regionalized and formed a convex polygon shape in contact with the neurons, the muscle and the intestine precursor cells (Pharynx ③-④). In the developing virtual embryo model, development process of precursor pharynx followed three major steps, i) the specification and proliferation of the progenitor cells, ii) the formation of a manifold 'sheet like' structure through regionalization in the anterior dorsal region of the embryo body, iii) the three -dimensional outgrowth of the convex shape pharynx precursor out of the manifold sheet-like structure. It has been demonstrated that the Wnt signaling pathway played the critical role in regulating pharynx development. The exact setting of the gene circuitry and the spatiotemporal configuration of the signaling network that govern each step remain to be fully understood.

The hypodermis and neuron cells (purple color coded, Supplementary movie 13) were specified in the ventral embryo body part along the V-D axis (Hypodermis and neuron ①-②). They contributed to the development of the future body wall and the sensory organs of the animal (Hypodermis and neuron ③-④).

Neuron and excretory progenitor cells (Supplementary Fig. 11., light red color coded, Supplementary movie 14) were originally specified in the anterior ventral body part of the embryo (Neuron and excretory ①).They then migrated collectively in a group from the ventral to the dorsal side of the embryo body over long distance (Neuron and excretory ②-③). After reaching the destination domain, they became stabilized and self-assembled into the polarized 'sheet' like structure in contact with the other cell types of the neurons, the pharynx and the muscle progenitors (Neuron and excretory ④). It has been reported that the Wnt signaling gradient regulate the migration of several neuron cell types along the A-P body axes over long distance. There are relative fewer studies however on the collective migration of the cells along the V-D body axes of the embryo. It remains a question that instead of through local cell fate specification, why the small set of neuron and excretory progenitor cell were genetically programmed to migrate collectively from the ventral to the dorsal side over such long distance, which seems to be a quite error-prone and energy consuming step.

During the process of myogenesis, initially, the muscle progenitor cells (Supplementary Fig. 11., yellow color coded, Supplementary movie 11) were specified and began with proliferation in the central and the posterior region of the embryo body (Muscle ①-③), then went through the process of convergent extension, forming a 'U' shape morphological structure in contact with the other precursor cell types in pharynx and intestine (Muscle ④-⑧). It is worth noting that these muscle progenitor cells tend to undergo a cell sorting process, similar to what has been described in the 'cell focusing' model [29]. The cells established the order of position along the three body axes out of a random configuration. The exact identities of the molecules that encode the address information of the cells for spatial sorting, as suggested in the 'cell focusing' model, are to be elucidated.

In the development of the embryonic gut, starting with just one cell, the intestine progenitor cells (Supplementary Fig. 11, cyan color coded, Supplementary movie 15) were specified in the posterior position of the embryo (Intestine①). They then proliferated, became polarized, and ultimately self-aligned to form the column structure of the future intestine in contact with the pharynx, the muscle, and the hypodermis populations (Intestine ②-④). Similar to the muscle progenitor cells that establish the spatial order out of randomness, likely through the cell sorting mechanism, it can be important to understand if the muscle, the intestine and other cell populations have a centralized global signaling regulatory network that operate at morphogenetic domain level in coordinating the development of multiple precursor organs for body planning.

Embryogenesis depends on the precisely coordinated work of multiple cell populations that express differential signature genes for signaling, transcriptional regulation and biological functions. To study the morphology features of the precursor organs, we generated the point cloud models from the 3-dimensional Voronoi diagrams for each progenitor ("founder") cell population [10,43]. In doing so, we were able to standardize and normalize the data for Biometrics and Bioinformatics analysis (Supplementary Fig.12A). The point cloud models were assembled, and color coded according to cell lineage identity. Three views of the virtual embryo model from different angles were presented (Supplementary Fig. 12B, Supplementary movie 16).

We generated *in silico* 2D histological section graphics with the point cloud model along the A-P, the L-R and the V-D body axis (Fig. 3. B, Supplementary Fig.12. C-D, Supplementary movies 17–19). The representative slice graphics images showed the complex tissue boundary pattern in the developing embryo, where each cell population interacted directly with another or multiple populations. The genes and the signaling pathways that were differentially expressed across such morphogenetic domains (Figs. 5-6, Supplementary Fig. 7-10) played the important roles in maintaining cell population identity and tissue boundary pattern formation.

Also, we generated the serial section graphics of the virtual embryo at different development stages. The representative slice graphics of the sectioning over time course were presented (Fig. 3. B, Supplementary Fig. 12C-D). The tissue boundary patterns are in concord with the spatiotemporal expression patterns of the genes, for example, pha-4 and hlh-1. The progenitor cell populations may communicate by signaling across the morphogenetic domains for coordinating behaviors in proliferation, collective migration, and the self-assembly into polarized tissues. (Supplementary movies 20–22).

3.3. Correlative image analysis identified 14 biological signaling pathways in the morphogenetic domains

Embryo development is governed by the dynamic interactions of signaling and transcription factors. Possible interactions can be predicted by identifying the signaling pathways that are statistically enriched amongst genes expressed in each position and time, relative to genes expressed in other positions or times. By overlaying the gene expression and the anatomy images for correlative analysis, we were able to interrogate which molecular pathways play the potential roles in regulating development of which organ precursor cell populations. Also, it should be noted that the biological activities of the computationally identified signaling pathways are subjected to future experimental validations through functional studies.

Specifically, we constructed a spatiotemporal map of the biology signaling pathways that may play some roles in regulating *C. elegans* morphogenesis. For instance, the well-studied expression pattern of the gene *pha-4*, a master regulator of pharynx development and a target gene of the Notch signaling pathway [15], was restricted to the progenitor cell populations of embryonic pharynx (green color coded) and intestine (cyan color coded) in the anterior and the posterior body parts respectively (Fig. 4A, Supplementary

Fig. 13A-B). The expression pattern of *hlh-1*, an important conserved muscle regulator, exhibited regional specificity of expression in the populations of body muscle (yellow color coded) and intestine muscle (yellow) (Supplementary Fig. 14). It was reported that the *Wnt/MAPK* signaling precedes the expression of *hlh-1* in activating myogenesis [58]. The downstream effectors of the two signaling pathways intertwine at the transcriptional level, regulating the identities and behaviors of the muscle and the pharynx progenitor cells at the tissue boundaries.

To identify groups of spatially coexpressed genes, we clustered the embryonic graphics images along the 3 body axes, A-P, V-D and L-R, in the most anterior, the central and the most posterior positions (Fig. 5, Supplementary Fig. 7-9, the center mean images of the 20 clusters identified were shown in the order from left-to-right and top-to-bottom). The cell populations at the same level of hierarchical tree with similar barcodes may make use of the same set of pathways for information processing and cellular function. Those in different clusters may make use of the differential combination of the pathways for executing population-specific biology functions. Thus, by correlating gene expression patterns across the anatomy, we identified signaling pathways that were highly enriched in each cell population and morphogenetic domain.

We then performed pairwise comparison with the pathway barcodes and constructed a map of the morphogenetic domain cell signaling co-occurrent network.

In the embryo central body position, along the A-P axis (Fig. 5A), we detected the mTOR signaling pathway in cluster 12 (neuron) but not cluster 17 (muscle). The *mTOR* signaling pathway regulates a wide variety of cell Biology Processes, such as lipid metabolism, autophagy, protein synthesis, ribosome biogenesis, cytoskeletal organization and cell survival [59]. Along the V-D axis (Supplementary Fig. 7B), ECM-receptor signaling pathway was detected in cluster 2 and 18 (muscle) but not cluster 17 (pharynx). Cell adhesion plays an important role in regulating collective cellular behaviors in proliferation, differentiation, migration, apoptosis and in the assembly of polarized tissues [60]. Mutating the extracellular matrix (ECM) receptor encoding gene, *PAT-2/PAT-3* (integrin) [61], leads to failed epithelial morphogenesis and defects in embryo body elongation [62]. Along the L-R axis (Supplementary Fig. 7C), we identified *TGF-β* pathway in cluster 5, 12 and 16 (hypodermis, intestine and pharynx). The *TGF-β* signaling pathway plays the critical role in body size regulation. However, its specific function in other cell populations during early embryogenesis remains to be fully elucidated [63]. In the central body position of the embryo, we detected the molecular pathways of *TGF-β*, *mTOR* and ECM-cell signaling in the muscle, the neuron, the pharynx, the hypodermis and the intestine populations.

In the most anterior position along the A-P axis (Supplementary Fig. 7A), we detected broad enrichment of the Notch signaling pathway, for instance, in cluster 2 (pharynx), cluster 5 (hypodermis and neuron) and cluster 10 (muscle), which is consistent with an important role of Notch in the regulation of cell fate and the coordination of progenitor cell populations during gastrulation and early organogenesis [64,65]. During the digestive tube morphogenesis, specifically, Notch signaling is required to prevent cross-fusion between pm8 (pharyngeal muscle) and *vpi-1*, the pharyngeal-intestinal valve cell that contributes to the formation of the tube structure linking pharynx to intestine [66].

Along the V-D axis (Supplementary Fig. 8B), the calcium signaling pathway was enriched in cluster 6 (muscle) and cluster 19 (pharynx) but not cluster 9, 14 and 17 (anterior region neuron). This pathway plays the roles in regulating muscle contraction and pharynx pumping [67], and contributes to control intercellular communication, germ layer formation and the establishment of body axes during gastrulation [68]. Along the L-R axis (Supplementary Fig. 8C), the *MAPK* signaling pathway is detected in cluster 1, 7, 8 and 9 (muscle) but not cluster 2 and 19 (anterior region neuron). The gene *mpk-1*, the *C. elegans* homolog of the conserved mitogen activated protein (*MAP*) kinase *ERK*, is the repressor of *par-1* gene that is required for cell polarity establishment and asymmetric division. The loss of function of *mpk-1* restores the viability of *par-1* mutant with corrected phenotypes in the asynchronous division of cells and the asymmetric distribution of cell fate specification markers [69].

In the most posterior body position of embryo, along the A-P axis (Supplementary Fig. 9A), we detected enrichment of the *PI3K/Akt* pathway, a member of phosphatidylinositol signaling system, in cluster 20 (neuron) but not in cluster 4 and 12 (hypodermis-and-neuron). During neuron development, *PI3K-Akt* is critical in mediating neuronal protrusion [36], and synaptic polarity [70]. Along the V-D axis (Supplementary Fig. 9B), we identified the 'ECM receptor interaction' pathway that was specifically expressed in cluster 13 (muscle), but not other clusters. Recently, it was reported that muscle functions as a connective tissue and source of extracellular matrix in planarians the flatworms [71]. It will be interesting to determine whether this role of muscle cells is evolutionarily conserved between these very distant species. Along the L-R axis (Supplementary Fig. 7F), we detected the *Jak-STAT* signaling pathway that plays the critical roles in regulating the development of neural progenitor cells (NPCs) and astrocytes after spinal cord injuries [72]. Not unexpectedly we detected it in cluster 1 (neuron) but not cluster 7 and 20 (posterior neuron-hypodermis domains).

Multiple signaling pathways can act together to coordinate tissue morphogenesis. In *C. elegans*, there are few reports on the computational identification of cell signaling networks that are co-expressed over morphogenetic domains. Through the analysis of a gene regulatory network model, we showed that components of the Notch and the Ephrin signaling pathways were highly expressed across multiple morphogenetic domains at the stages of gastrulation and early organogenesis (Fig. 6, Supplementary Fig.10, N and E stand for Notch and Ephrin signaling pathway components respectively in the circular markers. The white color indicates "detected" and black color indicates "not detected."). We identified differentially expressed gene sets in these morphogenetic domains by *K*-means clustering. The results of Gene Ontology analysis indicate that the Wnt, the Notch and the Ephrin pathway components were enriched in these gene sets. In our clustering, the three pathways, Wnt, Notch and Ephrin signaling are linked with each other through intermediate hub genes (Fig. 6A). To the center of the cell signaling network, the genes *plx-2*, *sma-6*, *mes-1* and *unc-129*, interact with all the three pathways of Wnt, Notch and Ephrin signaling, while the other three genes that form a regulatory module, including *nhr-66*, *sup-17* and *ego-1*, interact with the two signaling pathways of Notch and Ephrin. Functional genomics studies showed that the mutation of the gene *sup-17*, a coordinative regulator that has interaction with both Ephrin and Notch signaling pathways, caused embryonic lethality. The mutation of *mes-1*, the intermediate hub gene of all the three pathways, leads to defects in body elongation [73], symmetric division of germline cells P2 and P3 [74], and lack of intestine [75].

The gene *glp-1* is a central regulator of the Notch signaling pathway, as shown in the cell population signaling co-occurrent network in Fig. 6. By applying our tomographic graphics analysis to a mutant embryo dataset [37], it showed that *glp-1* mutants have severe abnormalities in tissue boundary pattern formation and cell fate specification. Similarly, mutation of the gene *skn-1*, a downstream effector of the Notch signaling pathway, resulted in the loss of tissue continuity and the malformed tissue boundaries in pharynx, muscle and intestine (Fig. 7, Supplementary Fig. 17). The morphogenetic domain crosstalk of the biology signaling pathways

through the intermediate hub genes is critical for coordinating the process of gastrulation and early organogenesis. Our correlative image analysis method provides a useful means for the quantitative modeling of cell population signaling networks at the level of morphogenetic domains.

3.4. Software tool: Embryo aligner version 1.0

To study the process of embryogenesis quantitatively and precisely, individual embryo samples have to be accurately aligned with the cell position information. Richards et al. [76] aligned individual samples by normalizing for embryo shape (based on the size and shape of an ellipse that encompasses the nuclei at the 100-cell stage) and rotated to match a common initial axis orientation. The algorithm has been applied to the analysis of cell position variability under the normal and the temperature stress conditions. The method worked effectively on the embryo samples that were generated from the same set of experiment when the time frame of video recording is synchronized. In this work, in addition to the application of camera time and cell number age for embryo time frame matching and alignment [76], we further proposed to make full use of the precisely annotated cell lineage information by biologists for accurate alignment. In doing so, we are able to perform cross-experiment comparisons on the embryo samples for data integration and meta-analysis.

We built the Software tool 'Embryo aligner version 1.0′ and made it publicly available for download from GitHub (https://github.com/csniuben/embryo_aligner/wiki). The program is in Open Source and has the Graphical User Interface. There are two function modules in the Software for, i) Spatiotemporal alignment of the embryo samples with cell position information across development stages, and ii) Generation of the virtual embryo model from aligned individual embryo samples. The input to the Software are the imaging scale factors along the A-P, V-D, L-R body axis, and the embryo sample coordinate files generated from the StarryNite program [8] which is most widely used by the *C. elegans* research community. The output of the Software is the virtual embryo's cell identity and position file in '.csv' format. Give the imaging scale factor, the Software can align the embryo samples across experiments or genotypes for quantitative phenotype analysis between Wild Type and mutant models.

We trained the Machine Learning model with the Random Forest Tree Regression algorithm and predicted cell locations in the aligned reference embryo at each development point from cell gene expression profiles. We performed 2-fold cross-validation test in lineage-informed and random shuffled mode of the training datasets respectively. When the cell lineage information was provided for model training, the prediction accuracy was improved in cell localization.

4. Discussion

In this study, we developed a computational framework for analyzing the putative cell population signaling networks that may regulate *C. elegans* embryo gastrulation and early organogenesis. We established a developing virtual embryo model and applied it to construct spatiotemporal expression profiles by data pooling from spatial single cell gene reporter assay experiments. We generated a computer image dataset of the virtual embryo model and performed correlative *in silico* image analysis on the dynamic embryonic gene expression and anatomy patterns. This analysis method helps us to dissect the relationship between gene expression regulation and tissue morphogenesis at cell population level. We trained the Random Forest Regression model that can predict accurately the cell positions from the gene expression profiles in the developing embryo. We performed *in silico* cell lineage tracing and identified the putative boundary patterns for each progenitor cell population. We detected the biology signaling pathways that were highly expressed in each morphogenetic domain and constructed the cell population signaling network model that may regulates some complex tissue boundary pattern formation through integrative Biometric and Bioinformatics analyses.

The current set of 254 samples (120 distinct genes) analyzed were selected from the EPIC Database, which contains patterns for ~10% of the transcription factor encoding genes in the *C. elegans* genome. Our computational pipeline is designed with scalability for genome-wide analysis as the EPIC database and other resource from laboratories world-wide keep updating, with ease for standardized processing, normalization and modeling.

Also, it is known that the relationship between cell position and gene expression is often nonlinear. Embryonic tissue morphogenesis is regulated by complex biological signaling networks and gene transcriptional circuitries that can hardly be modeled with linear equations. Owing to the complexity and the dynamic nature of the developing embryo, essentially a time-variant nonlinear open system, it is extremely challenging to predict causative relationship between gene expression and tissue morphogenesis without taking a data-driven modeling approach. Recently, researchers have made significant progresses in the area of Image Pattern Recognition and Image Understanding by Deep Learning. Deep Neural Networks can be trained to learn and extract useful knowledge from Big Image Dataset by leveraging the power of GPUs for parallel information processing. Deep Learning has been widely adopted for numerous real-world applications [30,31]. In Computational Biology, it has been applied successfully to image analysis at single cell level to study embryo development, for instance, cell membrane segmentation [56] and cell nuclear state classification [77,78]. At a cell population level, Deep Learning makes it possible to model large nonlinear complex systems directly from the raw sample images and gene expression profiles, without explicitly using Partial Differential Equations.

The developing virtual embryo model that we established produced two types of computer images on gene expression and embryo anatomy patterns. Future work will require the development of Deep Learning Neural Network models for such image analysis to study the nonlinear causative relationship between gene expression regulation, tissue morphogenesis in *C. elegans* early development. The method can also be extended to other animal models or applied to the study of human diseases in tumorgenesis and aging-related degenerative disorders through the collaborations with clinical experts.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work was supported in part by NIH/MH109665 (MQZ) and NCI/CA245294 (MQZ). MQZ would also acknowledge the Cecil H. and Ida Green Distinguished Chair in Systems Biology Science. The authors thank for Professor Chao Tang and his student Dr. Guoye Guan for their helps in data processing. The authors also thank for Dr. Zhuo Du for his constructive suggestions in reading and revising the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.05.058.

References

[1] Raspopovic J et al. Digit patterning is controlled by a Bmp-Sox9-Wnt Turing network modulated by morphogen gradients. Science 2014;345 (6196):566–70.
[2] Tan Z et al. Synergistic co-regulation and competition by a SOX9-GLI-FOXA phasic transcriptional network coordinate chondrocyte differentiation transitions. PLoS Genet 2018;14(4):e1007346.
[3] Cheng CW et al. Predicting the spatiotemporal dynamics of hair follicle patterns in the developing mouse. Proc Natl Acad Sci 2014;111:2596–601.
[4] Murray JD. Mathematical biology: I. An introduction. Interdisciplinary applied mathematics. Mathematical Biology. Springer; 2002.
[5] Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science, 1998. 282(5396): p. 2012-8.
[6] Sonnhammer EL, Durbin R. Analysis of protein domain families in *Caenorhabditis elegans*. Genomics 1997;46(2):200–16.
[7] Lai CH et al. Identification of novel human genes evolutionarily conserved in *Caenorhabditis elegans* by comparative proteomics. Genome Res 2000;10 (5):703–13.
[8] Murray JI et al. The lineaging of fluorescently-labeled *Caenorhabditis elegans* embryos with StarryNite and AceTree. Nat Protoc 2006;1(3):1468–76.
[9] Sulston JE, Horvitz HR. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. Dev Biol 1977;56(1):110–56.
[10] Sulston JE et al. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Dev Biol 1983;100(1):64–119.
[11] Cook SJ et al. Whole-animal connectomes of both *Caenorhabditis elegans* sexes. Nature 2019;571(7763):63–71.
[12] Brenner S. Nobel lecture. Nature's gift to science. Biosci Rep 2003;23(5–6):225–37.
[13] Carthew RW. Gene silencing by double-stranded RNA. Curr Opin Cell Biol 2001;13(2):244–8.
[14] Chalfie M. GFP: lighting up life (Nobel Lecture). Angew Chem Int Ed Engl 2009;48(31):5603–11.
[15] Girard LR et al. WormBook: the online review of *Caenorhabditis elegans* biology. Nucleic Acids Res 2007;35(Database issue):D472–5.
[16] Tintori SC et al. A Transcriptional Lineage of the Early *C. elegans* Embryo. Dev Cell 2016;38(4):430–44.
[17] Packer JS et al. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. Science 2019;365(6459).
[18] Ma X et al. A 4D single-cell protein atlas of transcription factors delineates spatiotemporal patterning during embryogenesis. Nat Methods 2021;18 (8):893–902.
[19] Vuong-Brender TT, Yang X, Labouesse M. *C elegans* Embryonic Morphogenesis. Curr Top Dev Biol 2016;116:597–616.
[20] Lardennois A et al. An actin-based viscoplastic lock ensures progressive body-axis elongation. Nature 2019;573(7773):266–70.
[21] Vergara HM et al. Whole-organism cellular gene-expression atlas reveals conserved cell types in the ventral nerve cord of *Platynereis dumerilii*. Proc Natl Acad Sci U S A 2017;114(23):5878–85.
[22] Hartmann J et al. An image-based data-driven analysis of cellular architecture in a developing tissue. Elife 2020;9.
[23] Hagolani PF et al. On the evolution and development of morphological complexity: A view from gene regulatory networks. PLoS Comput Biol 2021;17 (2):e1008570.
[24] Bao Z et al. Automated cell lineage tracing in *Caenorhabditis elegans*. Proc Natl Acad Sci U S A 2006;103(8):2707–12.
[25] Murray JI et al. Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. Nat Methods 2008;5(8):703–9.
[26] Santella A, Du Z, Bao Z. A semi-local neighborhood-based framework for probabilistic cell lineage tracing. BMC Bioinf 2014;15:217.
[27] Moore JL, Du Z, Bao Z. Systematic quantification of developmental phenotypes at single-cell resolution during embryogenesis. Development 2013;140 (15):3266–74.
[28] Giurumescu CA et al. Quantitative semi-automated analysis of morphogenesis with single-cell resolution in complex embryos. Development 2012;139 (22):4271–9.
[29] Schnabel R et al. Global cell sorting in the *C. elegans* embryo defines a new mechanism for pattern formation. Dev Biol 2006;294(2):418–31.
[30] Deng J et al. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[31] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84–90.
[32] Fawkner-Corbett D et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. Cell 2021;184(3):810–826.e23.
[33] Karaiskos N et al. The *Drosophila* embryo at single-cell transcriptome resolution. Science 2017;358(6360):194–9.
[34] Ebbing A et al. Spatial Transcriptomics of *C. elegans* males and hermaphrodites identifies sex-specific differences in gene expression patterns. Dev Cell 2018;47(6):801–813.e6.
[35] Segal D et al. Feedback inhibition of actin on Rho mediates content release from large secretory vesicles. J Cell Biol 2018;217(5):1815–26.
[36] Murray JI et al. Multidimensional regulation of gene expression in the *C. elegans* embryo. Genome Res 2012;22(7):1282–94.
[37] Li X et al. Systems properties and spatiotemporal regulation of cell position variability during embryogenesis. Cell Rep 2019;26(2):313–321.e7.
[38] Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007;315(5814):972–6.
[39] D'Haeseleer P. How does gene expression clustering work? Nat Biotechnol 2005;23(12):1499–501.
[40] Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 2002;97(458):611–31.
[41] Goodall C. Procrustes methods in the statistical analysis of shape. J Roy Stat Soc: Ser B (Methodol) 1991;53(2):285–321.
[42] Liaw A, Wiener M. Classification and regression by randomForest. R News 2002;2(3):18–22.
[43] Chen L et al. Establishment of signaling interactions with cellular resolution for every cell cycle of embryogenesis. Genetics 2018;209(1):37–49.
[44] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometr Intell Lab Syst 1987;2(1):37–52.
[45] The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res, 2019. 47(D1): p. D330-d338.
[46] Nakamura K et al. Wnt signaling drives WRM-1/beta-catenin asymmetries in early *C. elegans* embryos. Genes Dev 2005;19(15):1749–54.
[47] Torpe N et al. A protein disulfide isomerase controls neuronal migration through regulation of Wnt secretion. Cell Rep 2019;26(12):3183–3190.e5.
[48] Hutter H, Schnabel R. Establishment of left-right asymmetry in the *Caenorhabditis elegans* embryo: a multistep process involving a series of inductive events. Development 1995;121(10):3417–24.
[49] Chuang CF et al. An innexin-dependent cell network establishes left-right neuronal asymmetry in *C. elegans*. Cell 2007;129(4):787–99.
[50] Jacobson AG, Sater AK. Features of embryonic induction. Development 1988;104(3):341–59.
[51] Stapornwongkul KS, Vincent JP. Generation of extracellular morphogen gradients: the case for diffusion. Nat Rev Genet 2021;22(6):393–411.
[52] Zacharias AL et al. Quantitative Differences in Nuclear β-catenin and TCF Pattern Embryonic Cells in *C. elegans*. PLoS Genet 2015;11(10):e1005585.
[53] Katz WS et al. Different levels of the *C. elegans* growth factor LIN-3 promote distinct vulval precursor fates. Cell 1995;82(2):297–307.
[54] Alicea B, Gordon R, Portegys TE. Data-theoretical synthesis of the early developmental process. Neuroinformatics 2021.
[55] Zacharias AL, Murray JI. Combinatorial decoding of the invariant *C. elegans* embryonic lineage in space and time. Genesis 2016;54(4):182–97.
[56] Cao J et al. Establishment of a morphological atlas of the *Caenorhabditis elegans* embryo using deep-learning-based 4D segmentation. Nat Commun 2020;11 (1):6254.
[57] Labouesse M, Mango SE. Patterning the *C. elegans* embryo: moving beyond the cell lineage. Trends Genet 1999;15(8):307–13.
[58] Fukushige T, Krause M. The myogenic potency of HLH-1 reveals wide-spread developmental plasticity in early *C. elegans* embryos. Development 2005;132 (8):1795–805.
[59] Laplante M, Sabatini DM. mTOR signaling at a glance. J Cell Sci 2009;122(Pt 20):3589–94.
[60] Brown NH. Cell-cell adhesion via the ECM: integrin genetics in fly and worm. Matrix Biol 2000;19(3):191–201.
[61] Gettner SN, Kenyon C, Reichardt LF. Characterization of beta pat-3 heterodimers, a family of essential integrin receptors in *C. elegans*. J Cell Biol 1995;129(4):1127–41.
[62] Labouesse M. Role of the extracellular matrix in epithelial morphogenesis: a view from *C. elegans*. Organogenesis 2012;8(2):65–70.
[63] Dineen A, Gaudet J. TGF-β signaling can act from multiple tissues to regulate *C. elegans* body size. BMC Dev Biol 2014;14:43.
[64] Priess JR. Notch signaling in the *C. elegans* embryo. WormBook 2005:1–16.
[65] Djabrayan NJ et al. Essential role for Notch signaling in restricting developmental plasticity. Genes Dev 2012;26(21):2386–91.
[66] Rasmussen JP et al. Notch signaling and morphogenesis of single-cell tubes in the *C. elegans* digestive tract. Dev Cell 2008;14(4):559–69.
[67] Alvarez-Illera P et al. Long-term monitoring of Ca2+ dynamics in *C. elegans* pharynx: an in vivo energy balance sensor. Oncotarget 2016;7(42):67732–47.
[68] Webb SE, Miller AL. Calcium signalling during embryonic development. Nat Rev Mol Cell Biol 2003;4(7):539–51.
[69] Spilker AC et al. MAP kinase signaling antagonizes PAR-1 function during polarization of the early *Caenorhabditis elegans* embryo. Genetics 2009;183 (3):965–77.
[70] Kimata T et al. Synaptic polarity depends on phosphatidylinositol signaling regulated by myo-inositol monophosphatase in *Caenorhabditis elegans*. Genetics 2012;191(2):509–21.
[71] Cote LE, Simental E, Reddien PW. Muscle functions as a connective tissue and source of extracellular matrix in planarians. Nat Commun 2019;10(1):1592.
[72] Wang T et al. The role of the JAK-STAT pathway in neural stem cells, neural progenitor cells and reactive astrocytes after spinal cord injury. Biomed Rep 2015;3(2):141–6.
[73] Wang S et al. A high-content imaging approach to profile *C. elegans* embryonic development. Development 2019;146(7).
[74] Berkowitz LA, Strome S. MES-1, a protein required for unequal divisions of the germline in early *C. elegans* embryos, resembles receptor tyrosine kinases and

is localized to the boundary between the germline and gut cells. Development 2000;127(20):4419–31.
[75] Bei Y et al. SRC-1 and Wnt signaling act together to specify endoderm and to control cleavage orientation in early *C. elegans* embryos. Dev Cell 2002;3(1):113–25.
[76] Richards JL et al. A quantitative model of normal *Caenorhabditis elegans* embryogenesis and its disruption after stress. Dev Biol 2013;374(1):12–23.
[77] McDole K et al. In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. Cell 2018;175(3):859–876.e33.
[78] Stegmaier J et al. Real-time three-dimensional cell segmentation in large-scale microscopy data of developing embryos. Dev Cell 2016;36(2):225–40.