EPICViz: An interactive visualization of *C. elegans* embryogenesis and gene expression

Melissa Chiasson chiasson@uw.edu

Timothy Durham tdurham@uw.edu

Andrew Hill ajh24@uw.edu

Ning Li ningli30@uw.edu

ABSTRACT

Caenorhabditis elegans (C. elegans) is a widely used genetic model, yet we still lack fundamental knowledge about how gene expression shapes its development and cell fate. We developed a visualization of gene expression during *C. elegans* embryogenesis to address this gap in knowledge. The dataset used consisted of coordinates for each cell, its diameter, its cell type and lineage, and binarized expression values for 227 genes. Our final version tracks development through time using a 3D plot of the embryo, 2D projections of the 3D embryo, a lineage tree, a principal components analysis (PCA) plot of gene expression values, and a heatmap of gene expression values. Users can manipulate the 3D plot to change views, highlight specific tissue types and lineages, and create gene expression enrichment reports for cells of interest. This visualization can be used by researchers in the *C. elegans* community to explore gene expression patterns during development and facilitate hypothesis generation.

Author Keywords

Caenorhabditis elegans, *C. elegans*, development, embryogenesis, gene regulation, gene expression.

INTRODUCTION

C. elegans is a small roundworm used widely as a model organism in genetics and genomics. Its development has been well studied; each worm takes around 14 hours to grow from a single fertilized cell to a hatched larvae with 558 cells. This process of embryonic development progresses in a stereotyped pattern that follows an invariant cell lineage; the same branches in this tree always produce the same tissues in the hatched worm [1] (Figure 1).

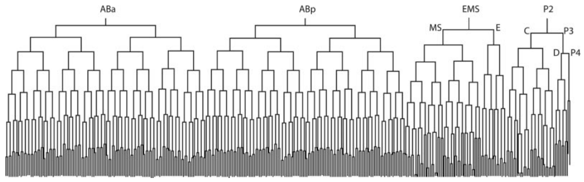


Figure 1. Lineage tree of the first 350 cells in *C. elegans* embryogenesis. Adapted from [3].

Development is the process by which cells derived from a single fertilized egg divide and differentiate to form the diverse tissues of the adult organism. In this process, cells start out very similar to each other, but over time express different sets of genes that specify the unique fate of each cell. Some of development is deterministic, encoded in the cells themselves; however each cell's fate is also influenced by the local environment and interactions with its neighbors. Thus, to understand development, we must consider the gene expression patterns of each cell within the broader context of the developing embryo.

Despite intense study of *C. elegans*, currently there does not exist a resource where one can interactively visualize *C. elegans* embryo development and interrogate how gene expression patterns change with time, lineage, or cell type. This limits exploration of data that could yield novel insights into the gene regulation of development.

RELATED WORK

C. elegans Gene Expression and Development Analysis and Visualizations

Because *C. elegans* development follows an invariant pattern, much work has focused on making computational tools that can track and annotate cells from microscopy images or movies [2], [3]. These tools are helpful for generating datasets like the one we implemented (described in **METHODS**) by identifying cells and tracking their positions over developmental time. Automated tracking and annotation of cells enables higher-level analysis of biological characteristics, like gene expression, that influence development.

Indeed, computational analysis of a subset of the data we use in our visualization yielded insights into *C. elegans* embryogenesis [4]. By analyzing expression of 127 genes, the study found gene patterns associated with terminal tissue type and spatial positions in the embryo; it also found sequential patterns of gene induction that could be indicative of regulatory cascades. No visualization accompanied this study, however, making it difficult for other researchers to explore gene expression patterns in development. It is this kind of analysis that we hope our visualization can make more accessible to the broader *C. elegans* community.

In addition to developing embryogenesis cell tracking and gene expression software, 3D modeling of *C. elegans* anatomy has been explored. The OpenWorm project is an interactive 3D model of *C. elegans*' muscular system and neural network [5]. Users can query specific cells and visualize the connections between them to better understand how neurons and muscles coordinate. We implemented a similar framework in our visualization.

Work in Other Model Organisms

Gene expression visualizations have been explored in other model organisms. ViBE-Z is a software package that generates graphics of gene expression in the zebrafish larval brain [6]. This helps in identifying colocalization of specific genes in specific substructures of the brain, but does not allow for interactive exploration. In *Drosophila*, a tool called MULTEESUM was developed to compare spatial and temporal gene expression data at a specific stage in embryogenesis [7]. At this stage, the drosophila embryo is shaped like a hollow tube, so the authors could convert a 3D embryo to two dimensions. In the visualization, users can select specific cells in this 2D rendering and view small multiples that display expression patterns for genes of interest over time. Both of these tools include features, like small multiples and brushing and linking, we sought to include in our own visualization.

METHODS

Data

The Expression Patterns In *C. elegans* (EPIC – http://epic2.gs.washington.edu/Epic2) project has generated a dataset that describes the spatial orientation of every cell during the first ~350 minutes of *C. elegans* embryogenesis, its developmental lineage and cell fate, and expression measurements for a set of 227 genes. These values were derived from confocal microscopy movies of a developing *C. elegans* embryo (Figure 2). In these movies, histones in the nucleus have been tagged with green fluorescent protein (GFP), while the genes of interest have been tagged with mCherry, a red fluorescent protein. Therefore, expression of a gene is indicated by the presence of red fluorescence.

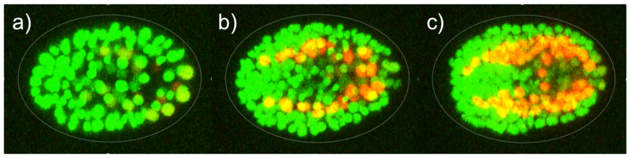


Figure 2. Expression of gene *hnd-1,* a muscle-specific transcription factor, at a) 0 minutes, b) 100 minutes, and c) 200 minutes. Notice the bilateral patterning of expression.

Data from EPIC was reformatted. Expression values in the EPIC dataset were expressed as relative fluorescence units, and because these values are hard to compare across experiments (due to experimental variability), values were normalized and binarized to indicate whether a gene was expressed or not. Also, due to limitations of the data collection methods the data sets in the visualization do not contain information about when genes turn off. Thus, after a gene is indicated as expressed this value is propagated to that cell's progeny throughout the remainder of the time series.

3D Plot and Small Multiples

Most of our understanding of cell-cell connections and spatial relationships in development comes from two-dimensional representations, either views through a microscope or in other representations like the lineage tree. While these representations can be very effective, we sought to address an important limitation with our 3D approach. Embryogenesis is a biological process that takes place in three dimensions, and the orientations and connections among cells play an essential role in this process. Being able to identify cells and to watch them undergo divisions and migrations in three dimensions, from any orientation, can greatly facilitate our understanding of which lineages are close together and which cell-cell connections might be important in forming different tissues.

To that end, we used the HTML library x3dom to build a three-dimensional plot of all cells. Cells were represented at each point in time as a row in an array which included its cell name, three-dimensional coordinates, size, and binarized gene expression values.

Lineage Tree

We also planned to include a lineage tree, as this graphic is widely used in the *C. elegans* community and provides context for cell origins and fate. By showing the lineage tree the user can quickly see how cells are lineally related to each other and can see how related cells are positioned in the 3D view. A collapsible tree diagram was implemented in d3 to represent the lineage tree. To image the entirety of the tree while providing detail, we used Cartesian distortion (https://github.com/d3/d3-plugins/tree/master/fisheye).

Gene Expression Plots

We sought to give the user the ability to assess global expression changes to find interesting patterns, and then provide ways to access more detailed information about the genes involved in these patterns of interest. One challenge was dealing with the scale of the data set. At the final time point in the series there are almost 550 cells, and we must display expression information for 227 genes, which gives on the order of 120,000 data points to

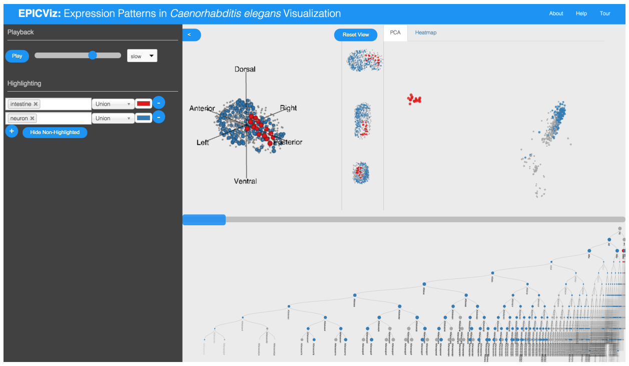


Figure 3. Overview of EPICViz.

show. We decided to use principle components analysis (PCA; performed in R) to cluster the cells by expression pattern and provide a high level summary of the patterns. To allow users to further investigate a cluster of interest, we provide an additional view showing the gene expression as a matrix with genes on the rows and cells on the columns. Any gene that is expressed in a particular cell is represented by a filled-in cell in the matrix. Thus users can see which genes are driving the PCA clustering.

RESULTS

Overview

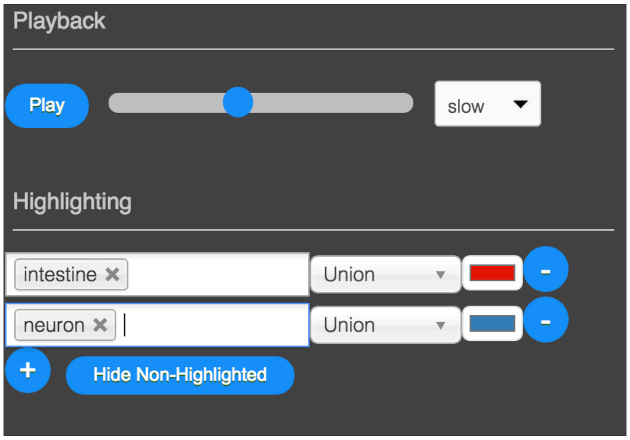
Our final visualization, EPICViz, consists of a user selection menu, 3D plot of the embryo, 2D small multiples of the embryo, a lineage tree, a PCA plot of gene expression, and a heatmap showing expression for all cells and genes (Figure 3). We describe each element in detail below.

User Selection Menu

In a collapsible menu on the left hand side of our visualization, users will see options for playback and highlighting. Under playback, users can play and pause the time slider at specific time points, as well as select the speed at which the visualization will animate.

Under highlighting, users can select lineages, cell types, and tissue types of interest by typing or scrolling through a drop-down menu. Cells within these categories will then be highlighted in the color of the user's choice. Multiple highlights can be implemented, and the user can also specify whether he or she wants the intersection or union of two selections. By default, cells that are not highlighted will appear as smaller grey spheres, but the user also has the choice to hide all non-highlighted cells.

As an example of how EPICViz can be used to understand gene expression and development, our figures will include highlights of intestine (in red) and neurons (in blue) (Figure 4).



intestine ✖ | Union ▾ | 🟥 | -

neuron ✖ | Union ▾ | 🟦 | -

Figure 4. User selection controls.

3D Plot

The 3D plot shows cells as spheres, which divide and migrate as development progresses (Figure 5, panel A). The user can drag the 3D plot to see it from a variety of views. Mousing over a cell reveals the cell name and its coordinates in a pop-up. Clicking on a cell emphasizes it with a yellow outline and the representations of that cell in all of the other plots are similarly highlighted, allowing the user to quickly reference corresponding information in the different facets of the visualization.

In our example, intestinal cells are clustering towards the posterior pole of the embryo and are mostly internal. Neurons, on the other hand, are distributed more widely across the embryo and are concentrated on the exterior of the embryo.

2D Small Multiples

The 2D small multiples lie to the right of the 3D (Figure 5, panel B). These are linked to the 3D plot both in terms of time and in terms of brushing. The three views are dorsal-posterior, anterior-right, and dorsal-right. Clicking on a small multiple moves the 3D plot to that view.

The small multiples show again the wider distribution of neurons on the exterior of the embryo in blue and the internal clustering of the intestinal cells towards the posterior in red.

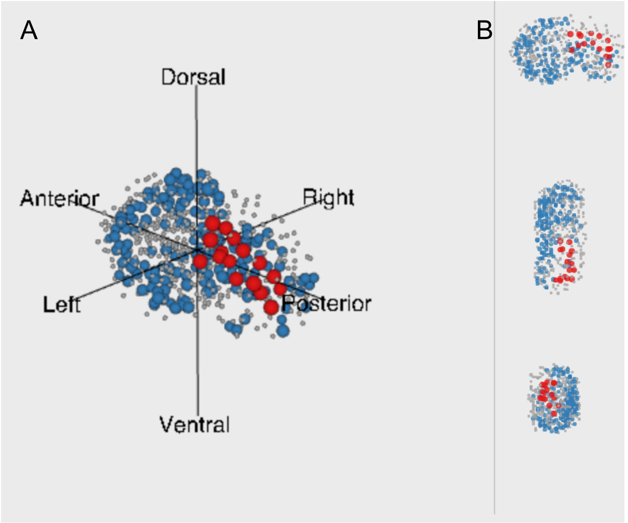


Figure 5. A) 3D plot and B) 2D small multiples. Intestine cells highlighted in red, neuron cells in blue.

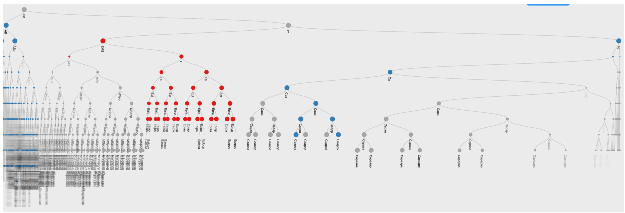


Figure 6. Lineage tree with intestinal cells in red, neurons in blue.

Lineage Tree

Cells within the lineage tree are populated as developmental time progresses. Highlights from the user selection menu are also linked to the tree, so cells of interest are easy to discriminate (Figure 6). Cells at the leaves of the lineage tree can be selected, resulting in a yellow highlight around the cell. As with the 3D view, these yellow highlights are propagated to all other plots to allow the user to quickly find their cell of interest throughout the visualization.

From the lineage tree, we can see how closely related intestinal cells and neurons are. Intestinal cells are descended from the EMS cell, which shares a parent with both P2-descended and AB-descended neuron cells.

PCA Plot of Gene Expression

Because our heatmap has values for approximately 500 cells and 227 genes at the end of the development time frame, we wanted a way to display gene expression information with reduced dimensionality. For this, we performed PCA for each cell and its 227 gene expression values.

Each dot within the PCA represents one cell, and users can watch as cells cluster into specific areas with developmental time (Figure 7). If a user clicks on a cell with a particular highlight (as defined in the user selection menu), a gene report pop-up is generated, which lists the gene name, WormBase ID (which links out to WormBase, a *C. elegans* database which has more detailed information on each gene), the fraction of cells in the selection expressing that gene, and a p-value to indicate how specific that gene's expression pattern is for the selection. P-values are calculated using a hypergeometric test on the number of selected cells expressing a particular gene versus the number of cells in the entire population expressing that gene and are corrected for multiple hypothesis testing using a Bonferroni correction.

Intestinal cells form a distinct cluster in the PCA relative to the more widely distributed neurons. Intestinal cells form part of the endoderm, the earliest evolving cell layer, which has well-conserved gene expression patterns, as seen here.

Heatmap of Gene Expression

On a tab next to the PCA plot, users can select to see a matrix of gene expression patterns (Figure 8). Every row in the heatmap is a gene, and every column is a cell. Like in the PCA plot, users can select colored highlights of interest, especially those exhibiting differential expression relative to other cells, and generate a gene report showing those genes that are most enriched among that cell selection.

Intestinal cells and neurons express many of the same genes, but there are some rows that are unique to each. This is seen by clicking on each tissue type's highlights and comparing the resulting gene reports (Figure 9).

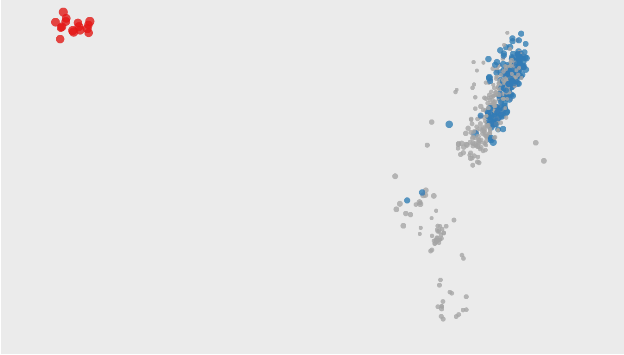


Figure 7. PCA plot of gene expression values. Intestinal cells (red) cluster differentially relative to neurons (blue).

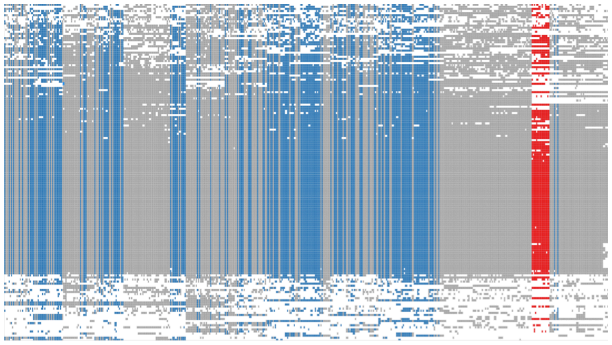
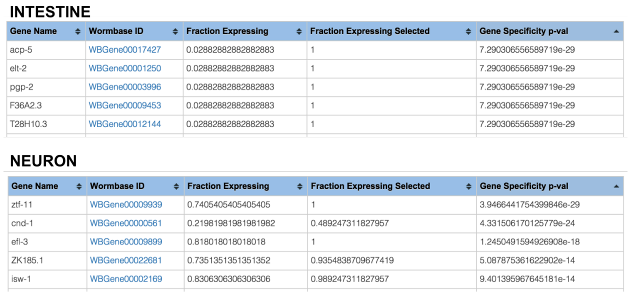


Figure 8. Heatmap of gene expression. Each row is a gene, each column is a cell. A filled square indicates expression. Intestines (red) and neurons (blue) highlighted.



INTESTINE

Gene Name	Wormbase ID	Fraction Expressing	Fraction Expressing Selected	Gene Specificity p-val
acp-5	WBGene00017427	0.02882882882882883	1	7.290306556589719e-29
elt-2	WBGene00001250	0.02882882882882883	1	7.290306556589719e-29
pgp-2	WBGene00003996	0.02882882882882883	1	7.290306556589719e-29
F36A2.3	WBGene00009453	0.02882882882882883	1	7.290306556589719e-29
T28H10.3	WBGene00012144	0.02882882882882883	1	7.290306556589719e-29

NEURON

Gene Name	Wormbase ID	Fraction Expressing	Fraction Expressing Selected	Gene Specificity p-val
ztf-11	WBGene00009939	0.7405405405405405	1	3.9466441754399846e-29
cnd-1	WBGene00000561	0.21981981981981982	0.489247311827957	4.331506170125779e-24
efl-3	WBGene00009899	0.818018018018018	1	1.2450491594926908e-18
ZK185.1	WBGene00022681	0.7351351351351352	0.9354838709677419	5.087875361622902e-14
isw-1	WBGene00002169	0.8306306306306306	0.989247311827957	9.401395967645181e-14

Figure 9. Gene reports with top five highest enriched genes for intestine and neuron.

DISCUSSION

We envision EPICViz being used by the *C. elegans* community to explore gene expression patterns in cells, tissues, and lineages of interest. While computational analysis of a subset of the EPIC dataset has been performed (as in [4]), many potential users of this data lack the computational background to carry out such sophisticated analysis. In contrast, EPICViz can be used by anyone, regardless of computational experience. A user can select a cell, tissue, or lineage of interest, see where the selection lies in the developing embryo, and interrogate what genes are enriched in that population. This allows for straightforward identification of gene expression patterns, much simpler than watching the microscopy movie for each gene and trying to compare expression across samples.

In addition to identifying gene expression patterns among user selections, EPICViz can serve as a tool to generate hypotheses about genes and their functions. For those genes in the dataset that are poorly characterized, EPICViz can provide better data about where and when these genes are expressed during embryogenesis.

FUTURE WORK

We are currently outlining ways to improve user selection, interface, and data management in EPICViz. With this version, users can select cell populations of interest by cell/tissue type and lineage to brush and link throughout all plots, but it is not possible to select a gene of interest. To facilitate gene-guided exploration of the data, we are planning to add another drop-down menu in the user selection menu so that users can specify a gene (or genes) of interest. In addition, at this point in time the screen is divided between all three major plots relatively equally; it is not possible for a user to resize a plot within the window. We are working to allow the user to expand or contract each element of the visualization so that he or she can focus on the plot(s) they find most informative. Finally, data loading and rendering performance at this point in time is sub-optimal; we are working on building efficient data structures that will quickly load and render.

CONCLUSION

We have developed EPICViz, an interactive visualization of *C. elegans* embryogenesis that allows users to explore gene expression patterns in cells of interest. This tool facilitates exploration of development and gene regulation, especially for users who lack the computational skills to analyze raw expression data.

REFERENCES

[1] J. E. Sulston, E. Schierenberg, J. G. White, and J. N. Thomson, “The embryonic cell lineage of the nematode Caenorhabditis elegans,” *Dev. Biol.*, vol. 100, no. 1, pp. 64–119, Nov. 1983.

[2] F. Long, H. Peng, X. Liu, S. K. Kim, and E. Myers, “A 3D digital atlas of C. elegans and its application to single-cell analyses,” *Nat. Methods*, vol. 6, no. 9, pp. 667–672, Sep. 2009.

[3] Z. Bao, J. I. Murray, T. Boyle, S. L. Ooi, M. J. Sandel, and R. H. Waterston, “Automated cell lineage tracing in Caenorhabditis elegans,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 8, pp. 2707–2712, Feb. 2006.

[4] J. I. Murray, T. J. Boyle, E. Preston, D. Vafeados, B. Mericle, P. Weisdepp, Z. Zhao, Z. Bao, M. Boeck, and R. H. Waterston, “Multidimensional regulation of gene expression in the C. elegans embryo,” *Genome Res.*, vol. 22, no. 7, pp. 1282–1294, Jul. 2012.

[5] S. K. Andrey Palyanov, “Towards a virtual C . elegans : A framework for simulation and visualization of the neuromuscular system in a 3D physical environment,” *In Silico Biol.*, vol. 11, pp. 137–147, 2012.

[6] O. Ronneberger, K. Liu, M. Rath, D. Rueβ, T. Mueller, H. Skibbe, B. Drayer, T. Schmidt, A. Filippi, R. Nitschke, T. Brox, H. Burkhardt, and W. Driever, “ViBE-Z: a framework for 3D virtual colocalization analysis in zebrafish larval brains,” *Nat. Methods*, vol. 9, no. 7, pp. 735–742, Jul. 2012.

[7] M. Meyer, T. Munzner, A. DePace, and H. Pfister, “MulteeSum: A Tool for Comparative Spatial and Temporal Gene Expression Data,” *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 908–917, Nov. 2010.