# Trajectory Analysis in Zebrafish Embryo Development

Author: Derek Fulton

Supervisor: Leila Muresan

7 August, 2019

Word Count: 5,350

# Declaration

I hereby declare that this dissertation entitled *Trajectory Analysis in Zebrafish Embryo Development* is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of this dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I confirm that I have read and understood the Faculty of Mathematics Guidelines on Plagiarism and the University-wide Statement on Plagiarism.

Derek Fulton

7 August, 2019

# Contents

# 1 Introduction

## 1.1 Gastrulation

During animal development, the embryo transforms from a single fertilized cell into a fully-functional organism (assuming all goes according to plan). Gastrulation, a critical early step in development, is the separation of the single-layered embryo (called the blastula) into three layers (called the endoderm, mesoderm and ectoderm). In some organisms (dipoblastic organisms), gastrulation only separates the blastopore into two layers [1]. The zebrafish, however, is not dipoblastic. It separates into endoderm, mesoderm and ectoderm during gastrulation like most animals. The tissue under investigation in this report is the zebrafish tailbud, which belongs to a specific part of the mesoderm called the paraxial mesoderm.

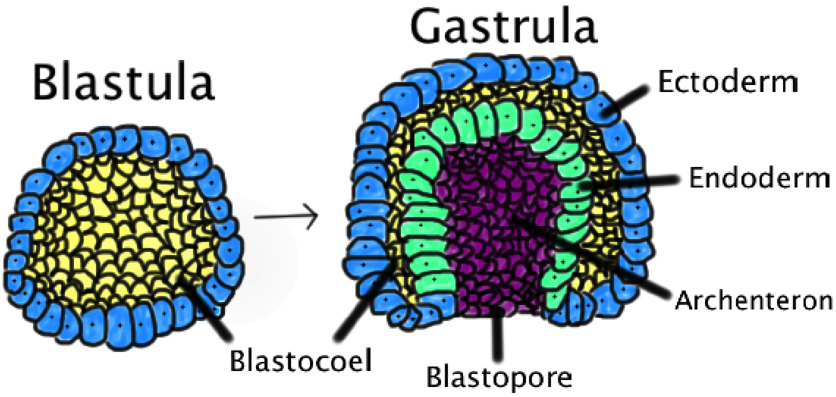


Figure 1: Diagram of the blastula and gastrula with the blastocoel labeled in both. The arrow represents the process of gastrulation, which transforms the blastula on the left (which contains only one layer, blue) into the gastrula on the right (which contains three layers: the ectoderm, endoderm and mesoderm).

Gastrulation is a critical step for development as it marks the beginning of cell specificity into distinct cell lineages. The eventual functions of these subsequent cell lineages go on to underpin all of animal life. The zebrafish is an excellent model organism for the study of gastrulation because it has translucent skin and a short development time (on the order of three days) [6]. This short development time allows for imaging of the embryo at all stages of development and quick iteration times for experiments, a benefit usually associated with other model organisms such as *Saccharomyces Cerevisiae* or *Escherichia Coli*, neither of which are animals. Hence, the zebrafish is an excellent animal model organism for the study of development.

## 1.2 Zebrafish Tailbud and Somites

The zebrafish tailbud exists at the most posterior end of the embryo, and extends in a posterior direction as the embryo develops. Though tailbud formation is not considered a core feature of gastrulation, the movement of the cells in the zebrafish tailbud is highly reminiscent of many movements seen during gastrulation [5]. For this reason, the study of the tailbud may provide insight into the underlying rules governing cell movement in gastrulation. The following quote illustrates the interdependence of the many constituent movements of gastrulation:

> "Although gastrulation may be conveniently divided into a number of particular movements for convenience of analysis, it is essentially a phenomenon of the whole. Each movement depends directly and indirectly on every other. Its cardinal feature is integration. For this reason, it is the process par excellence in which it will ultimately be necessary to understand each movement in relation to the others, in order to have a really meaningful comprehension of each one separately."
>
> — John Philip Trinkaus, 1969

Somites are the primitive segments in the developing embryo which later give rise to vertebrae. Somitogenesis is the process by which somites are formed in the developing embryo. As the tailbud extends, it leaves somites in its wake. Therefore tailbud extension is not only useful for the study of gastrulation, it is also useful for the study of somitogenesis. The growth dynamics of the somites which give rise to the spinal cord are highly conserved, making the study of the budding relevant to all vertebrate life[2].

We next turn to a primer on the imaging modality which makes this investigation possible.

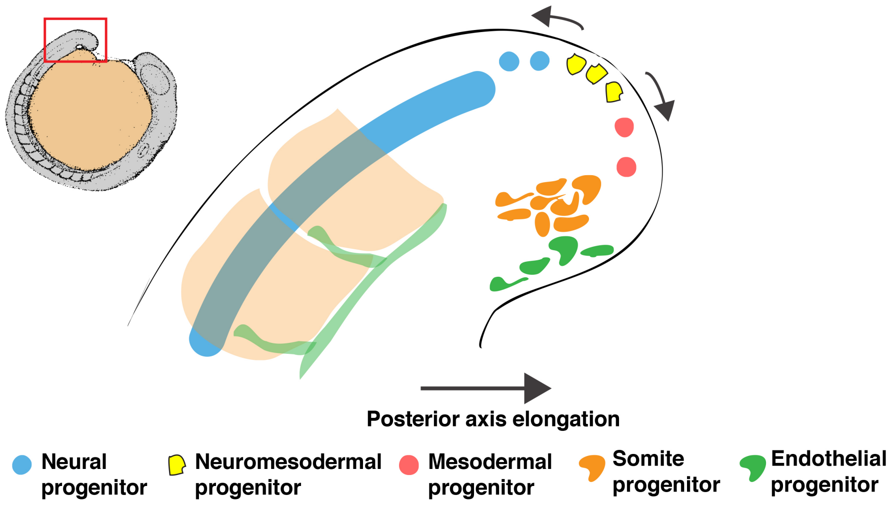


Figure 2: A diagram of the zebrafish tailbud with the various constituent groups of cells labeled. The progenitor cells are on the posterior as they organize themselves to give rise to the neuromesoderm (yellow), mesoderm (red), somites (orange) and endothelium (green). This diagram closely corresponds with the region imaged in this investigation.

## 1.3 Light-sheet microscopy

Light sheet microscopy is a type of fluorescence light microscopy that offers all the benefits of fluorescence microscopy (such as high resolution and the ability to keep the sample alive) while minimizing the pernicious photobleaching effect by illuminating only one ”sheet” (one $xy$ plane at a certain $z$-index) of the sample at a time (Figure 3)[3]. For this reason, it is ideal for the study of the zebrafish embryo, which requires roughly 72 hours from fertilization to complete development into a larva, which is usually too long to safely avoid photobleaching. The raw data used in this investigation is a series of three-dimensional time-lapse images from light sheet microscopy, which illuminates the sample in a direction parallel to the objective.

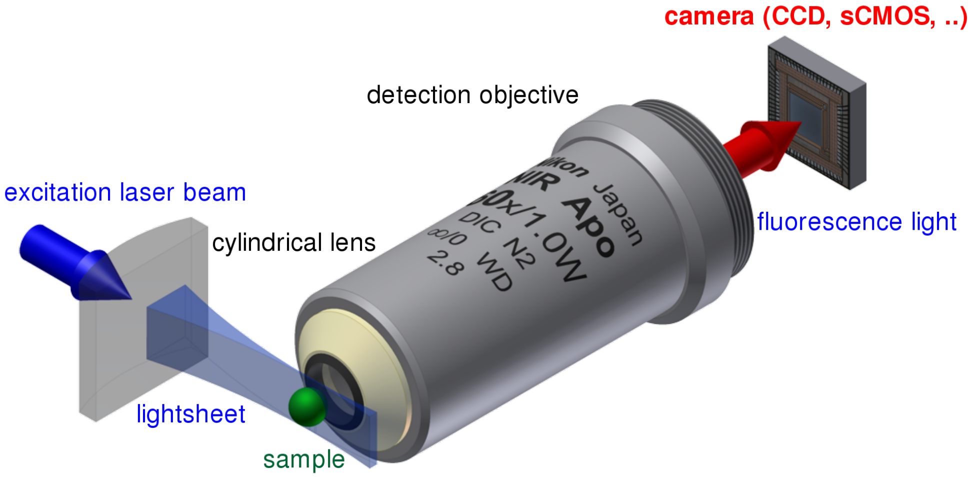


Figure 3: Diagram of light-sheet microscopy. The sample is illuminated with ”sheets” of light (translucent blue region on the left) directed orthogonally to the objective. Because a sheet of light only illuminates a small layer of the sample, the non-illuminated part of the sample is safe from photobleaching. Light-sheet microscopy allows for high-resolution fluorescence images over a longer timespan than traditional fluorescence microscopy.

### 1.4 Imaris Tracking Data

Imaris is a a software package from Oxford Instruments Group, designed for the visualization and analysis of three-dimensional microscopy data [4]. It offers built-in tracking which, from the raw data consisting of light intensity values at each $xyz$ coordinate in the sample, and for each timestep in the sample, outputs the $xyz$ coordinates (in $\mu$m) of each center of each well-defined cell based on fluorescently-labeled nuclei. It also assigns each tracked cell a ”Track ID”, which allows for the analysis of not only the point cloud of cells in each frame, but also the trajectories of the cells over time. Thus, each row in the Imaris output represents one tracked cell for one timestep, illustrated in Table 1, below:

| **x** | **y** | **z** | **t** | **TrackID** |
|---|---|---|---|---|

Table 1: One row of the output from Imaris tracking software, excluding columns irrelevant to this investigation.

Two embryos are analyzed in this investigation. **Embryo A** was imaged once per minute for 165 minutes. **Embryo B** was imaged once every four minutes for 80 minutes. We shall see later that **Embryo A** (one minute resolution) is easier to analyze due to the nature of the tracking software, however **Embryo B** is sometimes used to confirm results from **Embryo A** where appropriate.

Thus, we have two datasets of tracked cells, one from each embryo. Recall that the temporal resolution is one minute for Embryo A and four minutes for embryo B.

### 1.5 Online Videos

Development occurs over time (in the case of the zebrafish, about 72 hours). The data used in this investigation is time-lapse microscopy and trajectory data. Due to the time-dependent nature of the biological process under investigation and the resulting data, it is often far more informative to view video plots than still frames. Still frames are pictured in this investigation (as Figures 8-12) and the captions help provide an idea of the time-dependence of the data. The movies which complement Figures 8-12 in this report can be found at **ddfulton.github.io**.

## 2 My Work

### 2.1 Track Quality Analysis

Because Imaris is not perfect and downstream analysis relies on consistent tracks, it is useful to quantitatively analyze tracking quality before doing anything else. This way, we can capitalize on the highest quality sections of data while avoiding any artefacts which may arise from the lowest quality sections of data.

#### 2.1.1 Track Count

First, we ask the trivial question: how many cells were successfully tracked in each frame, regardless of whether the cells appear in the previous or subsequent frame? Figure 4 shows this information.

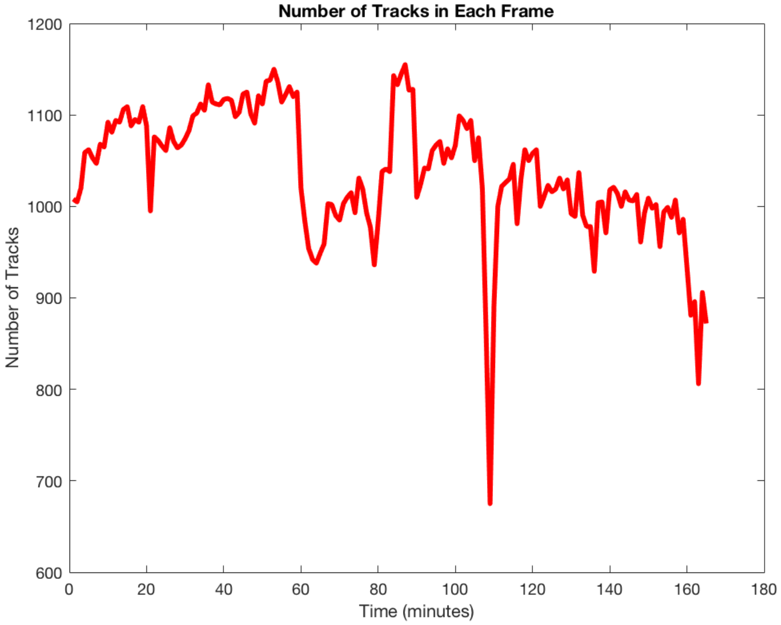


Figure 4: Number of successfully tracked cells in the data over time. Notice the large drop in number at 108 minutes. This plot pays no attention to whether a cell was tracked for multiple frames, but shows only the number of cells tracked at each individual frame.

In Figure 4 we see that there are roughly one thousand tracked cells per frame. This is in rough agreement with the raw image data (and is roughly the expected number of cells in the tailbud region of the embryo), and exposes no glaring flaws in the data.

However, notice the substantial dips in track count at roughly $t = 61$ minutes and especially at $t = 108$ minutes due to a snapping event of the embryo, where a large chunk of the embryo suddenly straightens out in unison (clearly visible in the movies). This is useful to bear in mind as it will help us contextualize results from later sections of this investigation. We will see later that for much analysis it is best to ignore these dips and focus on the higher-quality frames of $t$ between 1 and 45.

While it is good that we have over 1000 tracks for many individual frames, Figure 4 says nothing about continuity of tracks over time. What we are really interested in is the behavior of cells as they move during tailbud development over longer time periods, preferably an hour or more, forming long trajectories. It is not useful to know where a cell is for just one frame, because then there is no trajectory to analyze.

### 2.1.2 Track Lifespan

The next question we next ask: at each frame, how many tracked cells are there that also existed in the previous frame? Though a track lifespan (number of consecutive frames in which a track exists) of two is hardly more helpful than a track lifespan of one, the results shown in Figure 5 confirm the results shown in Figure 4 and affirm our idea of which frames have the highest-quality data for downstream analysis.

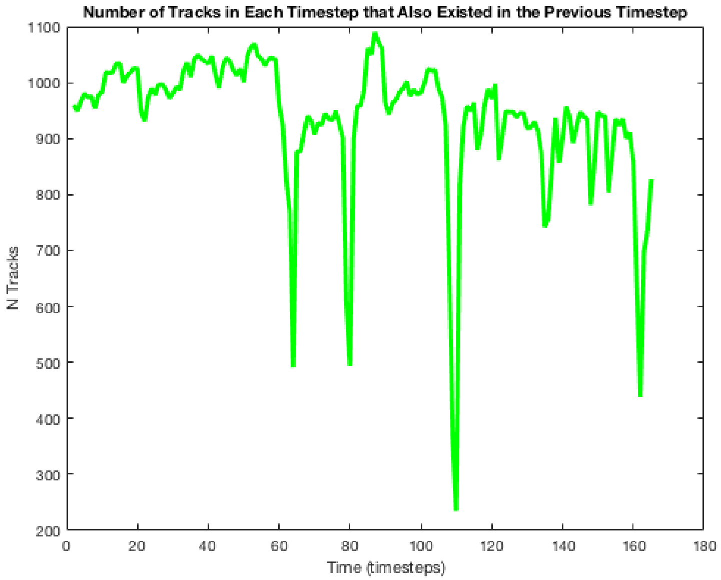


Figure 5: Number of tracked cells in each frame that also existed in the previous frame only. Similar to Figure 4 but excludes cells which were tracked for only one frame.

In Figure 5, we again see the same dips at roughly $t = 61$ and $t = 108$, but they are now more severe than those in Figure 4. We also see some new dips throughout the track, some nearly as severe as those at $t = 61$ and $t = 108$. However, there are still approximately one thousand tracks which now display some continuity before $t = 60$. The longest uninterrupted segment is roughly $t = 1$ to $t = 60$ minutes, which will be the main focus of section 2.3.

## 2.2 Registration

In visualizing and measuring cell movement it is not always obvious what a cell (or group of cells) is moving relative to, especially in light of the Trinkaus quote above (section 1.2). Since we are studying the tailbud formation in isolation on the posterior axis, it's often nontrivial to determine whether the tailbud cells are moving in a posterior direction or whether the embryo as a whole is elongating and the tailbud cells are more stationary (relative to the center of the embryo). We are most interested in movement of tailbud cells within the embryo, and how this movement gives rise to development of the entire organism.

In an effort to isolate the tailbud for visualization, one solution is to manually impose a reference frame at the tip of the tailbud in the images (which is a feature offered by Imaris). This way, the posterior-most tip of the tailbud is always at $(0, 0, 0)$ in $xyz$ space. For study of the tailbud in isolation, this makes intuitive sense. However, manually setting the reference frame at each timestep is not only tedious, but imprecise. These small imprecise displacements are inflicted upon entire frames which might compromise the fidelity of any downstream trajectory analysis sensitive to small displacements of a couple microns.

One way to correct any small inaccuracies incurred from manual registration is to register the tracks to themselves by minimizing the distance between all tracks that exist in two adjacent frames. In other words the point-cloud consisting of all the $(x, y, z)$ coordinates cells in frame $t$ will be registered to the point-cloud from the frame at $t - 1$ based on all the tracks that exist in both frames. Use of this method ensures that each cell's trajectory will be dominated by the cell's intrinsic motion and not affected by any shaky global motions of the entire embryo. Further, as the tailbud changes very little each minute, the general behavior of the embryo in the video will remain "smooth." Any cells displaying biologically relevant motion will now be easier to identify because the visualization of the embryo is smoother.

Take, for example, two frames, $S_1$ and $S_2$ (from $t_1$ and $t_2$). To register frame $S_2$ to frame $S_1$, three translations (along the $x$, $y$ and $z$ axes) and two rotations (about the $x$ and $y$ axes) are performed on frame $S_2$. We do not need rotation about the $z$-axis because the final result of any rotation about the $z$-axis can be expressed as one rotation about each of the other two axes. The combination of these translations and rotations allows the embryo to be shifted into any position and orientation, and therefore allows the closest registration possible. No stretching or shrinking were ever performed, only translation and rotation.

To assess quality of registration we employ the sum of squared differences as a "loss" function. Equation (1) shows the loss function we want to minimize (using the *optim* function from **R**) in the registration of one frame, $S_2$, to the prior frame, $S_1$.

$$\mathcal{L}(dx, dy, dz, \theta_x, \theta_y) = \sum_{i=1}^{N} (\hat{S}_2^i - S_1^i)^2 \tag{1}$$

Where each frame is a set of $N$ points where $\hat{S}_2^i$ is the translated and rotated cell $i$ from frame $S_2$ and $S_1^i$ is the original cell $i$ from frame $S_1$. The translations are performed with simple arithmetic, $\hat{S}_2^i = S_2^i + < dx, dy, dz >$, and the rotations are performed using matrix multiplication between the data matrix $S_2$ and the relevant three-dimensional rotation matrices $\mathbf{R}_x$ and $\mathbf{R}_y$ to rotate about the $x$

and $y$ axes, respectively,

$$\hat{S}_2^i = \mathbf{R}_y \mathbf{R}_x (S_2^i + < dx, dy, dz >), \tag{2}$$

where $\mathbf{R}_x$ and $\mathbf{R}_y$ are constructed using $\theta_x$ and $\theta_y$:

$$\mathbf{R}_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta_x) & -\sin(\theta_x) \\ 0 & \sin(\theta_x) & \cos(\theta_x) \end{pmatrix} \tag{3}$$

$$\mathbf{R}_y = \begin{pmatrix} \cos(\theta_y) & 0 & \sin(\theta_y) \\ 0 & 1 & 0 \\ -\sin(\theta_y) & 0 & \cos(\theta_y) \end{pmatrix} \tag{4}$$

## 2.3 Clustering

### 2.3.1 Hierarchical Clustering Overview

Hierarchical clustering is a method of clustering data points based on their respective distances from each other [7]. It works in multiple dimensions and, for this investigation, utilizes the Euclidean distance. There are two types of hierarchical clustering:

- Agglomerative (bottom-up)
- Divisive (top-down)

The clustering in this investigation is agglomerative (using the **hclust** function from the R programming language). Each data point starts in its very own cluster (one leaf on bottom of dendrogram in Figure 6) and is merged with other data points as the minimum required distance is inleavescreased.

Hierarchical clustering is often used in biology when constructing phylogenetic trees and cell fate trees. In phylogenetic trees, organisms are often clustered based on the similarity between conserved genes. In embryology cells are often clustered based on a visual biomarker such as a organ scar fractions [2].

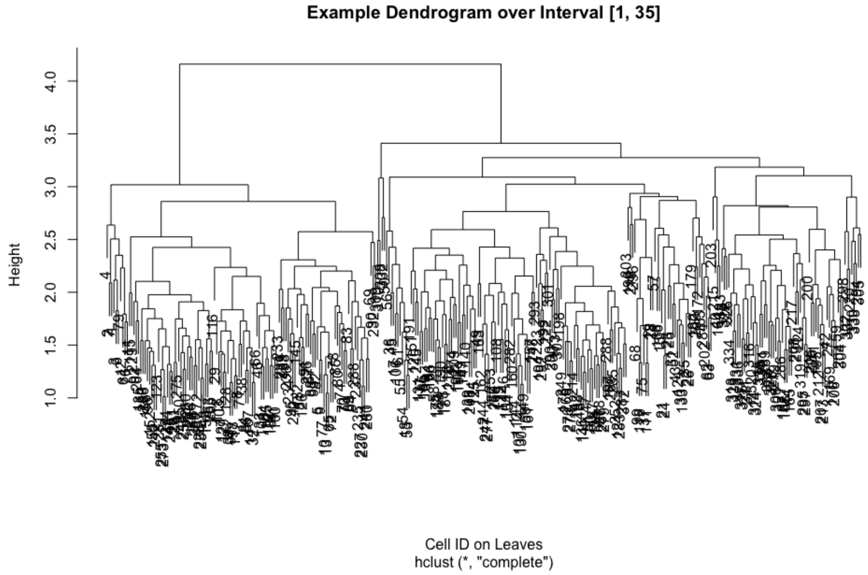


Figure 6: Dendrogram output from agglomerative hierarchical clustering on all cells in the first 35 frames of the data. As the height decreases (by moving down on the vertical axis) the number clusters increases. The two biggest clusters (from cutting the tree at a height of 3.5) will later be shown to represent the anterior and posterior ends of the embryo.

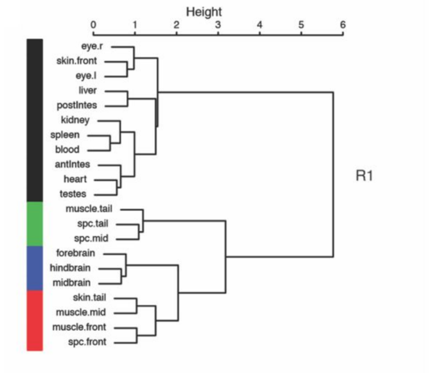


Figure 7: Hierarchical clustering dendrogram of cell fate in the zebrafish embryo based on organ scar similarity, a genetic marker. The mathematics of the clustering performed in this investigation is equivalent (as it is also agglomerative clustering). However the data clustered is different as it is based only on simple Euclidean cell trajectory.

Hierarchical clustering allows us to choose how many clusters we would like to see, depending on at which height the dendrogram is "cut." For example, cutting the dendrogram of Figure 6 at a height of 3.75 will yield two clusters which are the left cluster and the right cluster in the dendrogram. In the supplementary videos, the number of clusters is between 2 and 6 and is always displayed in the title and filename of each video.

Hierarchical clustering is useful in this study because the number of clusters can be manipulated to fit the problem. If we were attempting to cluster two distinct tissues, for example ectoderm and mesoderm, we would cut the dendrogram such that it produces two clusters. In this problem we perform hierarchical clustering solely based on trajectories of well-tracked cells and are agnostic to what we find. A good scenario would be finding many clusters, all of which appear to correlate strongly to a certain kind of embryonic tissue already discovered. This would link the cellular identities to trajectory data, which would provide a useful new quantitative perspective. Unfortunately, this is not the case so we are open to any number of clusters provided they are visually recognizable as clusters.

### 2.3.2 Input Data for Hierarchical Clustering

In this investigation, one data point is the trajectory of a cell over a predetermined number of frames. One single change in position for one cell $i$ is a 3-vector containing the $x$, $y$ and $z$ displacements of that cell between frames $S_t$ and $S_{t-1}$:

$$\Delta X_t^i = S_t^i - S_{t-1}^i \tag{5}$$

If clustering is performed over only the first 2 frames, $S_1$ and $S_2$ then there is only 1 displacement per each of the $N$ cells $(S_2 - S_1)$, so the data matrix will have 3 columns (one for each dimension) and $N$ rows. If it is performed over 10 frames the data matrix will have $3 * (10 - 1) = 27$ columns. For $M$ frames, this generalizes to $3 * (M - 1)$ columns. Cells are only included if they have been successfully tracked in all frames used for clustering.

As we learned in section 2.1, few, if any, cells are tracked throughout the entire 165 frames. Though a handful are, there are nowhere near enough to make sense in the context of the entire tailbud. We must have enough cells to discern the general shape of the tailbud and especially to distinguish anterior from posterior, but we also must cluster over enough frames to scrutinize longer-term cell behaviors as they may relate to tailbud formation and gastrulation. While clustering over five minutes (five frames) yields hundreds of persistent cells, five minutes is a relatively short time in the context of the three-day development process of the zebrafish. This interplay between number of persistent cells and amount of time to be clustered is discussed in section 3.2.

Striking this balance between cell count and frame count leads us to primarily focus on the embryo before time $t = 45$ and after time $t = 140$ due to the substantial loss of tracks in the middle of the data (Figures 4 and 5). Even the frames after $t = 140$ display far fewer persistent cells than frames between $t = 1$ and $t = 45$.

To visualize the result of the clustering in a way other than the dendrogam, it is desirable to assign each cluster a color and scatter plot the tracked cells. This allows us to visualize the clusters as they evolve throughout the course of tailbud development, and to notice any distinct spatial features or patterns unique to one cluster. These plots are displayed in the remainder of section 2.3.

### 2.3.3 Anterior and Posterior Sides of Embryo Successfully Clustered

Watching the movies corresponding to the single frame shown in Figure 8, we see a clear separation between anterior and posterior (over twenty minutes of development). This makes sense as the anterior and posterior ends of the tailbud are known to exhibit different biological features and contain different cell types (Figure 2).

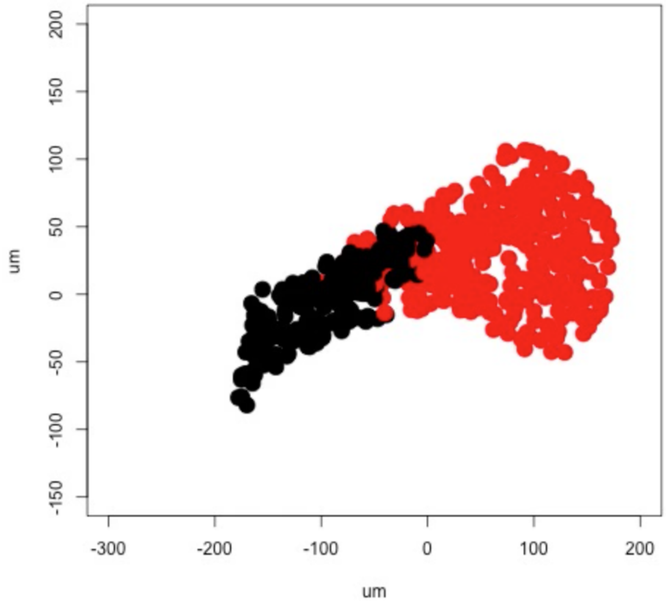


Figure 8: The two biggest clusters visualized in red and black. The black is the anterior cluster and the red is the posterior cluster. These anterior and posterior clusters correspond to the two biggest clusters in Figure 6 (the two clusterings resulting from cutting the dendrogram at a height of 3.75). The reader is encouraged to view the video on ddfulton.github.io

The posterior cells display more movement than the anterior cells because the multiple types of progenitor cells (shown in Figure 2) are still assembling themselves into tailbud. The anterior cells have formed into less motile structures (such as somites and neural progenitors).

### 2.3.4 Embryo B Confirms Section 2.3.3

Embryo B has only one fourth the temporal resolution of embryo A and therefore cannot be analyzed directly alongside embryo A for many analyses (as discussed in Section 1.4) . However, because the differences between the anterior and posterior sides of the embryo are so fundamental and well-captured by two-cluster clustering in embryo A, it is worthwhile to perform this particular analysis on embryo B to see whether the results match.

As embryo B (one image captured every four minutes) has worse temporal resolution than embryo A (one image captured every one minute), one might think that we should cluster a movie of the same length in minutes (not frames, recalling that embryo A has four times as many frames per minute than embryo B) from embryo B as we did in embryo A if we are to compare the two. However, these led to poor results and the number of frames seems to be more important in reproducing the results than the number of minutes the frames span. Hence we cluster over 20 frames for each embryo, knowing that this represents 20 minutes of development in embryo A but 80 minutes of development in embryo B.

Given this, we now cluster the first 20 frames (80 minutes) of embryo B and plot the two largest clusters, in a manner similar to Figure 7. This two-cluster visualization of embryo B is shown in Figure 9.

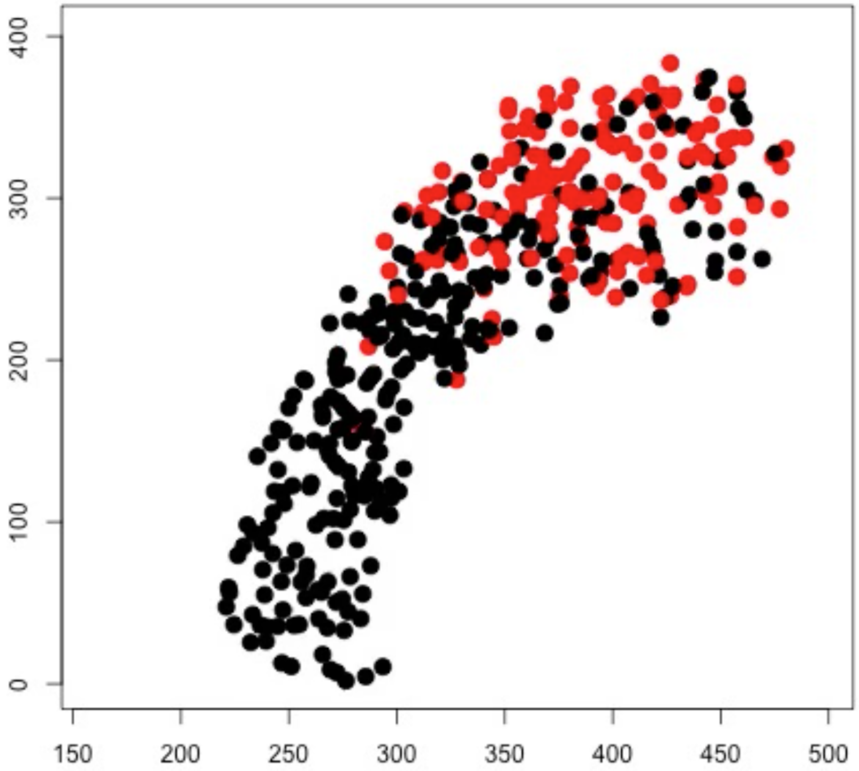
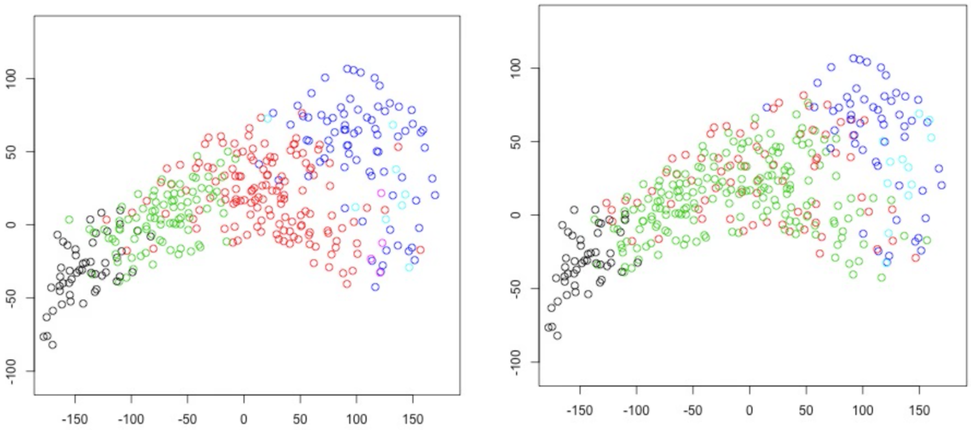


Figure 9: Two biggest clusters from embryo B. The shape of the embryo is slightly different but we see the same divide between anterior and posterior, confirming our results. The posterior is the top-right direction and the anterior is the bottom-left direction.

In Figure 9, we see a considerable number of black-clustered cells existing in the posterior end. While we don't see this in Figure 8, the red/black separation is clear as the red cluster dominates the posterior end and the black cluster dominates the anterior end. It is noteworthy that this similar clustering result persists over 80 minutes of development in embryo B and only 20 minutes of development in embryo A. This confirms our prediction that the difference between the trajectories of anterior cells and posterior cells, as captured by two-cluster agglomerative clustering, are fundamental enough to persist over different amounts of time of development.

### 2.3.5 Normalization of Displacements Has Small Effect

Previously, all clustering in this section has been performed on the three displacements (along the $x$, $y$ and $z$ components), where each entry is displacement in microns. The results presented in section 2.3.3 show a clear separation between anterior and posterior cells (Figure 7). It is quite a common practice among many clustering algorithms (including hierarchical clustering) to normalize the data before clustering. To investigate whether the normalization of data would have an effect on the clusters, the first 35 minutes from embryo A were clustered and plotted with and without normalization. Because normalization changes the magnitude of distances between two cells, different height cutoffs for each cluster were chosen so that, in both cases, the five biggest clusters are shown in colors. Figure 8 shows one frame of this difference. As always, the reader is urged to view the videos at **ddfulton.github.io** to get a clearer picture of the cluster behavior over time.



(a) Not normalized

(b) Normalized

Figure 10: Top 5 clusters of embryo A clustered over first 35 minutes after (a) no normalization of cell displacements and (b) normalization of cell displacements to the interval $(0, 1)$

## 2.4 Neighborhood Analysis

We now turn to one of the simpler and more famous analyses of spatial data: nearest neighbor analysis. Unlike hierarchical clustering, the displacements calculated in this section are only between sets of two frames, and will never require the same neighborhood, defined as the group of a given cell's $k$ nearest neighbors, to be successfully tracked in three or more adjacent frames.

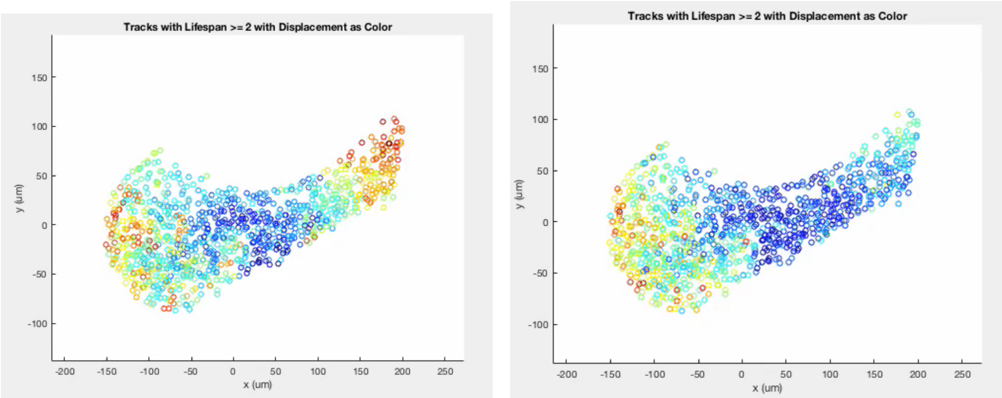
Analyzing the neighborhood of each cell will help us to visually identify special outlier neighborhoods which may be relevant to the underlying biology of gastrulation. Some neighborhoods may be more or less dense than others which may correspond to distinct regions of the embryo.

### 2.4.1 Average Neighborhood Displacement

First we ask: for each cell, what is the average displacement of its neighborhood? As the anterior-most cells represent the budding tailbud, and the posterior-most cells represent the more organized part of the embryo, we might predict the anterior-most region would exhibit the greatest average displacement of constituent neighborhoods as the cells display more motility while organizing. This is also visually apparent by simply examining the movie visually, however it is nice to confirm quantitatively. The equation for a cell's average neighborhood Euclidean displacement based on its $k$ nearest neighbors, between two frames $S_t^i$ and $S_{t-1}^i$, is

$$d\bar{S}_t^i = \frac{1}{k}\sum_{i=1}^{k} \|S_t^i - S_{t-1}^i\| \tag{6}$$

In the plots below (Figure 4, (a) and (b)), the color represents the average displacement of each tracked cell's neighborhood at two different neighborhood sizes (5 and 12). In order of increasing value the colors are: blue, cyan, green, yellow, orange, red (corresponding to MATLAB's colormap *jet*). Further, only tracks that exist for at least two consecutive frames are included.



(a) Neighborhood of size 5.

(b) Neighborhood of size 12.

Figure 11: Snapshot of neighborhood displacement movie based on neighborhood of size 5 (left) and size 12 (right). In these movies and images the anterior is right and the posterior (tailbud) is on the left. The budding tailbud is on the anterior (left).

It is essential to view the videos to get an idea of the embryo's behavior over time (though these snapshots are fairly representative).

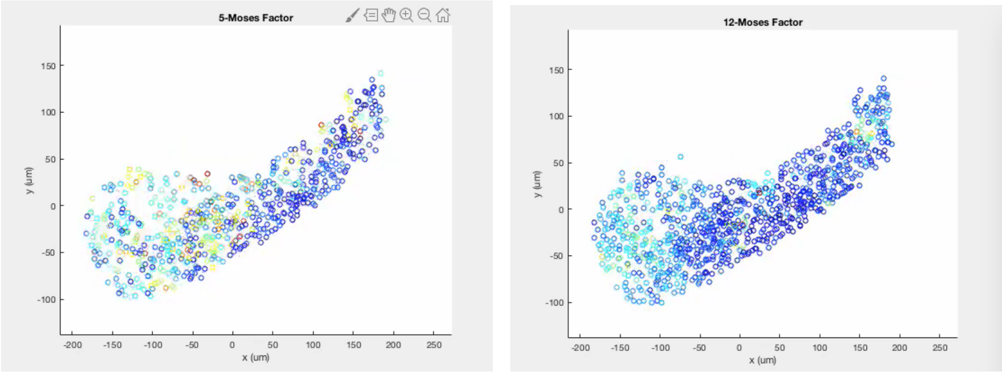
### 2.4.2 Divergent Cells

We next ask: do any cells stand out from their neighborhood? Cells tend to move in clusters, so it might be interesting to see if any cells show behavior highly divergent from the average behavior of their neighborhoods. The formula for a cell's divergence from its neighborhood is equivalent to the norm of the cell's displacement minus the average neighborhood displacement (from Equation 2)

$$\|dS_t^i - d\bar{S}_t^i\|, \tag{7}$$

where $dS_i$ is cell $i$'s displacement and $d\bar{S}_i$ is the average displacement of cell $i$'s neighborhood.

Optimistically, we will see highly divergent cells which signal some biological phenomenon of interest, for example one of the cell type identities shown in Figure 2. More realistically, the cell's divergence from its neighborhood will most likely indicate that some neighborhoods are simply more cohesive than others, as we saw in Figure 11. Figure 12 shows two snapshots of these movies, which are accessible at **ddfulton.github.io**.



(a) Neighborhood of size 5.

(b) Neighborhood of size 12.

Figure 12: Snapshot of divergence from neighborhood displacement movie based on neighborhoods of size 5 (left) and size 12 (right). In these movies and images the anterior is right and the posterior (tailbud) is on the left.

The frames shown in Figure 12 (and their corresponding movies) no obvious outlier cells which behave differently from their neighborhood due to any known biological phenomenon. However, as predicted, this continues to quantitatively confirm that there is greater cell motility and less cohesive neighborhoods in the posterior end.

# 3 Conclusion and Future Work

## 3.1 Conclusion

Though imaging data from embryo A is limited, there is about one hour of persistent tracks at the beginning of the data. A few things, none of which were revolutionary, were accomplished:

- Registration of adjacent frames to produce smoother video
- Hierarchical clustering based on each cell's trajectory which produced biologically verified results over two separate embryos
- Nearest neighbor analysis which further corroborated the assumption that cell neighborhoods of the posterior end of the embryo displays more cell motion, on average, than the anterior end

Though nothing new has been discovered, we have affirmed some basic assumptions about development of the zebrafish tailbud with quantitative methods. Further, we have done this based solely on trajectory data with no attention paid to genetic markers or raw image data. It is even more encouraging that the clustering worked, in one case, on a second embryo imaged at a very different temporal resolution. We have also critically analyzed the tracking quality of Imaris, a popular software for biological researchers

## 3.2 Future Work

### 3.2.1 Cell Tracking Quality

Perhaps the biggest limitation to this investigation was the quality of the tracking data. The longest period of time for which we were able to analyze the tracks of persistent cells was about 45 minutes. While this is not bad and zebrafish development is fast compared to many other vertebrates, the tailbud forms for many hours. Cell tracking and segmentation has been a persistent challenge in the field for many years, and is a well-known problem. In fact, there is even a running competition for cell-tracking published in Nature Methods [8].

However, cell tracking is a challenging task and it is limited by microscopy methods and computational resources, and, in a sense, the resolution limits imposed by wavelength of light (for *in vivo* light microscopy, obviously). Regardless, this investigation was performed anyway with the expectation that in several years, as cell tracking improves, the same analysis can be run on better tracks which span longer periods of zebrafish development. As zebrafish development lasts in the neighborhood of 72 hours, having persistent tracks for less than one hour was a limitation.

### 3.2.2 Analysis on Multiple Embryos

An interesting follow-up project would be a tool that automatically replicates this investigation on new embryos. However, as we have already seen embryos are not all imaged in the same way and at the same spatiotemporal resolutions. Using the known ground truths about development to standardize the embryo data would be quite useful toward the goal of applying the following analysis and accompanying code to new embryos. Trajectory interpolation would be one crude solution to the problem of differing temporal resolutions.

### 3.2.3 Clustering Based on Incomplete Tracking Data

The hierarchical clustering in section 2.3 successfully clustered the anterior and posterior of the tailbud in both embryos. However, one issue with this clustering method is that it requires all cells to exist in all frames under investigation.

As we have seen, the first 35 minutes (out of 165 available) provide enough persistent cells to produce clusters which look cohesive and not random. Although the number of persistent cells decreases later in the data and with number of frames clustered, the number of cells remains around 1000 throughout the data (in embryo A). To cluster their trajectories we need to make direct comparison between cells and this would be quite difficult to do with incomplete trajectory data. If there were a way to do this, the quality of the clustering might be increased and the visualization would be enriched due to the greater number of cells present.

### 3.2.4 Amazon Mechanical Turk

As we have seen, the limitations of the automated Imaris tracking data restrict our analysis. Manual tracking is often feared, however would provide much higher quality data. While manual tracking is tedious, expensive and not scalable, the high quality of the data would be useful (even if it was only generated from one embryo). Using a tablet, it is possible to track cells at a fairly rapid pace. For 165 frames, assuming 1000 cells per frame, tracking $165,000$ cells would take a person two weeks at the most. If we had this high-quality data, its results from our clustering analyses could be constantly compared to new Imaris tracking data to to verify results based on incomplete data. Although research labs do not tend to outsource work in this way, the overall cost would be trivially covered by the average research university grant size and is a possible avenue for future work in this field. Scalable methods that require far more data, however, must rely on automated tracking.

# References

[1] Steventon lab website.

[2] Andrea Attardi, Timothy Fulton, Maria Florescu, Gopi Shah, Leila Muresan, Jan Huisken, Alexander van Oudenaarden, and Benjamin Steventon. Neuromesodermal progenitors are a conserved source of spinal cord with divergent growth dynamics. *bioRxiv*, 2018.

[3] Tanner C Fadero, Therese M Gerbich, Kishan Rana, Aussie Suzuki, Matthew DiSalvo, Kristina N Schaefer, Jennifer K Heppert, Thomas C Boothby, Bob Goldstein, Mark Peifer, et al. Lite microscopy: Tilted light-sheet excitation of model organisms offers high resolution and low photobleaching. *J Cell Biol*, 217(5):1869–1882, 2018.

[4] Oxford Instruments. Imaris tracking software.

[5] John P Kanki and Robert K Ho. The development of the posterior body in zebrafish. *Development*, 124(4):881–893, 1997.

[6] Charles B Kimmel, William W Ballard, Seth R Kimmel, Bonnie Ullmann, and Thomas F Schilling. Stages of embryonic development of the zebrafish. *Developmental dynamics*, 203(3):253–310, 1995.

[7] Robin Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 16(1):30–34, 1973.

[8] Vladimír Ulman, Martin Maška, Klas EG Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al. An objective comparison of cell-tracking algorithms. *Nature methods*, 14(12):1141, 2017.