Biophysical Journal

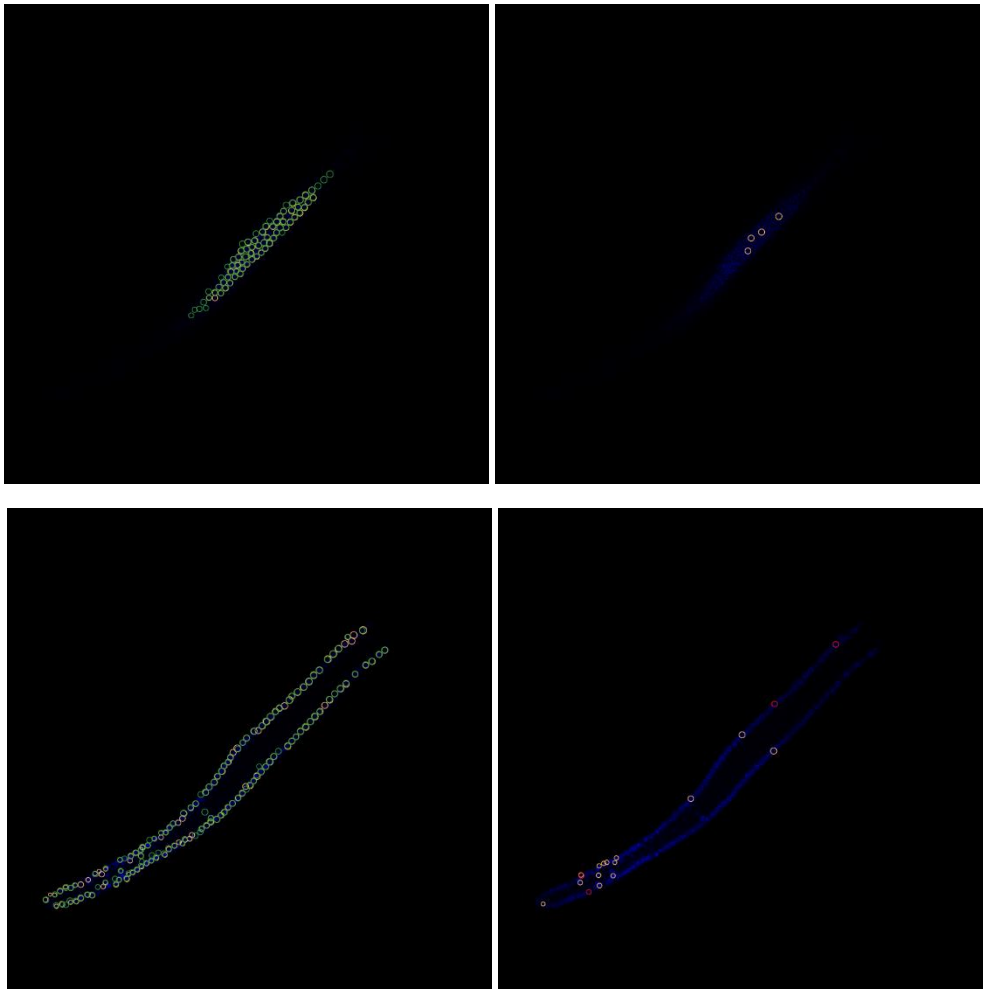
A Transport Model for Estimating the Time Course of ERK Activation in the *C. elegans* Germline

Henry H. Mattingly[1], Jessica J. Chen[2], Swathi Arur[2,*], Stanislav Y. Shvartsman[1,*]

[1]Department of Chemical and Biological Engineering and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, [2]UT Graduate School of Biomedical Sciences and Department of Genetics, University of Texas M.D. Anderson Cancer Center, Houston, TX

SUPPORTING MATERIAL

Section S1: Comparison of Manual and Automated Image Segmentation



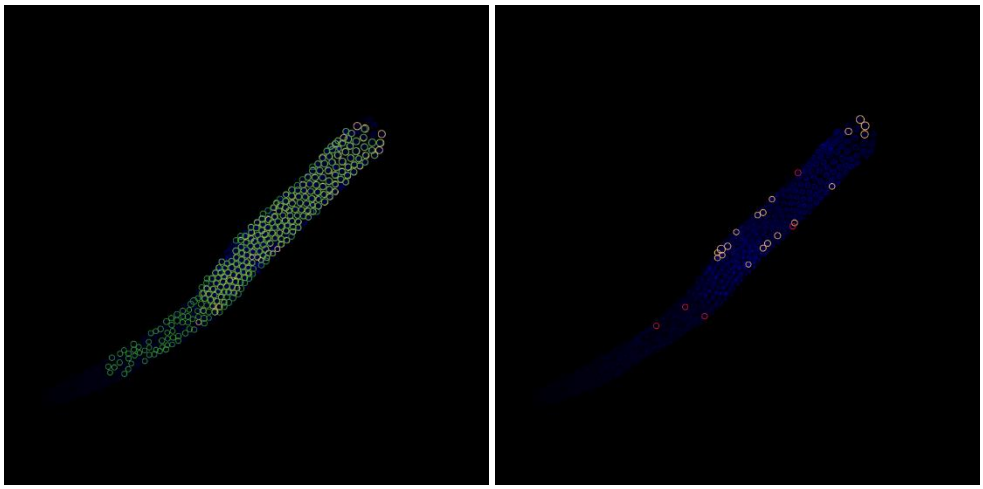


Figure S1: Detection of germ cells in z-stacks of germlines. Each row is an image of the same germline taken at a different depth. The distal tip is at the bottom-left of each image. The green circles in the left column are the cells that were correctly identified by the automatic segmentation algorithm, and the orange circles are the corresponding manually-segmented cells. Green circles without an orange partner correspond to cases in which the cell was first detected in that slice by the automatic segmentation algorithm, but first detected in the slice above or below that one by the manual segmentation. The red circles in the right column are the objects that were incorrectly identified as cells by the automatic segmentation, while the orange circles are the cells that were identified by the manual segmentation but not by the automatic one.

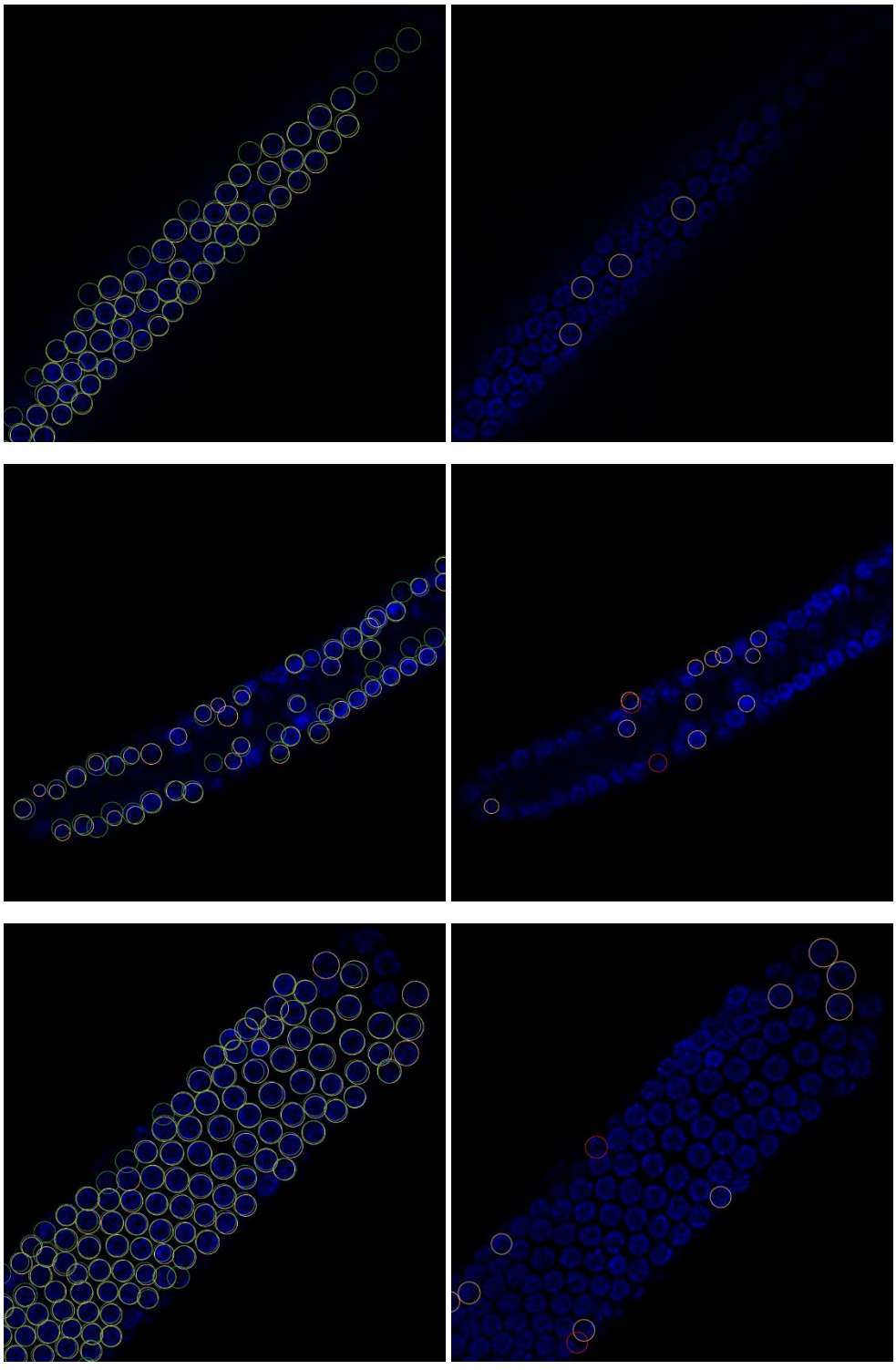


Figure S2: Zoomed-in regions of the images in Figure S1. Circle colors have the same meaning as in Figure S1. The second row is zoomed in on the distal tip. The third row is zoomed in on the pachytene region, just before the loop.

Section S2: Diffusion Maps Algorithm for Ordering Cells along the Center Line

In the Diffusion Maps algorithm (1), a random walk is constructed over a set of data points, with hopping probabilities between pairs of points determined by their pairwise distances and a kernel function. Here, the data points were the positions of the germ cells in a single germline. When the data points lie on a lower-dimensional manifold, the algorithm produces a robust ordering of the data along its principal nonlinear axis (or axes) on the manifold. In the case of the germline, the germ cells essentially lie on a 1D manifold, the center line of the germline. The algorithm was implemented using custom MATLAB code.

First, weights between data points (germ cell positions) were calculated by passing their pairwise distances through a Gaussian kernel function. Weights between data points are related to the probability of a random walker jumping from one of those data points to the other, with higher weights corresponding to higher hopping probabilities. The width of the Gaussian kernel determines the relevant scale of hopping. If the kernel width is much smaller than even the smallest distance between cells, then all weights between data points will be near zero, and a random walker cannot jump between any pair of data points. As the width is increased from zero, there is a scale at which the data appear 3D, then 2D, then 1D. At the 2D scale, a hopper can jump across the entire depth of the flattened germline in one jump, but not the diameter or length of the germline. At the 1D scale, the hopper can jump across the entire diameter of the germline in one jump, but not the length of the germline. If the kernel width is larger than the entire germline, all weights between data points will be close to one, and a random walker can jump between any two data points, no matter how far apart they are in space. At this scale, the data is essentially zero-dimensional from the point of view of the hopper, all collapsing to a single point. Previous work has developed an automated way of choosing the kernel width (2, 3). In practice, Diffusion Maps is not sensitive to the precise value of the kernel width, as long as it is in the correct dimensionality regime. We chose the kernel width for each germline so that the data "appeared" one-dimensional to a random walker.

The weights were then assembled into a symmetric matrix, with entry (i, j) containing the weight between germ cell i and germ cell j. The rows were normalized so that the sum of each row equaled one. This normalized matrix can be interpreted as a Markov transition matrix, with entry (i, j) containing the probability of a hopper located at data point i jumping to data point j in one time step. As the number of data points approaches infinity, the eigenvectors of this Markov matrix approach the eigenfunctions of the Laplace (diffusion) operator with Neumann (reflecting) boundary conditions (2). The first eigenvector of this matrix is a vector of ones, and contains no information. The first nontrivial eigenvector is one-to-one with and parameterizes the principal nonlinear axis of the data, the center line of the gonad tube. Element i of this eigenvector is associated with germ cell i; therefore, the monotonic ordering of the elements gives the ordering of the cells according to their positions along the center line. Diffusion Maps does not give the arc length positions of the cells, only their ordering.

Section S3: Estimating the Apoptosis Term, $\boldsymbol{R(x)}$

Evaluation of Equation 5 of the main text requires an expression for $R(x)$, which accounts for cell death in the germline. It is common to observe several germ cells undergoing apoptosis in a given fixed germline. Apoptotic germ cells are recognizable because their chromatin condenses and they undergo cellularization (4, 5). The former causes the cells to exhibit strong fluorescence when stained for DNA, while the latter causes the cells to exhibit essentially no fluorescence when stained for dpMPK-1 (Figure S1). From the time that a germ cell first shows symptoms of apoptosis to the time that the cell is removed from the germline by sheath cells is about 1 hour (4). In this time, a dying germ cell can only travel about one cell diameter before being cleared from the germline (6). As a result, the frequency at which cell corpses are observed at a given position is essentially the same as the frequency at which cells undergo apoptosis at that position.

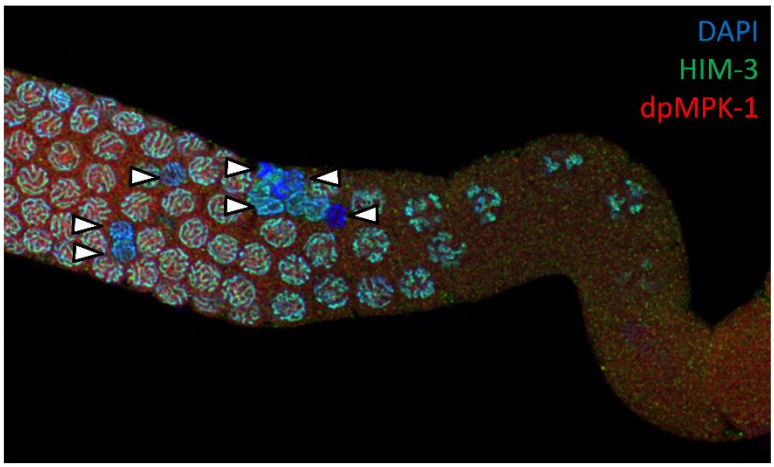


Figure S3: Image of the loop region of a *C. elegans* germline. Apoptotic cells (denoted by white arrowheads) are recognizable by their strong DAPI signal and lack of dpMPK-1 signal.

By this argument, we assumed that the death/clearance rate $R(x)$ is proportional to the number of cell corpses observed at x. We can decompose $R(x)$ into a shape function times a constant of proportionality that determines its scale. Under our assumptions, the shape function is given by the probability distribution of cell corpse locations (Figure S2). The constant of proportionality is r, the total rate of cell death, which can be estimated from data in the literature, as discussed in the main text. Note that the source term $S(x)$ could be estimated in a similar manner by looking at the relative frequency of cell divisions as a function of arc length across germlines, but this was not explored here.

However, in a given germline there are too few corpses to estimate the shape of $R(x)$ accurately. To address this, we pooled corpse counts from multiple fixed germlines to estimate the average shape of $R(x)$ over multiple germlines. The pooling process, itself, requires aligning spatial positions across different germlines. Here we will describe a method for registering arc length positions across different germlines. This is the only part of our approach that requires averaging across germlines.

Section S4: Registering Arc Length Positions across Germlines

Since there are so few cell corpses in a given germline, we need to estimate the shape of $R(x)$ from multiple germlines. Germlines come in different sizes, so pooling corpse positions across germlines requires that arc length positions in different germlines be registered or transformed to a common axis. Registering positions between two germlines is equivalent to determining an invertible function that maps positions in one germline to corresponding positions in the other. We assume that germ cells at "corresponding positions" are at the same developmental maturity, are the same age, and have spent the same amount of time in their respective germlines. Under the assumption that germ cells are arranged according to their maturity, this invertible function exists. The mapping will locally stretch or compress positions in one germline, like an accordion, to match the corresponding positions in the other. There will be a different mapping between each pair of germlines.

If x_1 refers to arc length positions in the first germline and x_2 refers to those in the second. The goal is to estimate an invertible function that maps x_1 to x_2, i.e. $x_2 = g(x_1)$ and $x_1 = g^{-1}(x_2)$. In the Derivation section, we introduced the probability density of germ cell arc length positions, $f_X(x)$, which quantifies the local "concentration" of germ cells in a particular germline. A related quantity is the cumulative distribution of arc length positions, $F_X(x)$, which quantifies the cumulative fraction of germ cells located at or before x.

Since x_1 and x_2 are related by an invertible function, their cumulative distributions must satisfy:

$$F_{X_1}(x_1) = F_{X_2}(x_2). \tag{S1}$$

Proof of Equation S1:

1. $F_{X_1}(x_1) = P(X_1 \leq x_1)$ (definition of a cumulative distribution function)
2. $= P(g(X_1) \leq g(x_1))$ (applying $g(\cdot)$ to both sides, and noting that $g(\cdot)$ is invertible, one-to-one and onto, and monotonically *increasing*)
3. $= P(X_2 \leq x_2)$ (using $X_2 = g(X_1)$ and $x_2 = g(x_1)$)
4. $= F_{X_2}(x_2)$. (definition of a cumulative distribution function)

Therefore, the invertible function we are seeking is $x_2 = g(x_1) = F_{X_2}^{-1}\left(F_{X_1}(x_1)\right)$. For each germline, $F_X(x)$ and its inverse are both measurable from data, meaning the mapping between any pair of germlines is measureable.

Using this approach, germ cell corpse positions from all germlines were transformed to their corresponding positions in a single germline. The particular germline used does not affect the result. The shape of $R(x)$ was estimated from these corpse positions the same way as $f_X(x)$, by kernel density estimation. This shape function for $R(x)$ was then transformed back to the arc length axis of each germline.

Note that the local source term $S(x)$ in Equation 1 of the main text can also be estimated in a similar fashion by identifying the positions of mitotically dividing cells in multiple germlines and registering the positions across germlines. Recent work suggests, though, that production throughout the mitotic region is roughly uniform (7).

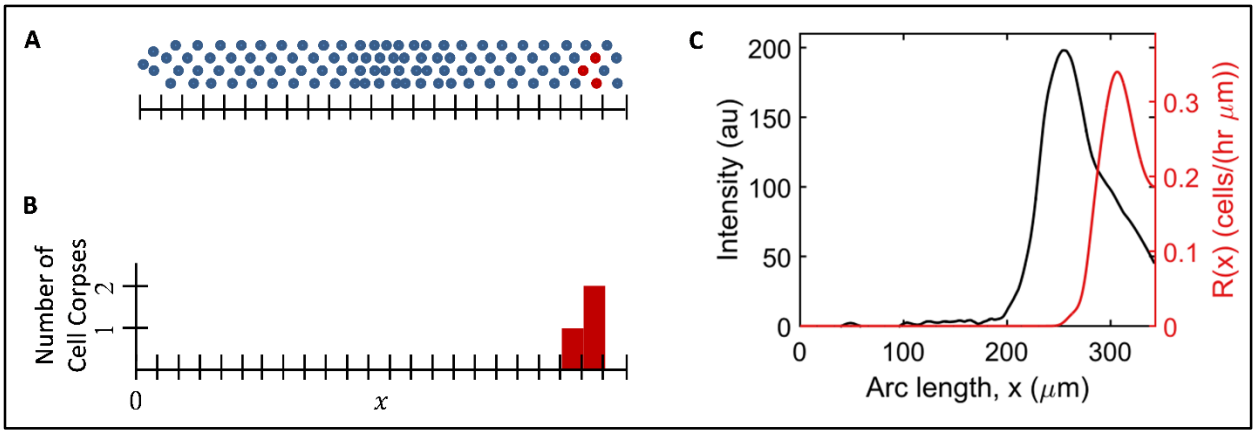


Figure S4: Estimation of $R(x)$. Assuming that the time rate at which cells undergo apoptosis at a given location is proportional to the frequency with which corpses are observed at that location, the shape of the spatially-dependent sink term $R(x)$ is the probability density function of cell corpse locations. **A)** Shows a schematic of a germline, where red cells are undergoing apoptosis. **B)** The histogram of cell corpse positions. With enough corpses, normalizing this histogram would give a good approximation of the shape of $R(x)$. Since there are not enough corpses in a single germline to estimate $R(x)$, corpse positions from multiple germlines must be aligned and pooled. **C)** Plotted in red is $R(x)$, estimated from 63 cell corpses pooled from 6 germlines and plotted against arc length position in a particular germline. The spatial dpMPK-1 profile from the same germline is shown in black. This shows that the peak rate of cell death occurs spatially (and temporally) after the peak of the dpMPK-1 pulse.

Section S5: Error Analysis

Uncertainty in $f_X(x)$

Assuming that we can measure germ cell arc length positions accurately, uncertainty in the probability density functions $f_X(x)$ for each germline can be approximated by their root mean squared error (RMSE) in the asymptotic limit of many samples. The mean squared error (MSE) of the estimate of $f_X(x)$ is the squared bias of the estimate (introduced by oversmoothing the true function) plus the variance of the estimate (introduced by estimating the function from a finite set of observations). When kernel density estimation is used to estimate probability density functions, the expression for the MSE is (8):

$$\delta f_X(x)^2 = \frac{h^4\left(f_X{}''(x)\right)^2}{4} + \frac{R f_X(x)}{nh}. \qquad \text{(S1)}$$

The first term in the sum is the squared bias of the estimate of $f_X(x)$, and the second term is the variance of the estimate. h is the bandwidth of the smoothing kernel used in the density estimation, and n is the number of observations (here, the number of cells in a germline). ${f_X}''(x)$ is the second derivative of the density, meaning that regions of the density function with larger curvature are more difficult to estimate accurately. This quantity was calculated by fitting the estimates of $f_X(x)$ with splines (MATLAB *csapi*) and taking the second derivative of the spline (MATLAB *fnder*). Finally, R is a property of the kernel function used in density estimation; for a kernel function $g(u)$, $R = \int_{-\infty}^{\infty} g(u)^2 du$. Here, an approximately Gaussian kernel was used, for which $R = 1/2\sqrt{\pi}$. Technically, density estimation was done via solving a diffusion equation, which acts much like a Gaussian smoothing kernel, but with better estimates of the density near the boundaries of the domain. This expression for the MSE should overestimate the error of the estimate near the boundaries.

Uncertainty in $t(x)$

Uncertainty in the estimates of $t(x)$ propagate from: uncertainty in $f_X(x)$, uncertainty in the values of the parameters in the model, and uncertainty in the shape of the apoptosis function $R(x)$.

Uncertainty in the model parameters was accounted for by uniformly sampling the literature ranges for N_{tot}, s, and the rate of ovulation (used to calculate r). Sampling was performed using a Latin Hypercube design (MATLAB's *lhsdesign*) to generate 100,000 samples. The value of τ produced by each parameter combination was calculated; if the value of τ was outside of the literature reported range for τ (48-54 hrs (9)), then the parameter set was discarded. After this pruning, 31,552 parameter sets remained. This collection of acceptable parameter sets was sampled from during the next step.

To estimate the effect of uncertainty in the shape of the apoptosis function $R(x)$, 5,000 samples were bootstrapped per germline from the collection of corpse observations. The shape of $R(x)$ was calculated for each randomly-sampled set of corpses. Then, for each sample, a parameter set was drawn at random from the collection of acceptable parameter sets, with replacement. Finally, $t(x)$ was calculated for that set of corpses, that parameter set, and that germline. The result was 5,000 estimates of $t(x)$ for each germline, the distribution of which accounted for uncertainty in the model parameters and the shape of $R(x)$. We denote the standard deviation of this distribution, as a function of x, $\delta t_{boot}(x)$.

The total uncertainty in $t(x)$, for each germline, is given by:

$$\delta t(x)^2 = \delta t_{boot}(x)^2 + \left(\frac{dt}{df_X}\right)^2 \delta f_X(x)^2, \tag{S2}$$

where $\frac{dt}{df_X}$ is, from Equation 5 in the main text, $\frac{dt}{df_X} = \int_0^x \frac{N_{tot}}{s - \int_0^w R(u)du} dw$, and $\delta f_X(x)$ is the RMSE of $f_X(x)$. This quantity was calculated separately for each germline.

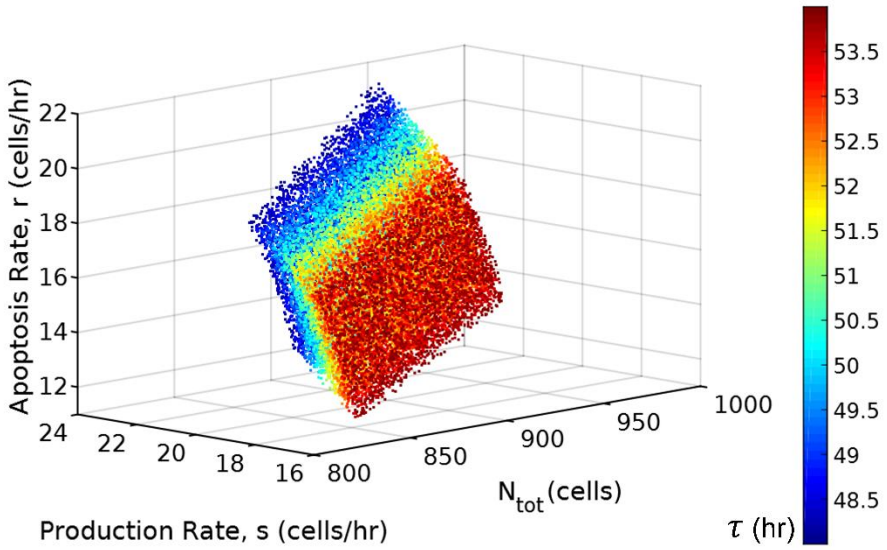


Figure S5: The collection of acceptable samples of N_{tot}, s, and r. Colors correspond to the value of τ, in hours, calculated from that parameter set, using Equation 7 of the main text. Axes limits are the ranges consistent with the literature. Using knowledge from the literature of all four parameters significantly reduces the volume of acceptable parameter combinations. The resulting region is called the feasible set (10).

Uncertainty in MPK-1 activation dynamics

Uncertainty in the dynamics of dpMPK-1 estimated from fixed samples arose from measurement uncertainty of the antibody staining and propagation of uncertainty from the time estimates. The measurement uncertainty was taken to be the standard deviation of the nuclear dpMPK-1 intensity measurements around the smoothed dynamics for that germline. If $y(t)$ is the fluorescence intensity of dpMPK-1 with respect to time and $\delta y_{measure}$ is the measurement uncertainty, then the total uncertainty in $y(t)$ is:

$$\delta y(t)^2 = \delta y(t)^2_{measure} + \left(\frac{dy}{dt}\right)^2 \delta t(x)^2, \qquad \text{(S3)}$$

where dy/dt is the derivative of $y(t)$ with respect to t, and δt is the uncertainty in t. This derivative was calculated by fitting $y(t)$ with splines (MATLAB *csapi*) and taking the derivative (MATLAB *fnder*). This calculation was performed for each germline.

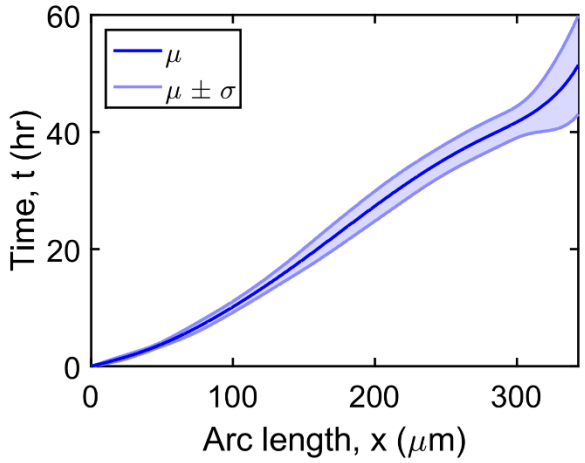


Figure S6: Sample $t(x)$ from a germline, plus and minus one standard deviation $\delta t(x)$, which accounts for uncertainties propagated from errors in estimating $f_X(x)$, uncertainty in the parameter values, and uncertainty in the shape of $R(x)$.

Section S6: Additional Figures

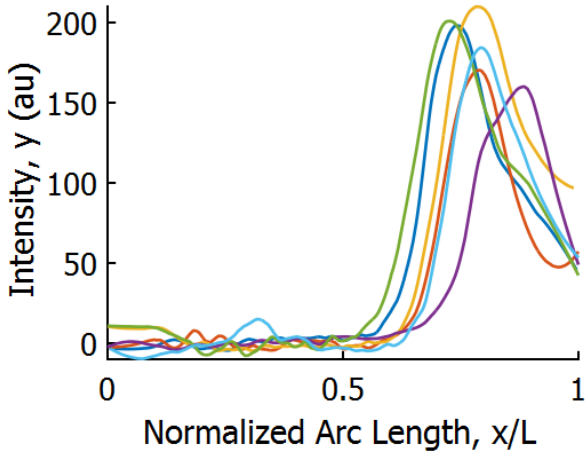


Figure S7: dpMPK-1 fluorescence intensity versus normalized arc length (arc length divided by the total distance from the distal tip to the loop) for multiple germlines. The images were acquired in the same experiment, at a set microscope condition. Plotting this way does not cause the spatial dpMPK-1 profiles from multiple germlines to collapse.

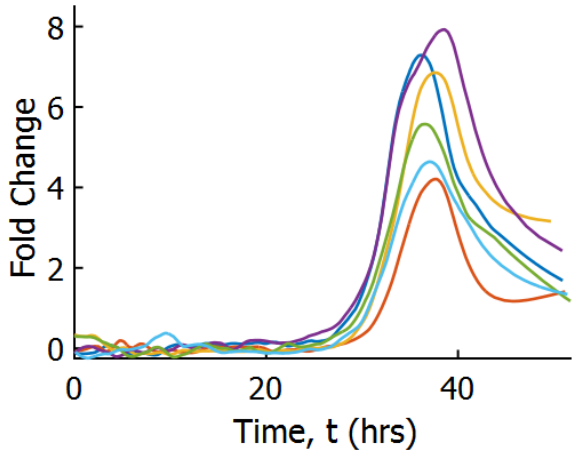


Figure S8: The fold change in dpMPK-1 fluorescence intensity, relative to background levels, as a function of time. Background fluorescence was calculated by averaging the dpMPK-1 intensity in the mitotic and early meiotic region, where there should be much less active MPK-1 than in the pachytene region.

SUPPORTING REFERENCES

1. Coifman, R. R., and S. Lafon. 2006. Diffusion maps. Appl Comput Harmon A 21:5-30.
2. Coifman, R. R., Y. Shkolnisky, F. J. Sigworth, and A. Singer. 2008. Graph Laplacian tomography from unknown random projections. IEEE T Image Process 17:1891-1899.
3. Ferguson, A. L., A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis. 2010. Systematic determination of order parameters for chain dynamics using diffusion maps. P Natl Acad Sci USA 107:13597-13602.
4. Gumienny, T. L., E. Lambie, E. Hartwieg, H. R. Horvitz, and M. O. Hengartner. 1999. Genetic control of programmed cell death in the Caenorhabditis elegans hermaphrodite germline. Development 126:1011-1022.
5. Lant, B., and W. B. Derry. 2013. Methods for detection and analysis of apoptosis signaling in the C. elegans germline. Methods 61:174-182.
6. Crittenden, S. L., K. A. Leonhard, D. T. Byrd, and J. Kimble. 2006. Cellular analyses of the mitotic region in the Caenorhabditis elegans adult germ line. Mol Biol Cell 17:3051-3061.
7. Fox, P. M., and T. Schedl. 2015. Analysis of germline stem cell differentiation following loss of GLP-1 notch activity in Caenorhabditis elegans. Genetics 201:167-184.
8. Hansen, B. 2009. Lecture Notes on Nonparametrics. University of Wisconsin. Web. Oct. 2015.
9. Jaramillo-Lambert, A., M. Ellefson, A. M. Villeneuve, and J. Engebrecht. 2007. Differential timing of S phases, X chromosome replication, and meiotic prophase in the C-elegans germ line. Dev Biol 308:206-221.
10. Frenklach, M., A. Packard, P. Seiler, and R. Feeley. 2004. Collaborative data processing in developing predictive models of complex reaction systems. Int J Chem Kinet 36:57-66.