From heterogeneous datasets to predictive models of embryonic development

Sayantan Dutta[1], Aleena L. Patel[1], Shannon E. Keenan[1], Stanislav Y. Shvartsman[1,2,3,4,*]

[1]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ-08544
[2]Lewis Sigler Institute of Integrative Genomics, Princeton University, Princeton, NJ-08544
[3]Department of Molecular Biology, Princeton University, Princeton, NJ-08544
[4]Center of Computational Biology, Flatiron Institute, New York, NY-10010
[*]Corresponding author: stas@princeton.edu

Abstract

Modern studies of embryogenesis are increasingly quantitative, powered by rapid advances in imaging, sequencing, and genome manipulation technologies. Deriving mechanistic insights from the complex datasets generated by these new tools requires systematic approaches for data-driven analysis of the underlying developmental processes. Here we use data from our work on signal-dependent gene repression in the fruit fly, *Drosophila melanogaster*, to illustrate how computational models can compactly summarize quantitative results of live imaging, chromatin immunoprecipitation, and optogenetic perturbation experiments. The presented computational approach is ideally suited for integrating rapidly accumulating quantitative data and for guiding future studies of embryogenesis.

Modern studies of embryonic development started in the 1980s, when mutagenesis screens in the fruit fly revealed associations between individual genes and external structures of the larva, which hatches approximately one day after egg fertilization [1]. By the end of the 20th century, molecular biologists established that most genes identified by genetic approaches encode transcription factors and signaling enzymes, and these genes are conserved across the animal kingdom. While early molecular studies of these components relied mainly on qualitative approaches, the current ongoing research is increasingly quantitative and multi-pronged, harnessing advances in imaging, sequencing, and genome manipulation technologies [2,3]. Extracting mechanistic insights from the heterogeneous datasets generated by these new experimental techniques, each of which reveals a different aspect of development, requires systematic strategies for data integration and analysis [4–6]. Based on a large body of work on nonliving chemical systems, we propose that computational models can serve as both compact repositories of data and as platforms for quantitative evaluation of candidate mechanisms of embryonic development [7]. Here we illustrate this point, using datasets from genetic, biochemical, and imaging studies of the signaling enzyme Extracellular signal Regulated Kinase (ERK) and its substrate, a transcription factor Capicua (Cic), both of which are critical for normal embryogenesis and are deregulated in human diseases.

Cic is a gene repressor that controls its targets by binding to specific sequences in their regulatory DNA. Repression is relieved when Cic is phosphorylated by ERK, an enzyme activated by signaling from cell surface receptors [8–10]. ERK counteracts gene repression by Cic by controlling its DNA binding as well as its sub-cellular localization. The relative contributions of each mechanism on gene transcription remain unclear, mainly because of the differences among the experimental systems and lack of a quantitative framework for data analysis. The fruit fly embryo, a versatile developmental system where Cic was originally identified in one of the mutagenesis screens, offers unique opportunities for elucidating how individual regulatory steps affect gene transcription. The embryo uses ERK to pattern the terminal structures of the larva and can be used to study gene regulation at multiple levels of biological organization, from molecular interactions to tissue morphogenesis [11].

We have established quantitative approaches for live imaging of Cic and its transcriptional effects in the embryo (Fig. 1c, 1e, and supplementary figure 1). These real-time studies reveal dynamics of molecular and cellular processes at single-cell resolution and can be coupled with biochemical assays that quantify concentration and genome-wide DNA binding patterns of Cic [12] (Fig. 1d). The power of live imaging and biochemical assays is further extended by combining them with optogenetic perturbations of the signaling cascade that culminates in ERK activation [12–14]. Here we demonstrate that datasets emerging from such varied studies (Fig. 1a) are already sufficient for constraining the parameters in a computational model that accounts for the key dynamic processes in the early embryo. In addition, we show that the resulting model provides insights into the relative time scales and importance of the processes through which ERK relieves gene repression by Cic.

Our computational model is based on a three-tiered biophysical description of the following processes: nuclear divisions that establish a random nuclear packing under the common plasma membrane of the early embryo, signal-dependent nucleocytoplasmic shuttling, and transcription (Fig. 1b). Dynamic nuclear packing generates a heterogeneous medium for Cic, which is synthesized in the cytoplasm and shuttles in and out of the nuclei with rates that depend on Cic phosphorylation, which is in turn controlled by ERK. When in the cytoplasm, Cic is degraded with first order kinetics; in the nucleus, Cic reversibly binds to DNA, with phosphorylation-dependent rate constants. The Cic target gene tailless (*tll*) switches between transcriptionally active and inactive states. The integrated model has both stochastic and deterministic components: dissipative particle dynamics for nuclei (tier 1) [15], mean field kinetics for nucleocytoplasmic transport and phosphorylation (tier 2), and a Markov chain model for the Cic target gene (tier 3) [16] (Materials and Methods).

Since nuclear divisions are unaffected by nucleocytoplasmic shuttling and transcription, parameters for the first tier of the model can be constrained independently of the other two, by matching the experimental and pre-

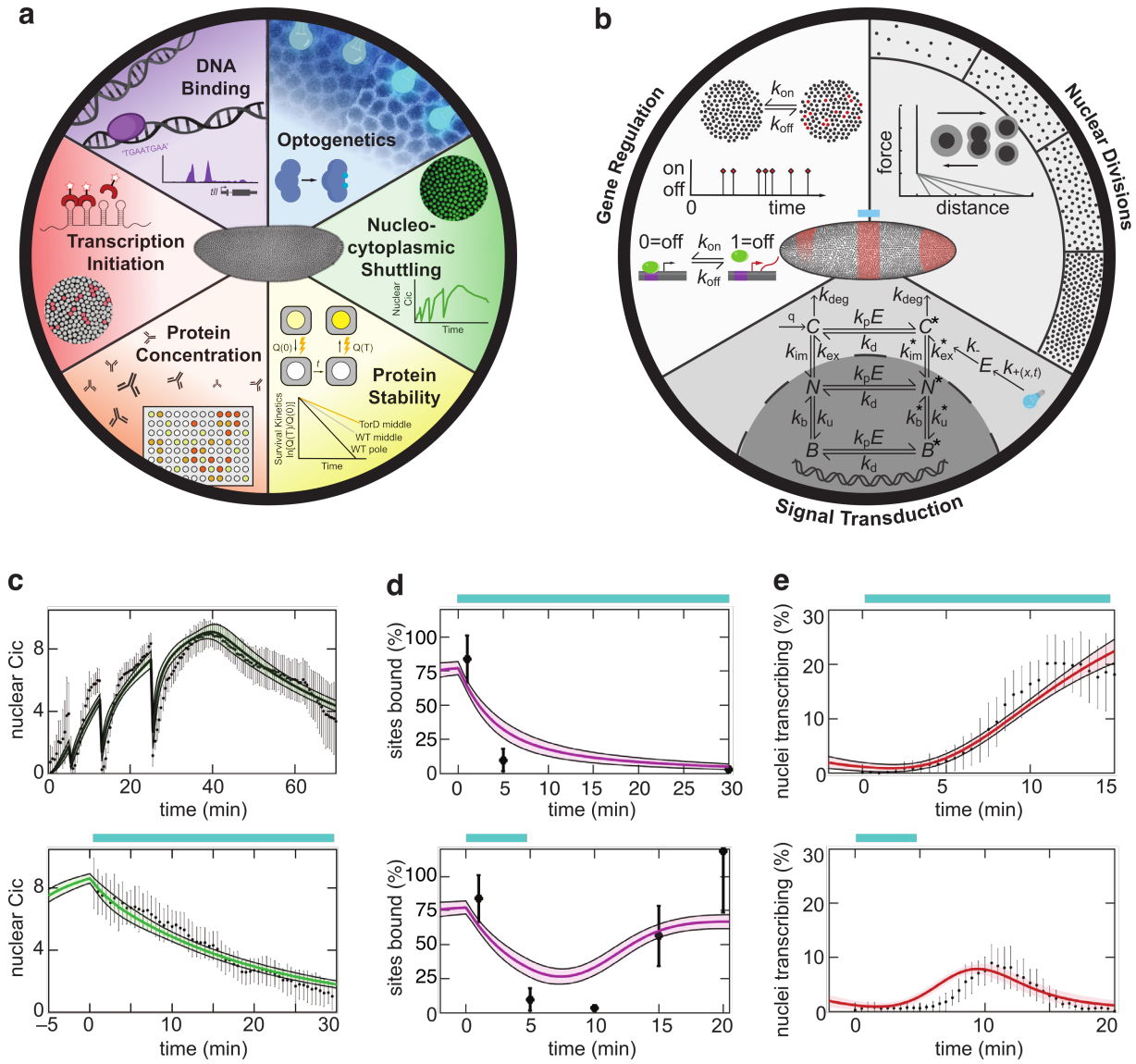

Figure 1: (a) A schematic of the experiments used in building the model– optogenetic stimulation, live imaging of fluorescently tagged Cic, estimation of protein lifetime using photo-switchable fluorescent proteins, concentration estimation by ELISA, visualization of transcription via MS2-MCP reporters, and DNA affinity by ChIP-Seq/ChIP-qPCR (clockwise). (b) A schematic of different elements of the model: A force-field model of nuclear divisions, a kinetic model describing the sub-cellular localization and biochemical state of Cic in response to ERK signals, and a stochastic model of transcription of Cic target genes (clockwise order). (c) Dynamics of nuclear Cic in the middle of a wild type embryo in cycle 11-14 (top) and in an embryo stimulated with optogenetic ERK signals in cycle 14 (bottom) for the simulation (green) compared to real data from fluorescence imaging (black) ($N = 3$ samples for both the experiments). Both data are normalized to cytoplasmic Cic level in cycle 14. (d) The dynamics of the fraction of Cic bound to DNA for the simulation in cycle 14 in response to step (top) and pulsed (bottom) optogenetic ERK signals. Results from the simulation (purple) are compared to chromatin immunoprecipitation data (black) (e) Transcription dynamics of the reporter of gene tailless (*tll*) in the middle of the embryo in response to step (top) and pulsed (bottom) optogenetic ERK signals. Percentage of nuclei with active transcription in cycle 14 in the simulated (red) and real embryo (black) are shown. In (c)-(e) the cyan lines in the top border of the plots refer to the time when optogenetic stimulation was on. The solid colored lines represent the mean simulation output from the 1000 parameter sets obtained from MCMC simulation and shaded colored regions represent one standard deviation from the mean. Errorbars denote standard deviation across all the experiments.

dicted statistics for internuclear distances, as described elsewhere [15]. The remaining part of the model has 12 free parameters. We used a stochastic optimization approach to obtain an ensemble of parameter vectors that minimizes the mismatch between model predictions and datasets from measurements of Cic dynamics and its effects on transcription ([17], Materials and Methods) . Data from live imaging of nuclear Cic (supplementary figure 1) and time resolved chromatin immunoprecipitation measurements of Cic binding to the regulatory DNA [12]

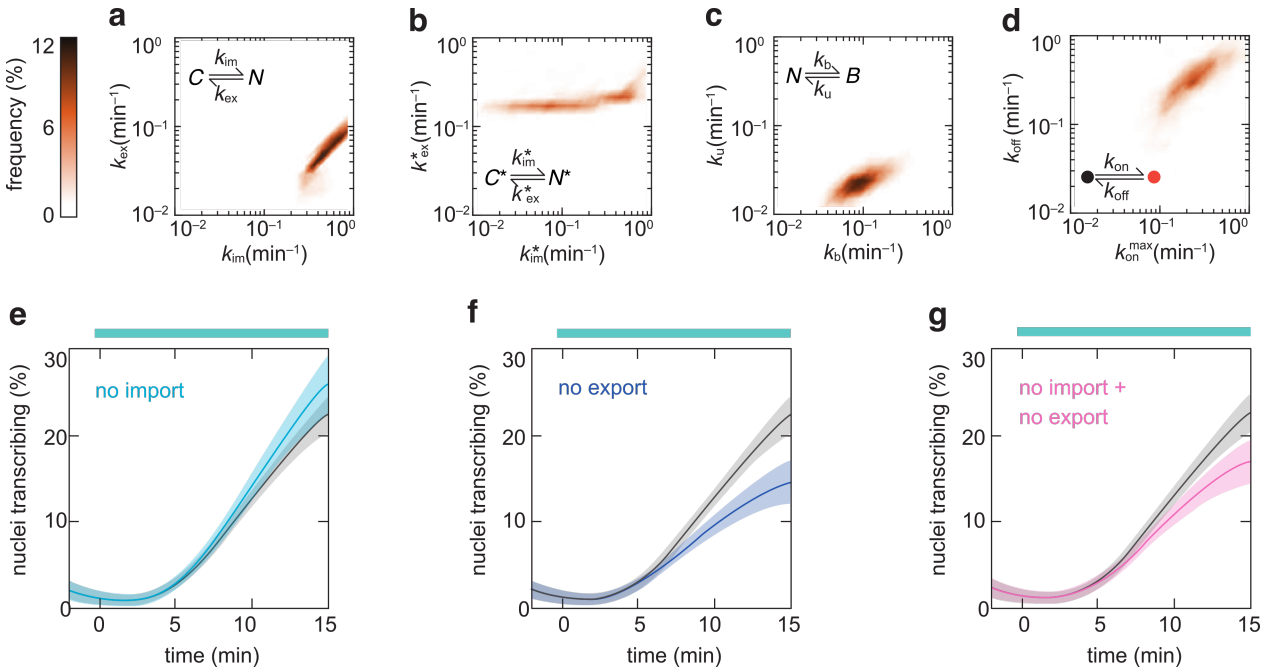

Figure 2: Histograph of the rate of import and export of unphosphorylated (a) and phosphorylated (b) Capicua, the rate of binding and unbinding of Capicua with DNA (c), and the rate of turning on and turming off of transcription of the Cic target gene tailless (*tll*) (d), obtained from 1000 MCMC simulations. The bins are in a logarithmic scale. The color represents the probability of the respective pair of data points to be in a particular bin on the 2D plot. The schematics in the inset show the processes the rate constants correspond to. Transcriptional responses in the middle of the embryo in the full model (black lines) (e-g) compared to the models with inhibited import (light blue lines) (e), inhibited export (dark blue lines) (f), and inhibited import and export (pink lines) (g), after the ERK signal was turned on (cyan bar) in cycle 14. In each plot, the dark line represents the mean and the light shaded region around it represents a region of 1 standard deviation across the mean. In the absence of all nucleocytoplasmic transport, transcription is reduced slightly (g), whereas in the absence of only export it is reduced significantly (f) and in the absence of only import it is increased slightly (e).

constrained the parameters related to tier 2 of the model (i.e. phoshphorylation, nucleocytoplasmic shuttling and DNA binding). Parameters for the transcription part of the model were constrained by live imaging of nascent mRNA production [12] controlled by the regulatory region of *tll*, visualized with single nucleus resolution via the MS2-MCP system. All of these measurements were done in both unperturbed embryos and embryos stimulated with optogenetically induced steps and pulses of ERK activation.

As shown in Figure 1, our computational model is quantitatively consistent with the most salient features of Cic regulation and function, revealed by imaging and biochemical assays (Fig. 1c-e, supplementary figure 2). The individual parameters of the model, most notably those related to the time scales of Cic nucleocytoplasmic shuttling and Cic-DNA binding, as well as parameters for transcription kinetics are either well constrained or have well defined upper bounds (Fig. 2a-d, supplementary figure 3). Furthermore, all parameter sets display a clear hierarchy among the time scales of Cic-DNA unbinding, nuclear export, and degradation. Consistent with these findings, model predictions over the parameter ensemble show that while changes in the nuclear localization of Cic are not essential for relieving its repressor function, they provide quantitative control of transcriptional output on the longer timescale (Fig. 2e-g). These results demonstrate how our computational model can generate credible predictions for the outcomes of perturbations that can not be realized experimentally, such as selective disruption of Cic nuclear import or export.

In conclusion, we argue that the rapid convergence of time-resolved quantitative data, advanced genetic perturbations, and biophysical modeling has reached the point when models can make credible quantitative predictions regarding multiple aspects of embryogenesis, especially those that are too costly and difficult to explore experimentally. The presented framework is ideally suited for rigorous evaluation of the mechanistic consistency of heterogeneous datasets. For example, we can show that the same model we have trained on data reporting transcription of a known target gene is also consistent with a Cic-dependent synthetic reporter (supplementary figure 4). Furthermore, this model gives us an opportunity to quantitatively evaluate the importance of a particular experiment by evaluating how a new dataset constrains and changes the existing model parameters (supplementary figure 5). Ultimately, the model can be extended to processes where coupling between different regulatory tiers is not unidirectional, as commonly happens

in real systems. For example, the gene products are themselves transcription-regulating proteins that introduce more complex feedback loops [18]. Future efforts should focus on strategies for integrating data from independent laboratories, similar to what has been successfully accomplished by the "data collaboration" framework in combustion research [7]. If successful, such a framework will not only advance our fundamental understanding of development but will also lead to more coordinated and cost-effective research efforts.

Materials and Methods

Cic-sfGFP live imaging

To track endogenous nuclear Cic intensity in the middle of the embryo, flies with Cic Crispr-tagged with sfGFP (previously described in [12]) were crossed to flies containing a nuclear marker, histone-RFP. Embryos from this cross have maternally deposited Cic-sfGFP and his-RFP. Collected embryos were manually dechorionated and mounted onto a slide. The slide was placed on a live imaging chamber with halocarbon oil. All imaging was done on a Leica SP5 point scanning confocal microscope using the 63X oil objective. A central region of the embryo was imaged encompassing about 500 nuclei using 20 % and 10 % laser powers for the 488 nm and 561 nm lasers respectively. A Z-stack of 12 images (step size of 1 micron) was captured every 30 seconds from nuclear cycle 12 through to gastrulation. From the red channel capturing the nuclear marker, we segmented the images using the imbinarize function of the MATLAB and found the nuclear Cic levels by averaging the intensity from the green channel capturing the sf-GFP.

To track how quickly Cic leaves the nucleus in response to a light stimulus (in other words, ERK activation), we imaged embryos expressing Cic-sfGFP and OptoSOS [19]. OptoSOS expresses SOS tagged with mCherry, indicating that a red histone marker could not be used in this case. Additionally, blue light activates Opto-SOS, so simply imaging Cic-sfGFP with the 488 laser activates ERK. We mounted embryos in the same way as previously described. Within the first 5 minutes of nuclear cycle 14 we imaged Cic-sfGFP (and subsequently activated the optogeneic construct) using 20 % 488 nm laser. A Z-stack of 12 images (step size of 1 micron) was captured every 30 seconds until gastrulation. From the OptoSOS-mcherry channel, we segmented the images using the imbinarize function of the MATLAB and identify the pixels corresponding to the membranes. We assumed all other pixels correspond to nuclei (in NC 14 nuclei are tightly packed in the cell) and found the the nuclear Cic levels by averaging the intensity from the channel capturing the sf-GFP.

Model Description

In our proposed model, Capicua (Cic) molecules have three different localizations– cytoplasmic (C, C^*), nuclear but unbound to the DNA (N,N^*) and nuclear and bound to the DNA (B, B^*). The molecule in each localization can be in two states– unphosphorylated (C, N, B) and phosphorylated (C^*, N^*, B^*) (Fig. 1b). The transition between these two states in each region depends on level of active ERK (E), which switches reversibly to it's inactive form from the total constant pool of ERK $E_{\max}$. k_+ and k_- are the rates of transition from inactive to active forms of ERK and vice-versa. Equation (1) describes the dynamics of ERK signals. $k_+ = k_{\text{endo}}(x) + k_{\text{opto}}(x,t)$, where k_{endo} has an endogenous profile from pole to the middle of the embryo and k_{opto} represents the optogenetic activation of ERK. In the model, ERK levels are same in the nucleus and cytoplasm. In equations (2)-(7), we describe the dynamics among two states and three localization of Cic. The rate of phosphorylation of Cic is proportional to active ERK (E) and k_{p} is the rate constant of phosphorylation, and k_{d} is the rate constant of dephoshorylation. We used the endogenous level in the pole of the WT embryo as the unit signaling level of ERK. Cic is synthesised with a volumetric rate of q and degrades with a rate constant of k_{deg} in cytoplasm. The rate constants of degradation in cytoplasm are independent of the phosphorylation state of the molecule [9]. The phosphorylation changes the protein structure, so the rate constants of import and export (affected by affinity to membrane protein) are different for the unphosphorylated ($\hat{k_{\text{im}}}$, $\hat{k_{\text{ex}}}$) and phosphorylated ($\hat{k^*_{\text{im}}}$, $\hat{k^*_{\text{ex}}}$) states. Similarly, affinity to DNA is also different for the unphosphorylated (k_{b}) and phosphorylated (k^*_{b}) states. However, for the sake of simplicity we assumed the rate constants of unbinding to be same ($k_{\text{u}} = k^*_{\text{u}}$) and rate of binding to DNA of unphosphorylated molecule to be significantly higher than the phosphorylated ones ($k_{\text{b}}/k^*_{\text{b}} = 1000 >> 1$). $S^{\max}$ is the total number of Cic binding sites in the *Drosophila* genome and $C_N^{\max}$ is the maximum concentration of nuclear Cic in NC 14.

$$\frac{\mathrm{d}E}{\mathrm{d}t} = k_+(x,t)(E_{\max} - E) - k_- E; \tag{1}$$

$$\begin{aligned}\frac{\mathrm{d}(V_{\text{c}}^i C)}{\mathrm{d}t} &= qV_{\text{C}} - k_{\text{deg}} V_{\text{c}}^i C - \hat{k_{\text{im}}} A_{\text{n}}^i C \\ &\quad + \hat{k_{\text{ex}}} A_{\text{n}} N - k_{\text{p}} E V_{\text{c}} C + k_{\text{d}} V_{\text{c}}^i C^*.\end{aligned} \tag{2}$$

$$\begin{aligned}\frac{\mathrm{d}(V_{\text{c}}^i C^*)}{\mathrm{d}t} &= -k_{\text{deg}} V_{\text{c}}^i C^* - \hat{k_{\text{im}}}^* A_{\text{n}}^i C^* \\ &\quad + \hat{k^*_{\text{ex}}} A_{\text{n}}^i N^* + k_{\text{p}} E V_{\text{c}}^i C - k_{\text{d}} V_{\text{c}}^i C^*\end{aligned} \tag{3}$$

$$\begin{aligned}\frac{\mathrm{d}(V_{\text{n}}^i N)}{\mathrm{d}t} &= \hat{k_{\text{im}}} A_{\text{n}}^i C - k_{\text{ex}} A_{\text{n}} N^i - k_{\text{p}} E V_{\text{n}}^i N + k_{\text{d}} V_{\text{n}}^i N^* \\ &\quad - k_{\text{b}}(N/C_N^{\max})(S^{\max} - V_{\text{n}}^i(B + B^*)) + k_{\text{u}} V_{\text{n}} B\end{aligned} \tag{4}$$

$$\frac{\mathrm{d}(V_{\mathrm{n}}^i N^*)}{\mathrm{d}t} = k_{\mathrm{im}}^{\hat{*}} A_{\mathrm{n}}^i C^* - k_{\mathrm{ex}}^* A_{\mathrm{n}}^i N^* + k_{\mathrm{p}} E V_{\mathrm{n}} N - k_{\mathrm{d}} V_{\mathrm{n}}^i N^* - k_{\mathrm{b}}^*(N^*/C_N^{\max})(S^{\max} - V_{\mathrm{n}}(B+B^*)) + k_{\mathrm{u}}^* V_{\mathrm{n}}^i B^* \tag{5}$$

$$\frac{\mathrm{d}(V_{\mathrm{n}}^i B)}{\mathrm{d}t} = -k_{\mathrm{p}} E V_{\mathrm{n}}^i B + k_{\mathrm{d}} V_{\mathrm{n}} B^* + k_{\mathrm{b}}(N/C_N^{\max})(S^{\max} - V_{\mathrm{n}}^i(B+B^*)) - k_{\mathrm{u}} V_{\mathrm{n}} B \tag{6}$$

$$\frac{\mathrm{d}(V_{\mathrm{n}}^i B^*)}{\mathrm{d}t} = k_{\mathrm{p}} E V_{\mathrm{n}} B - k_{\mathrm{d}} V_{\mathrm{n}}^i B^* + k_{\mathrm{b}}^*(N^*/C_N^{\max})(S^{\max} - V_{\mathrm{n}}^i(B+B^*)) - k_{\mathrm{u}}^* V_{\mathrm{n}} B^* \tag{7}$$

During time integration, we non-dimensionalized all the concentrations by (q/k_{deg}) and time as $\tau = k_{\mathrm{deg}} t$. In cycle 11, we initialized the concentration of all the Cic pools as zero, whereas, in cycle 12-14 we distributed the total pool of unphosphorylated and phosphorylated Cic in the cytoplasm of two daughter cell, as the nuclear membrane breaks apart during mitosis. We used the geometric parameters A_n^i, V_n^i and V_c^i in nuclear cycle (i) 11-14 from [20]. For a given set of parameters, we calculated $C_n^{\max}$ as the maximum value of $N + N^*$ by setting k_b, k_b^*, k_u, k_u^* to zero and treating the bound and unbound pools together.

We introduce a stochastic model of transcription, where each nucleus flips between a transcriptionally active and inactive state (Fig. 1b). The rate of turning on transcription in a nucleus is k_{on}, which is proportional to the probability of finding all the sites not bound by Cic, a non-linear function of fraction of Cic sites bound (f_{b}), defined as $V_{\mathrm{n}}^i(B+B^*)/S^{\max}$ (Fig. 1b).

$$k_{\mathrm{on}} = k_{\mathrm{on}}^{\max}(1 - f_{\mathrm{b}})^m. \tag{8}$$

The rate of transcription turning off in a nucleus is k_{off}. If transcription is off in a nucleus in time t, it switches to the on state in the next timestep at $t + \Delta t$, with a probability $\exp(-k_{\mathrm{on}}(t)\Delta t)$. Similarly, if transcription is on in a nucleus, it switches to the off state with a probability $\exp(-k_{\mathrm{off}}\Delta t)$.

Parameter Estimation

There are 12 free parameters of the model (k_{im}, k_{ex}, t_0, k_{im}^*, k_{ex}^*, k_{p}, k_{dp}, k_{b}, k_{u}, $k_{\mathrm{on}}^{\max}$, k_{off}, m). In this 12 dimensional space, we started from an initial guess $\mathbf{p_0}$ chosen randomly in the pre-specified range (Table 1). Then we ran the Markov Chain Monte Carlo (MCMC) simulations for M steps, where in each step j the following operations were done:

- The normalized error from each experiment ϵ_{kj} is calculated as $\epsilon_{kj}^2 = \sum_{l=1}^{T}(S_{kl} - \mu_{kl})^2 / \sum_{l=1}^{T} \sigma_{kl}^2$, where, μ_{kl} and σ_{kl} are the mean and standard deviation of the lth datapoint of the kth experiment and S_{kl} is the equivalent model output.
- The total error of the simulation E_j is evaluated as the norm of a vector of normalized errors from all the experiments– $E_j^2 = \sum_{k=1}^{N} \epsilon_{kj}^2$.
- A vector $\mathbf{V}$ of length δ of the same dimension as the parameter space is chosen in a random direction.
- The parameter vector $\mathbf{p}^g$ is updated such that the mth parameter is $p_m^g = p_m^j(1 + V_m)$.
- The total error E^g is calculated for the parameter vector $\mathbf{p}^g$.
- If $E_g < E_j$, the movement in the parameter space is accepted with a probability 1, else, the movement is accepted with a probability $\exp(-\beta(E_g^2 - E_j^2))$.

As our initial guesses were uniformly dispersed in the parameter space, we chose $\beta = 500 >> 1$ and $\delta = 0.1$ to find the local minimum around each initial guess. We present results from 1000 simulations, each of which ran for $M = 2000$ steps, when the error becomes steady (supplementary figure 6). We rejected the simulations, where the error from any experiment is 1.5 times the experimental error. In Fig. 2a-d, we plot the relative frequency of each pair of parameters in a bin in logarithmic scale.

Comparing data and model predictions

In every step of the MCMC simulations, we solved equations (1)-(7) for 4 nuclei–a nucleus at the middle of a wild type embryo ($E = 1$), a nucleus at the pole of a wild type embryo ($E = 0$), a nucleus at the middle of the embryo with a step activation of ERK in cycle 14, and a nucleus at the middle of the embryo with a pulse activation of ERK of 5 minute duration in cycle 14 (E changes dynamically). We elaborate below how we compare the model output for these nuclei to the experimental data below.

Live Cic imaging

Though there are four pools of Cic in the nucleus, the live imaging data of Cic-sfGFP doesn't distinguish between the proteins bound and unbound to DNA or it's state of phosphorylation. So, for comparison with the experimental data, we define the total dimensionless concentration of nuclear Cic as $\tilde{C_{\mathrm{n}}} = \tilde{N} + \tilde{N^*} + \tilde{B} + \tilde{B^*}$, where $\tilde{N}$, $\tilde{N^*}$, $\tilde{B}$, $\tilde{B^*}$ are dimensionless form of the respective concentrations. However, the nuclear Cic intensity obtained from live imaging are in arbitrary units. So, we further divide the non-dimensional nuclear Cic concentration $\tilde{C_N}$ from the model by the concentration of cytoplasmic Capicua

Fixed parameters		
parameter	*physical significance*	*value*
q	rate of synthesis	eliminated by non-dimensionalization
k_{deg}	rate of degradation	0.09 min^{-1} [9]
k_{opto}	rate of activation of light dependent ERK signal	10 min^{-1} (estimated from [13])
k_{-}	rate of deactivation of ERK signal	0.5 min^{-1} (estimated from [13])
$E_{\mathrm{ss}}(\mathrm{opto})/E_{\mathrm{ss}}(\mathrm{pole})$	ratio of steady state ERK signalling in an optogenetically activated embryo and at the pole	2 (estimated from [13])
V_n^i, A_n^i	area and volume of nucleus of in nuclear cycle i. $i = \{11, 12, 13, 14\}$.	data from [20]
$\frac{S^{\max}}{V_{\mathrm{n}}^{14} C_{\mathrm{n}}^{\max}}$	ratio of Cic site and maximum number of Cic molecule	0.1 (data from [12] and Liu Yang)
Free parameters		
parameter	*physical significance*	*range of initial guesses*
t_0	Time of stop of synthesis of Cic in cycle 14	[5 15] min. (linear scale)
$k_{\mathrm{im}} = \hat{k_{\mathrm{im}}} A_n^{14}/V_n^{14}$	rate of import of unphosphorylated Cic into nucleus in cycle 14	[0.1 10] min^{-1}. (log scale)
$k_{\mathrm{ex}} = \hat{k_{\mathrm{ex}}} A_n^{14}/V_n^{14}$	rate of export of unphosphorylated Cic from nucleus in cycle 14	[0.1 10] min^{-1}. (log scale)
$k_{\mathrm{im}}^* = \hat{k_{\mathrm{im}}^*} A_n^{14}/V_n^{14}$	rate of import of phosphorylated Cic into nucleus in cycle 14	[0.01 1] min^{-1}. (log scale)
$k_{\mathrm{ex}}^* = \hat{k_{\mathrm{ex}}^*} A_n^{14}/V_n^{14}$	rate of export of phosphorylated Cic from nucleus	[0.01 1] min^{-1}. (log scale)
k_{p}	rate of phosphorylation in unit ERK signal	[1 100] min^{-1}. (log scale)
k_{d}	rate of dephosphorylation	[1 100] min^{-1}. (log scale)
k_{b}	rate of binding of unphosphorylated Cic	[0.01 1] min^{-1}. (log scale)
k_{u}	rate of binding of unphosphorylated Cic	[0.01 1] min^{-1}. (log scale)
$k_{\mathrm{on}}^{\max}$	rate of turning on transcription in absence of Cic	[0.01 1] min^{-1}. (log scale)
k_{off}	rate of turning off transcription	[0.01 1] min^{-1}. (log scale)
m	effective number of Cic sites near gene	[0 20] (linear scale)

Table 1: The value of fixed parameters and and range of initial guesses of the free parameters of the model along with their physical significance.

$(C + C^*)$ at time t_0, when the synthesis of Cic stops in cycle 14. We normalized the fluorescence intensity of Cic such that the maximum value of Cic intensity in cycle 14 in the middle of the embryo is 9, the experimentally reported ratio of nuclear and cytoplasmic Cic in cycle 14 [9]. In this way, we effectively normalized both the experimental and the computational time-series of nuclear Cic by the typical cytoplasmic concentration of Cic at cycle 14 (Fig. 1c). For comparing the ratio of nuclear Cic at middle and pole we found the ratio of $\tilde{C_N}$ at time t_0 in cycle 14 and compare them with the results from [12] (supplementary figure 2a).

DNA binding

The ChIP-Seq data tells us the amount of genome bound to Cic from the embryo. Although this data is from the entire embryo, endogeneous ERK signal is active in less than 10 % of the embryo. So, we assumed the data to be reflective of nuclei without any endogeneous ERK signal such as one in the middle of the embryo within experimental error ($\approx 20\%$). Furthermore, to compare with the model output f_b, the fraction of Cic sites bound, we normalized all the experimental ChIP-Seq data to the wild type data assuming all sites are bound in the middle of the wild type embryo in cycle 14.

Transcriptional response

The experimental data for transcriptional activity in the middle of the embryo in response to optogenetic activation is the fraction of nuclei transcribing in a given area [12], which is an ensemble property. However, during the MCMC optimization, we ran the simulations on a single nucleus on the middle of the embryo. To extract ensemble statistics out of this, we found the fraction of nuclei f_t transcribing in the middle of the embryo by solving the following equation,

$$\frac{\mathrm{d}f_t}{\mathrm{d}t} = k_{\mathrm{on}}(t)(1 - f_t) - k_{\mathrm{off}} f_t, \tag{9}$$

where, we calculated k_{on} as a function of f_b (equation 8) for the nucleus situated in the center of the nucleus. The

initial condition is $f_t = 0$ after the mitosis. This comparison is based on the assumption that E, the only input to the model is constant in the imaging region, which is evident from the experimental ERK profiles [9, 11]. We demonstrate that the solution of the equation is significantly close to the output of stochastic simulation (supplementary figure 7).

Acknowledgements

We thank Michael Frenklach, Eric Wieschaus, Trudi Schüpbach, Rocky Diegmiller, and members of the Shvartsman lab for helpful discussions. We thank Gary Laevsky and Molecular Biology Core Confocal Microscopy Facility for imaging support. We thank Liu Yang for the ELISA experiment. We thank Nareg J.-V. Djabrayan and Gerardo Jimenez for the synthetic reporter CZC. We acknowledge the computing cluster in Lewis Sigler Institute of Integrative Genmics for computational resources.

Author Contributions

S.E.K. and A.L.P. carried out all the experiments. S.E.K., S.D. and A.L.P. analyzed the data. S.Y.S, S.D. and S.E.K. designed the model. S.D. implemented the model and performed the simulations. S.D., A.L.P., S.E.K., and S.Y.S. wrote the manuscript

References

[1] Wieschaus, E. & Nüsslein-Volhard, C. The Heidelberg screen for pattern mutants of *Drosophila*: A personal account. *Annual Review of Cell and Developmental Biology* **32**, 1–46 (2016).

[2] Karaiskos, N. *et al.* The *Drosophila* embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).

[3] Wunderlich, Z. *et al.* Quantitative comparison of the anterior-posterior patterning system in the embryos of five *Drosophila* species. *G3: Genes, Genomes, Genetics* **9**, 2171–2182 (2019).

[4] Chen, S. Y. *et al.* Optogenetic control reveals differential promoter interpretation of transcription factor nuclear translocation dynamics. *Cell Systems* **11**, 336 – 353.e24 (2020).

[5] Blum, Y. *et al.* Temporal perturbation of ERK dynamics reveals network architecture of FGF2/MAPK signaling. *Mol Syst Biol* **15**, e8947 (2019).

[6] Guignard, L. *et al.* Contact area–dependent cell communication and the morphological invariance of ascidian embryogenesis. *Science* **369** (2020).

[7] Frenklach, M. Transforming data into knowledge—process informatics for combustion chemistry. *Proceedings of the Combustion Institute* **31**, 125 – 140 (2007).

[8] Ajuria, L. *et al.* Capicua DNA-binding sites are general response elements for RTK signaling in *Drosophila*. *Development* **138**, 915–924 (2011).

[9] Grimm, O. *et al.* Torso RTK controls Capicua degradation by changing its subcellular localization. *Development* **139**, 3962–3968 (2012).

[10] Patel, A. L. & Shvartsman, S. Y. Outstanding questions in developmental ERK signaling. *Development* **145** (2018).

[11] Goyal, Y., Schupbach, T. & Shvartsman, S. Y. A quantitative model of developmental RTK signaling. *Developmental Biology* **442**, 80 – 86 (2018).

[12] Keenan, S. E. *et al.* Rapid dynamics of signal-dependent transcriptional repression by Capicua. *Developmental Cell* **52**, 794–801.e4 (2020).

[13] Johnson, H. E. *et al.* The spatiotemporal limits of developmental ERK signaling. *Developmental Cell* **40**, 185–192 (2017).

[14] Patel, A. L. *et al.* Optimizing photoswitchable MEK. *Proc Natl Acad Sci USA* **116**, 25756–25763 (2019).

[15] Dutta, S., Djabrayan, N. J.-V., Torquato, S., Shvartsman, S. Y. & Krajnc, M. Self-similar dynamics of nuclear packing in the early *Drosophila* embryo. *Biophys. J.* **117**, 743 – 750 (2019).

[16] Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **135**, 216 – 226 (2008).

[17] Brooks, S. *Handbook of Markov chain Monte Carlo* (CRC Press, Boca Raton London, 2011).

[18] Lengyel, J. & Liu, X. Posterior gut development in *Drosophila*: a model system for identifying genes controlling epithelial morphogenesis. *Cell research* **8**, 273—284 (1998).

[19] Goyal, Y. *et al.* Parallel imaging of *Drosophila* embryos for quantitative analysis of genetic perturbations of the Ras pathway. *Disease Models & Mechanisms* **10**, 923–929 (2017).

[20] Gregor, T., Wieschaus, E. F., McGregor, A. P., Bialek, W. & Tank, D. W. Stability and nuclear dynamics of the Bicoid morphogen gradient. *Cell* **130**, 141 – 152 (2007).