Investigating The On-Off Problem: The Effect of 'Invisible' Cells on Transcription Measurement

Jonathan Liu[1] **and Susanna Weber**[1]

***For correspondence:**

[1]Department of Physics, UC Berkeley

Abstract

Live imaging of transcription using tools such as MS2 or PP7 to label nascent transcripts allows for quantification of *in vivo* transcription dynamics, and greater insight into pattern formation and genetic control within the embryo. However, these techniques are hindered by the detection limit of the microscope used, as there is no reliable way to determine the difference between a cell that is not transcribing the target gene, and one that is active, but producing fluorescence below the detection level. Here, we investigate whether there is a significant number of such 'invisible' cells, and if so, what fraction of all cells fall into this category. We approach this problem by simulating the transcription of *hunchback*. Our results indicate that a significant fraction of all active cells produce a fluorescence below the detection limit, implying that when using fraction on as a metric of transcriptional activity, one needs to account for this population of undetected nuclei.

Main Text

Introduction

Imaging single molecules in cells reveals the process of gene expression from transcription to translation, which is essential for the study of development. The MS2 system for live imaging of nascent RNA production (***Garcia et al., 2013***; ***Lucas et al., 2013***) allows for visualization of this process in *Drosophila melanogaster* and other multi-cellular organisms with high spatial and temporal resolution. As an individual RNAP molecule transcribes a set of MS2 stem loops, MCP-GFP proteins bind to their respective loops. This results in sites of transcription within the nucleus that appear as fluorescent puncta under a laser-scanning confocal microscope. The localized fluorescence produced by GFP can then be used to visualize transcription dynamics at the single cell level, as well as along the anterior-posterior axis. The output pattern is caused by the accumulation of fluorescence protein at the target gene, revealing the spatiotemporal changes in transcription rates throughout the embryo.

The ratio of active to inactive cells can be quantified by measuring the fraction of all nuclei within a given anterior-posterior (AP) bin transcribing the target gene. This quantity can be expressed with the variable $f = \frac{n}{N}$, where n is the number of active cells within one AP bin, N is total number of cells per bin, and f is the fraction of all cells that are active. This quantity only measures whether a cell ever turns on during a nuclear cycle, and does not take into account whether it switches off

again. Thus, one can distinguish between 'on' cells, which are transcribing the target gene, and 'off' cells, which are not.

However, f, or 'fraction on', is only useful as a metric of transcription if all target gene transcription sites are being observed, that is, MCP-GFP is producing a detectable fluorescent spot at each transcription site (see Figure 1). Consider the scenario in Figure 2, where there are not only on and off cells, but active (or 'on') cells that are not producing enough fluorescence to be observable. This creates a third category of 'invisible active cells' - cells that are transcribing the target gene, and thus are contributing to f, but are not included in any measurements of f due to their undetectable fluorescence.

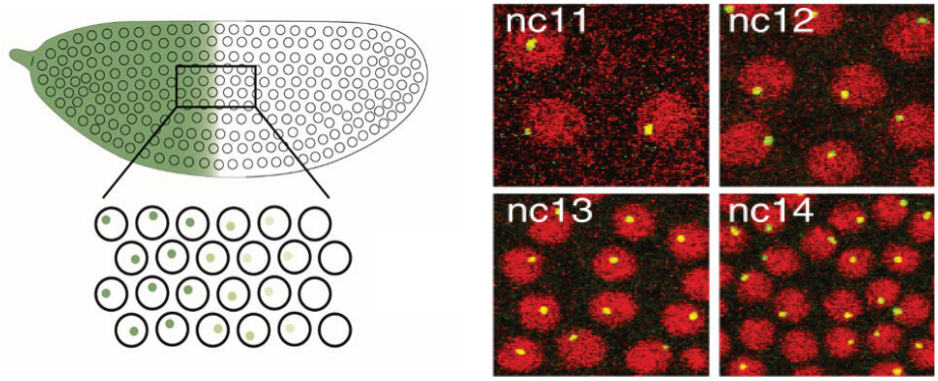

Figure 1. Left: A cartoon showing a field view of an embryo, where the green color represents sites of higher *hunchback* transcription activity. The black box is used to show a cartoon of the fluorescent puncta within individual nuclei. Fluorescence gradients like the one shown in this diagram are used to observe where and when transciption is happening within the embryo. Right: Fluorescent puncta within nuclei during nuclear cycles 11-14. Changes in these spots over time can be used to investigate transcription dynamics at the single cell level. Data taken from ***Garcia et al.*** (***2013***).

The goal of this project was to determine whether such 'invisible cells' really exist, that is, to determine whether current measurements of f using MCP-GFP are actually overlooking active sites of transcription. If so, the follow-up goal was to quantify what fraction of all active cells falls into this category, and how many transcription sites were going unrecorded.

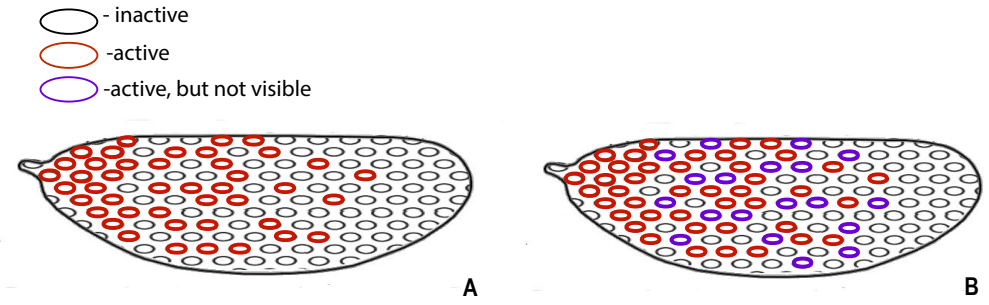

Figure 2. Representative fields of view (Histone-RFP) of nuclei in nc 14 in the activation (left) and transition (middle) regions. In this diagram, nuclei where transcription was detected at any point over the entire nc. are circled in red. The right graphic shows the effects of 'invisible' cells (purple), increasing the fraction of active cells, but not being recorded. Scale bars represent 10 mm.

However, detecting these 'invisible' cells cannot be accomplished by analyzing existing data sets

Variable Name	Mean Value	Standard Deviation	Description
t_{on}	4.2 min	1.50 min	Promoter turn on time
t_{off}	14.2 min	3.89 min	Promoter turn off time
r	see Equation 2	N/A	Polymerase loading rate
L	5.4 kb	N/A	Gene length
v	2.4 kb/min	0.76 kb/min	Elongation rate
p	0.2-0.8 x/L	N/A	AP Position

Table 1. Simulation variables. Figures showing the distribution of t_{on} and t_{off} can be found in Supplementary Information. The value for L is determined in ***Garcia et al.*** (***2013***) and the value for v in ***Eck et al.*** (***2020***).

since by definition, the category of cells being studied are not represented in any data. Therefore, this project uses MATLAB to simulate changes in f throughout the embryo. The goal was to use experimentally constrained parameters (such as promoter turn on time or gene length) to generate simulated data, and to use the statistical trends in this data to make inferences about unobservable behavior, such as variation in f. The MATLAB model of the embryo is not limited by microscope resolution, and is thus able to generate a more accurate picture of f at various AP positions. This approach allowed us to compare the simulated results for f with previously collected data. All data used as a comparison point in this project come from ***Garcia et al.*** (***2013***). Using this in combination with our model of the embryo, this method gives insight into the accuracy of f as a transcription metric.

Single Nucleus Model

Here, the P2 minimal enhancer and promoter of the *hunchback* gene during the 14th nuclear cycle of development were used to study changes in f. *hunchback* is transcribed in a step-like pattern along the anterior-posterior axis of the embryo.

The simulation of the nc 14 embryo functions by first modelling transcription of *hunchback* within a single nucleus. A simple toy model of transcription at the nuclear level was used for this purpose. To simulate single cell transcription, the model begins traversing the AP axis of the embryo, with the position of the cell currently being simulated given by the variable x. A given number of cells are simulated at every AP position, for which transcription is then simulated. Between the times t_{on} and t_{off}, when the promoter turns on and off, respectively, polymerase begin arriving at a rate of r molecules per minute. r is time dependent, with a mean value that differs for each cell based on AP position. The total value of r for each nucleus consists of this mean and time dependent fluctuations, which are discussed in more detail in Embryo Model. After a polymerase molecule binds to the gene, which has a given length L, it begins transcribing at an elongation rate v, which is held constant for each nucleus. Once the polymerase molecule reaches the end of the gene, it detaches immediately.

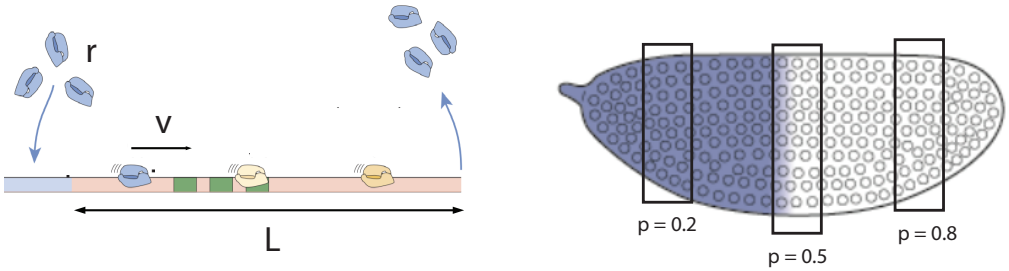

Figure 3. Left: Visualizing r, v, and L. Polymerase molecules arrive at the 5′ end of the gene at a rate r and transcribe the gene of length L at an elongation rate v. Right: Visualizing AP bins, tracked in the simulation by the variable p.

The gene being transcribed has 12 MS2 stem loops at its 5′ end, which the MCP-GFP protein then binds to. The total fluorescence is thus directly proportional to the number of polymerase transcribing the gene. This creates the rising edge in fluorescence shown in Figure 4, until the rate of polymerase arriving and leaving are equal, which is responsible for the plateau shape in the same figure. After the promoter turns off, and r, the rate of polymerase arrival labelled in 3 drops to zero, and the number of polymerase molecules on the gene begins to decrease. The corresponding decrease in fluorescence creates the falling edge that succeeds the plateau. The current version of the model only simulates this process until t_{off}, i.e, only the rising edge and plateau shape . This had a negligible effect on quantities such as average fluorescence and standard deviation of fluorescence for a large set of simulated traces, and simplified the model by not having to describe the more complicated falling edge (see Simulation Traces for details).

Embryo Model

In order to model the changing fraction of active nuclei, f, throughout the embryo, the modulation of transcription rates along the AP axis needs to be taken into account. This is controlled in part by the concentration of Bicoid, which determines the accumulation of *hunchback*, producing a decreasing fraction of active nuclei along the AP axis. Here, Bicoid is modelled with the decaying exponential

$$Bcd \sim e^{kx}. \tag{1}$$

The coefficient of the exponent, k, is equal to -2.67, and is derived from the measured decay rate of Bicoid, 23.5% per 10% x/L length AP bin (***Eck et al.*** (***2020***)).

The position dependent polymerase loading rate, r, is modelled with the equation

$$r = (r_{Max} - rMin) * \frac{([Bcd]/K_d)^5}{1 + ([Bcd]/K_d)^5} + r_{Min}. \tag{2}$$

Hill functions are often used to model the distribution of hunchback ***Gregor et al.*** (***2007***). Here, r_{Max} and r_{Min} are the maximum and minimum polymerase loading rate observed across all nuclei in ***Garcia et al.*** (***2013***), respectively. $[Bcd]$ is the position-dependent Bicoid concentration, as described above, and K_d is the dissociation constant, which was found by fitting Equation 2 to the ***Garcia et al.*** (***2013***) data showing the polymerase loading rate as a function of AP position (see Figure 5).

Variable Name	Value	Description
r_{Max}	20 mol/min	Maximum polymerase loading rate
r_{Min}	4 mol/min	Minimum polymerase loading rate
Bcd	see Equation 1	Bicoud concentration
K_d	4.247	Dissociation constant

Table 2. Simulation variables. r_{min} and r_{max} were determined in ***Garcia et al.*** (***2013***), Bcd was calculated using the decay rate determined in ***Eck et al.*** (***2020***), K_d was found by using a fit to polymerase loading rate data.

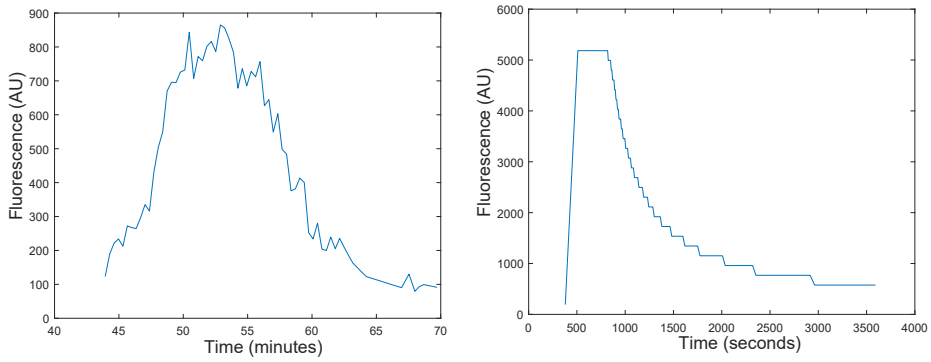

Figure 4. Left: Fluorescence over time recorded for a single nucleus in one of our datasets. Though these individual traces are noisy, the rising edge, plateau, and falling edge in the fluorescence are clearly visible. Right: A single nucleus trace produced by the simulation. It captures the same three features of f transcription as the trace on the right, though with far less noise.Note that in the final version of the simulation we ignore the data up to the beginning of the 'plateau'.

However, within each AP bin, there is still a noticeable amount of time dependent noise in the loading rate, which is not accounted for by the value calculated with Equation 2 (see Polymerase Loading Rate and S13 for examples of f as a function of position for different polymerase loading rates). The Hill equation sets the average mean loading rate for each AP bin, but each nucleus within an AP bin still has some variability in the loading rate (here, 'mean' refers to the time averaged loading rate, and 'average' to the mean across an AP bin). It is outside the scope of this project to calculate the exact dynamics of this transcription noise, so it is instead approximated with a Gaussian distribution of noise. Here the mean is the polymerase loading rate calculated with Equation 2. The standard deviation of the distribution models the nucleus-to-nucleus variability in the polymerase loading rate. This standard deviation is set to 25% of the mean, based heuristically on the shape of the data (see Polymerase Loading Rate for examples). This is a simplification of the actual noise in the loading rate, but provides a reasonable fit to the fraction on data.

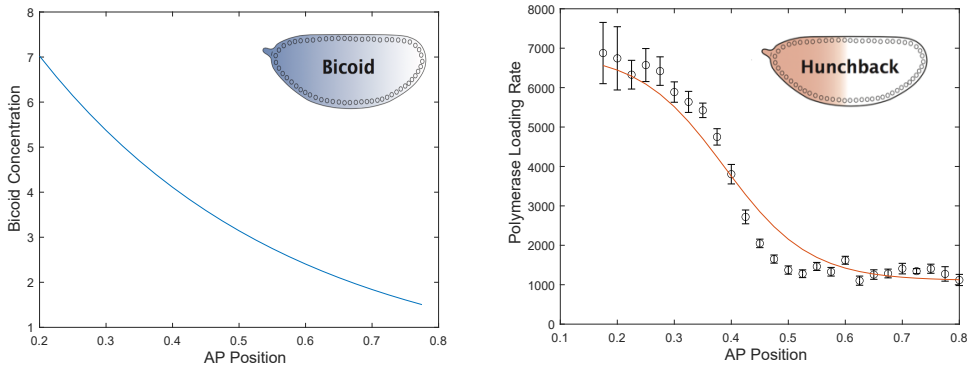

Figure 5. Left: Bicoid concentration as a function of AP position. Right: Polymerase loading rate as a function of AP position. Black data points are average loading rate at a given position, and the red line is a fit to the data using Equation 2.

Determining Fluorescence Cut Off

The model is now able to reproduce the modulation of *hunchback* transcription observed in the embryo, and the next step is thus to implement an artificial detection limit, dI, in the simulation. This approximates the detection limit of a microscope being used to take data, and allows the simulation to distinguish between active and visible, active and invisible, and inactive cells. This cut off is determined using the data in Figure S2 of ***Garcia et al.*** (***2013***), which shows the minimum fluorescence recorded for each nucleus in the data set. The mean of this set, equal to approximately 262.13 AU, is used as the detection limit in the simulation. This is implemented in the code as the fluorescence value which cells must produce a fluorescence output above in order to be recorded as 'on'. Any simulated cell that does not produce a fluorescence above this value for at least one time point is thus recorded as being 'off' (see Figure 7). By modelling the changing polymerase loading rate and implementing this limit, the simulation can now recreate the modulation of f (fraction on) along the AP axis.

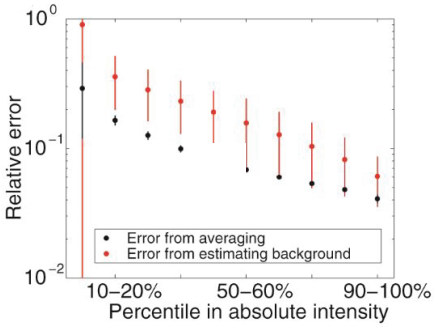

Figure 6. Figure S2G from ***Garcia et al.*** (***2013***). The error from estimating background (red) was used to estimate F_{noise}. See Including Fluorescence Noise and Adding Noise to Fluorescence Signal for details.

The microscope used to image the ***Garcia et al.*** (***2013***) data has a frame rate of 1 frame per 20 seconds. Thus cells must display fluorescence above the cut off for at least this long, in order to be observed at all. However, generally only cells that display fluorescence for multiple frames are considered to be active. Since there is no hard limit for how many frames are necessary, the limit is set to two frames in the simulation. As discussed in Frame Cut Off, within a reasonable range,

there are no significant differences between different frame cut offs.

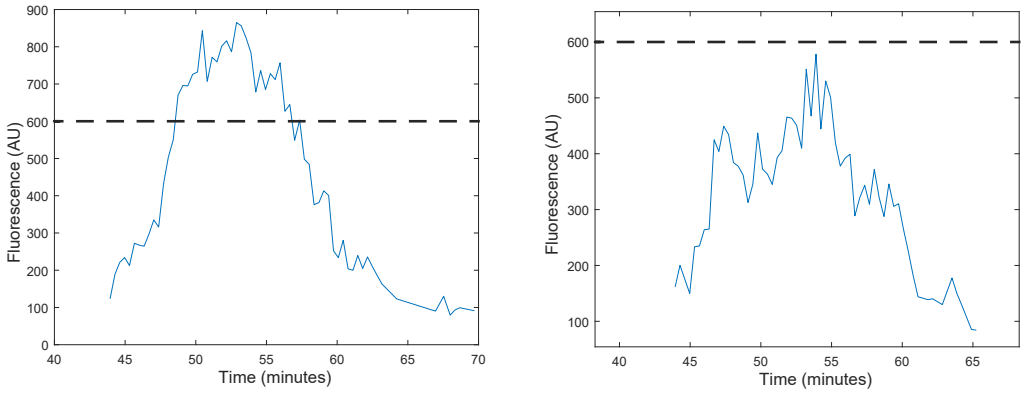

Figure 7. Visualization of the fluorescence cut off implementation. Here, a detection limit at 600 AU is shown. The trace on the left would produce an 'on' cell, since the fluorescence is above the cut off for multiple traces. The figure on the right is a trace from an 'off' cell, since the fluorescence is never above the detection limit. An individual nucleus must display fluorescence above the detection limit for at least 2 frames (~ 40 seconds) in order for the simulation to consider it an active cell. Note that the fluorescence produced does not need to be continuous, the nucleus only needs to exhibit it for a minimum of two frames.

Every time the simulation is run, the steps of the model are as follows:

1. The code stores its current AP position in the variable p and calculates the polymerase loading rate r for all of the cells at that position, using Equation 2.
2. Transcription at the singular nucleus level is simulated for each cell within the current AP bin. See Single Nucleus Model for a description of the model and a list of variables.
3. The model verifies whether each cell is 'on' by comparing its fluorescence output at every point in time to the detection limit dl
4. This process is repeated for all cells in the current bin in order to calculate f at that position
5. The model moves on to the next AP bin and repeats the process, which continues until the embryo has been traversed for all possible values of p

Results

Figure 8 shows the simulated f as a function of AP position in red, as well as the data recorded in ***Garcia et al.*** (***2013***), in blue. It is clearly visible that while simulation and data agree in the very anterior of the embryo (.2-.4 x/L), they begin to diverge towards the posterior. Notably, the simulation output is consistently larger than the recorded data, meaning that the model is expecting a higher fraction of active cells at every AP position than the data are. The shading in Figure 8 shows the implications of this - the area between the data and simulation curves represents the effects of the 'invisible' cells, that is, the simulation is expecting a higher output because it is recording the presence of these nuclei, while the data are not (light blue region in 8). Thus, the ***Garcia et al.*** (***2013***) data record a lower fraction of active cells than are actually present in the embryo. The model indicates a significant presence of 'invisible' cells, accounting for up to 50% of all active cells at its maximum, as shown in Figure 9. However, the fact that the simulation and data clearly follow the same general shape shows that the observed fraction on is not purely an artifact, and that there are truly 'off' cells.

The discrepancies between the simulation and data are also shown in Figure 9. Here the experimental fidelity, a metric of how well data and simulation, is defined. This quantity is calculated

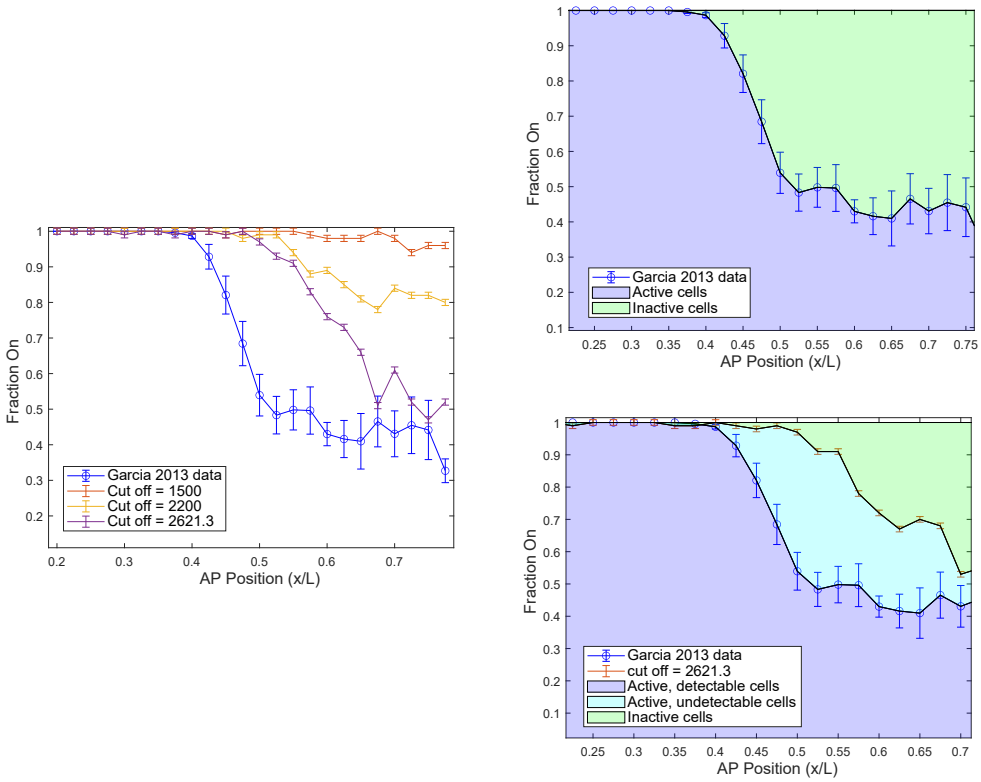

Figure 8. f as a function of position. In each subfigure, the blue line represents the measurements taken in ***Garcia et al.*** (***2013***). Left: the changing output for decreasing simulated detection limits in red, yellow and purple. As expected, when the limit is decreased, f increases. Top Right: shaded regions show the fraction of active and inactive cells as represented by the data set, where the plotted line separates the categories. Bottom right: categories of cells as separated by the lines representing f data and simulation output. The blue region between the two lines shows the fraction of invisible active cells. All error bars represent the standard error of the mean across 20 simulation runs. Here, we can see the differences between how the simulation and original data set categorize the data.

using $EF = \frac{f_{exp}}{f_{sim}}$, where f_{exp} is the fraction on value recorded in the ***Garcia et al.*** (***2013***) data, and f_{sim} is the fraction on value calculated by the simulation. This quantity shows what percentage of all actually active cells are captured by the ***Garcia et al.*** (***2013***) data set. For example, if the 2013 data records a value $f = 0.4$, and the simulation find that $f = 0.8$, then the experimental fidelity is 50%. At its minimum, the f value recorded in the data represents only 50% of all the active cells that the simulation is recording - that is, f, the fraction of all cells that are active in the ***Garcia et al.*** (***2013***) data, accounts for only half of all the active cells in the simulation.

Including Fluorescence Noise

These results can be refined by accounting for the effects of noise in the fluorescence signal itself (as well as the noise in the polymerase loading rate described in Embryo Model. This was done by using the data represented in 6. Specifically, the points showing error caused by estimating background were utilized. The relative error as a function of percentile shown in the figure is used to approximate the signal noise for each fluorescence measurement.

For a fluorescence measurement in a specific percentile, some value F_{Noise} is chosen from a normal distribution whose size is determined by the standard error shown in Figure 6. This is then

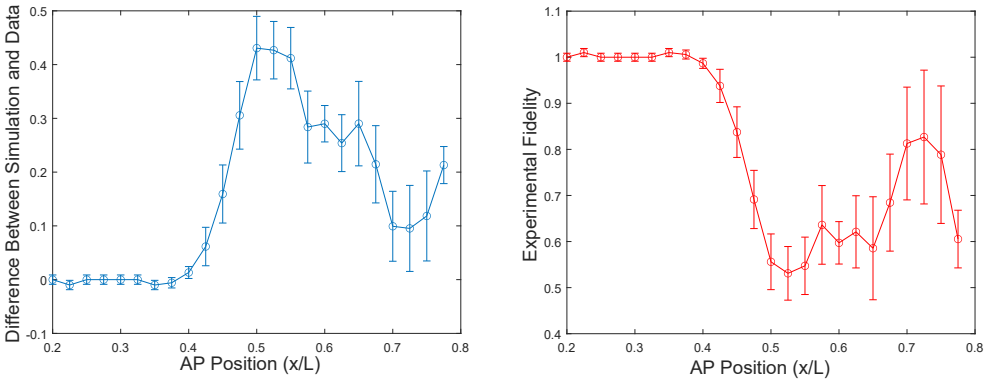

Figure 9. Left: Difference between simulation and data results for f as a function of AP position. In the very anterior, the data and simulation agree, but the difference between them peaks towards the middle of the AP axis. Right: Experimental fidelity as a function of AP position, where experimental fidelity is the fraction of all active cells that are visible in the ***Garcia et al.*** (***2013***) data. For example, if at AP Position .5, the data shows that .4 of all nuclei are active, and the simulation indicates .8 of all nuclei are active, the experimental fidelity is .5, i.e only 50% of the active cells are being represented in the data. All error bars represent the standard error of the mean across 20 simulation runs.

added to F_{Model}, the fluorescence noise calculated by the simulation before any noise is applied.

The total amount of noise, F_{Final} is thus given by

$$F_{Final} = F_{Model} + F_{Noise}.$$

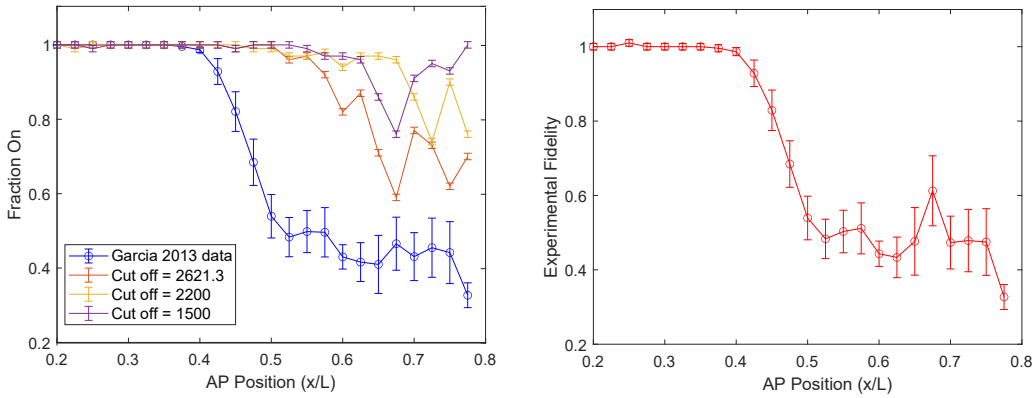

Figure 10. Left: f as a function of position, where f in this case includes the noise discussed in Including Fluorescence Noise. The changing output for decreasing simulated detection limits in red, yellow and purple, and ***Garcia et al.*** (***2013***) data in blue, when including fluorescence signal noise. All of the lines representing simulation output show a clear shift upward Right: Experimental fidelity as a function of AP position, for a cut off of 2621.3 AU, when fluorescence signal noise is accounted for. All error bars represent the standard error of the mean across 20 simulation runs.

F_{Final} thus takes into account the noise in the measurement of the signal as well as the fluorescence that the model calculates. See Adding Noise to Fluorescence Signal for further details.

As shown in Figure 10, when plotting f as a function of position, this extra noise increases the difference between simulation and data for all detection limits. Correspondingly, the experimental fidelity decreases at all points, reaching a minimum of 50%. Though this is not a very sophisticated

method for estimating noise, it indicates that increasing the noise in the simulation does not resolve discrepancies between simulation and data. Rather, it exacerbates them.

Discussion

Our results indicate that the pattern of *hunchback* concentration seen using MCP-GFP is not entirely an artifact of the detection limit - in other words, there is such a thing as a 'truly off' cell. However, a significant fraction of all cells transcribing *hunchback* go undetected in current measurements of f , with up to 50% of all active cells being missed when not accounting for signal noise. In other words, the model indicated the presence of active cells that are recorded as being off, due to their low fluorescence output. This value increases to 60% when this extra noise is added. This means that in order for f to be accurate metric of transcriptional activity, measurements need to account for this high percentage of invisible active cells.

Future improvements to the model could include incorporating observed changes in downstream cleavage of nascent transcripts as a function of position. After transcription, polymerases continue moving along the gene for some period of time, creating a 'waiting time' before they detach. This allows more polymerase than accounted for in the simulation to be present at any moment in time, with the exact number being dependent on the duration of the 'waiting time'. This value varies with AP position, and incorporating it into the simulation could thus change the f output at each point in the simulated embryo.

Currently, the simulation recreates the very specific dynamic of *hunchback* transcription at the single nucleus level. Another potential improvement is generalizing the simulation to model transcription of other genes, and checking whether the simulation still consistently outputs a higher f than the gene data. This generalization could also extend to modeling other nuclear cycles. In nc 12 and nc 13, for example, one would expect 100% f at all points in the embryo. This could be used to test the accuracy of the simulation by checking if it produces this same result.

Some adjustments to the simulation parameters can also be recreated with changes to experimental set up, in order to compare data and simulation output. For example, one could compare the results when taking data at full and half laser power, recreated in the simulation by simply halving the fluorescence constant.This should, in theory, just show half as many active cells. This and other set ups could provide further insight into the distribution and quantity of invisible active cells, as well as allowing us to test whether the simulation still follows the shape of the data at other laser powers.

Supplementary Information

S1 Simulation Parameters

Promoter Turn On Time (t_{on})

The simulation models the time that the promoter turns on varies using a Gaussian distribution. This distribution is supported by the data taken in ***Garcia et al.*** (***2013***) (see Figure S11). The mean and standard deviation of the distribution used in the simulation were taken from this data set as well. The data sets show similar coefficients of variation, supporting the fact that the normal distribution is an accurate model of turn on time.

Promoter Turn Off Time (t_{off})

The promoter turn off time was held constant at all AP positions within the embryo. This is a simplification, since it has been observed that the turn off time does in fact vary with position (see Figure S4E in ***Garcia et al.*** (***2013***)). However, when these changes were modelled, using the data

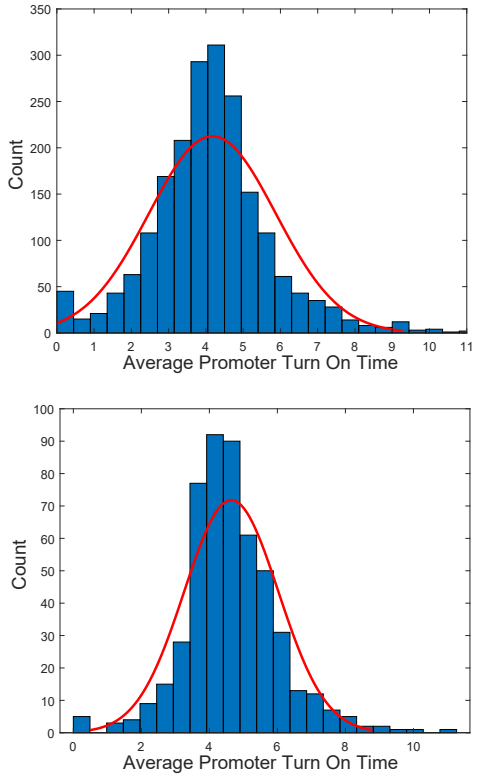

Figure S11. Average turn on time for two different data sets within ***Garcia et al.*** (***2013***). Both experiments show similar mean turn on time, coefficient of variation, and distribution, indicating that a Gaussian model of turn on time variation is appropriate.

in ***Garcia et al.*** (***2013***), there was no significant difference in the results for the f as a function of position, as shown in Figure S12. Correspondingly, there are also no significant changes in the experimental fidelity as a function of position. Therefore, the promoter turn off time was modelled with a constant value, for simplicity.

Gene Length and Elongation Rate (L & v)

The length of the *hunchback* gene being transcribed as well as the elongation rate are also held constant in all parts of the simulation. AP position dependent changes in elongation rate were not taken into account here, but this is an avenue for future research.

Polymerase Loading Rate

The polymerase loading rate is modelled using the Hill function below:

$$r = (r_{Max} - r_{Min}) * \frac{(Bcd/Kd)^5}{1 + (Bcd/Kd)^5} + r_{Min} \tag{S3}$$

A power five Hill function is used here. The resulting graph is show in Figure S13.

However, the polymerase loading rate also includes some amount of time-dependent genetic noise within each AP bin (see Figure S13). Modelling this noise is outside the scope of this project, so a Gaussian model of the noise was instead used as an approximation. The mean of this model is r (the results of equation 2), and the standard deviation is a percentage of this result, determined by the shape of the data. A standard deviation of 25% of r is used in the final version of the simulation (see S13B).

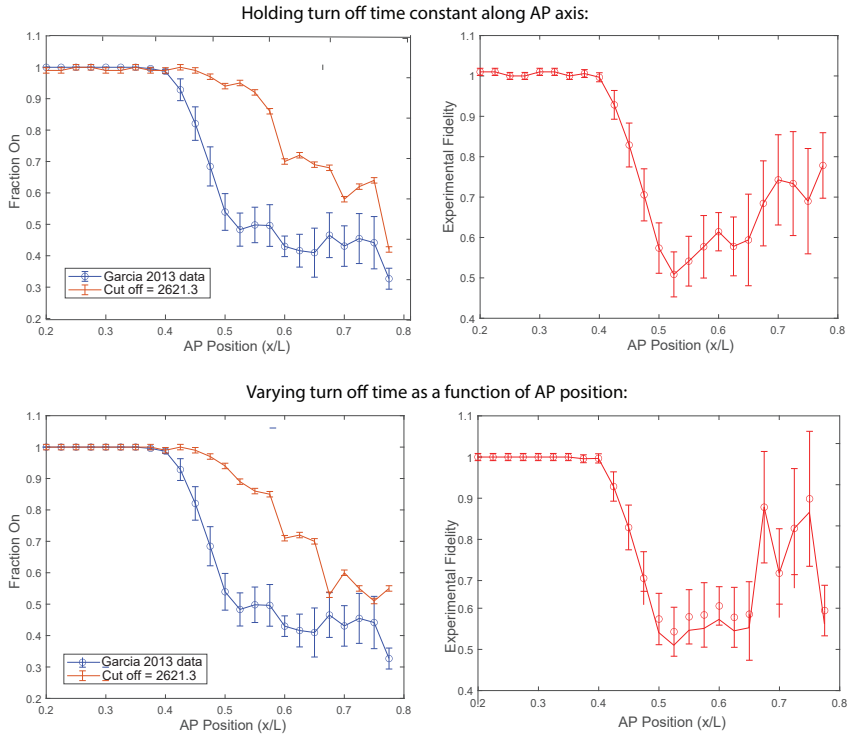

Figure S12. Top: Fraction on as a function of position and the corresponding experimental fidelity, when t_{off} is held constant along the AP axis. The corresponding experimental fidelity is shown on the top right. Bottom: Model output for f as a function of AP position, with data for comparison, when changes in turn off time with position are accounted for. There is no significant difference between this result and the results found when not considering variation in turn off time. Bottom: Experimental fidelity as a function of position when accounting for changes in t_{off}. The right graph shows the corresponding experimental fidelity.

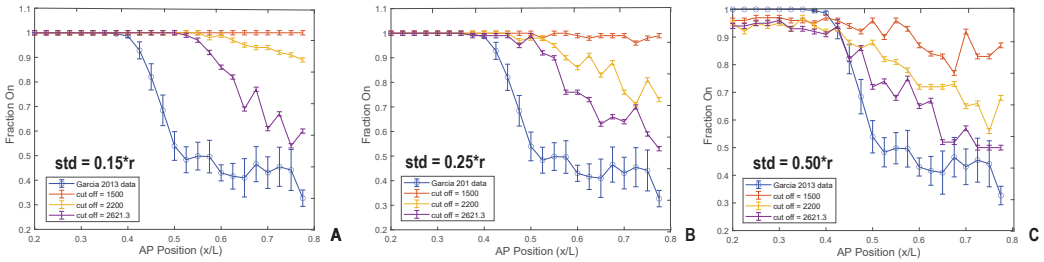

Figure S13. Fraction on for different amounts of noise in the polymerase loading rate. A shows a standard deviation $0.15 * r$ of the loading rate, where r is the value calculated using S3. B has a standard distribution of $0.25 * r$, and C $0.5 * r$. Standard deviations within the $0.2 * r - .25 * r$ range best followed the shape of the data. At higher values, the simulated curves begin to dip below the ***Garcia et al.*** (***2013***) data curve, a situation which does not have any meaningful interpretations.

Bicoid Concentration

The decay of the concentration of Bicoid along the AP-Axis is modelled with the equation

$$Bcd = 12 * e^{-2.679x} \tag{S4}$$

The coefficient of the exponent (-2.67) is derived from the measured decay rate of Bicoid, 23.5%

per AP bin, where each bin is 10% the length of the embryo. The coefficient (12) is an arbitrary value found through a fit to the data.

Maximum and Minimum Polymerase Loading Rate

The maximum and minimum polymerase loading rate values were determined from the data set shown in Figure 4a in ***Garcia et al.*** (***2013***), in molecules/minute.

Dissociation Constant

The dissociation constant was found through a fit to the polymerase loading rate data as a function of AP position from ***Garcia et al.*** (***2013***) (see Figure 5). Equation 2 was fitted to this data in order to find the best fit value for variable *Kd*.

S2 Frame Cut Off

In order for the simulation to categorize a nucleus as active, it must produce fluorescence above the detection limit for an extended period of time. Here, this time frame is set equal to two frames (60 seconds) of fluorescence, which is not necessarily continuous. This time frame is based on convention in analyzing GFP data. The time limit is implemented to ensure that cells that may make it above the detection limit for only a few seconds, due to noise, were not included in the final set of active of cells.

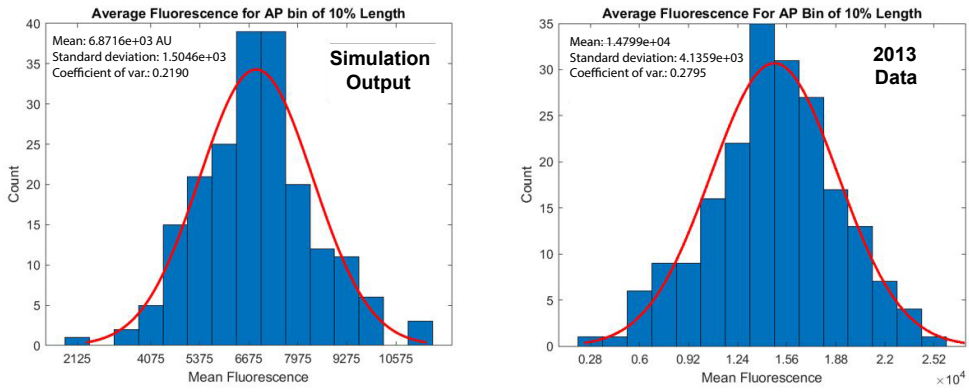

Figure S14. Histograms of average fluorescence of single nucleus traces, the left showing averages of simulation traces, and the right showing averages of actual data traces. The two sets have similar coefficients of variation (< .1 apart), indicating that the simulation, despite its lack of noise at the single cell level, still manages to capture the important features of single cell level transcription.

S3 Simulation Traces

Clearly, the simulated traces showing fluorescence produced by a nucleus as a function of time, include far less noise than their experimental counterparts (see Figure 4. However, the noisiness of the data is not that relevant to the features that this study focused on. Unless the trace is unusually noisy, the noise is unlikely to push the fluorescence above the detection limit. Therefore, capturing the turn on and off times, as well as the polymerase loading rate, is more important than the effects of the noise.

This is supported by the fact that the coefficient of variation of the the histograms showing the average fluorescence of the simulation and the data were very close, with a difference of less than 0.05 (see Figure S14).

Furthermore, the simulation only utilizes the rising edge and plateau of the simulation, i.e, only the fluorescence until the promoter turns off. Based on the traces in the data set, these 'half-traces' are still representative of the full sample. Histograms of the average fluorescence for only rising edge curves and full fluorescence curves have very similar means and coefficients of variation (see Figure S15).

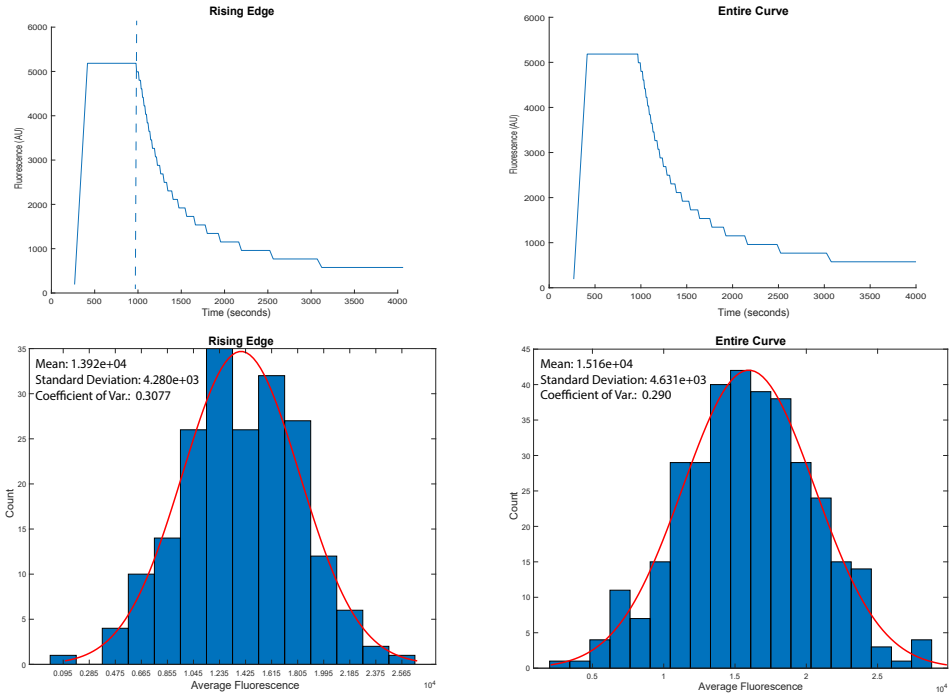

Figure S15. Left: Histogram of average fluorescence for the 'rising edge' of the single nucleus fluorescence traces (only the increase in fluorescence and plateau, excluding the decrease after the promoter turns off), shown above. Right: Histogram of average fluorescence for full single nucleus fluorescence traces (shown above). The data sets show similar mean, distributions, and coefficients of variation.

S4 Nuclei per AP Bin

During nc 14, there are 20 cells in each AP bin of 10% of the embryo length. As an idealization, each simulated AP bin contains 100 cells. Figure S16 shows the effects of increasing the number of cells per bin past their true value - though the noise decreases with the increasing cells number, the shape of the data, and the relationship between the data and simulation output remain largely the same. Thus, the idealized version with 100 cells was used in order to have access to a larger data set.

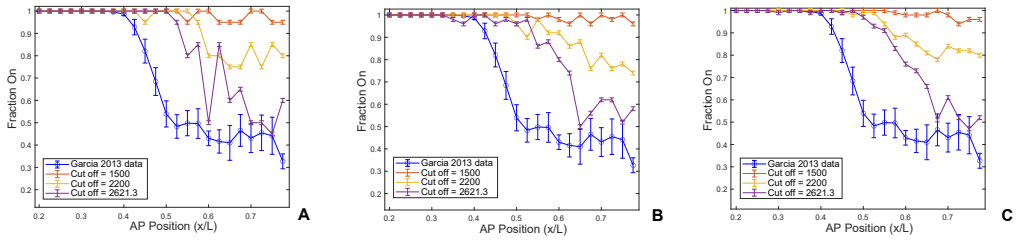

Figure S16. f as a function of position for three different simulated detection limits and a data set for comparison (blue). Figure A simulates 20 nuclei per AP bin, B 50 nuclei per AP bin, and C 100, the number used in the simulation to create the figures shown in the Results section. Apart from the amount of noise in the curves, there were no significant differences between the different figures.

S5 Error Bars on f

To calculate the error bars in Figure 8 the simulation was allowed to run 20 times, with only 20 simulated cells per AP bin (see Nuclei per AP Bin for further details). The standard error of this data set was then used to generate the error bars for the final results (the idealized version, using 100 cells per AP bin).

S6 Adding Noise to Fluorescence Signal

To simulate the noise in the fluorescence figure signal itself, the data showing relative error in fluorescence measurement as a function of percentile from ***Garcia et al.*** (***2013***) was used. This is shown in Figure S2G of the same paper (see Figure 6).

The red points in Figure 6 are measurements of error in fluorescence measurement. These errors were used to here to simulate noise in the fluorescence signal created by the simulation.

In Figure 6, the fluorescence error measurements are shown as a function of percentile in absolute intensity, so the fluorescence noise produced by the simulation is also dependent on the percentile of the fluorescence a given nucleus is producing.

The percentiles of fluorescence are fairly consistent across different data sets and simulation runs (for example, a certain fluorescence in AU corresponds to roughly the 30^{th} percentile every time the simulation is run). Thus, in order to cut down computing time, the simulation groups fluorescence values into percentiles before hand, based on previous results.

The noise is then applied after the fluorescence is calculated, based on this estimated percentile, rather than the simulation grouping the data into percentiles every time it is run. How much noise should be applied to each fluorescence value is based on the data in 6 - the 'Relative Error' on the y axis is used to calculate a distribution around the mean fluorescence, which is the value determined by the model before it considers noise. A new fluorescence value which accounts for noise is then chosen from this normal distribution round the average.

For example, assume fluorescence for a given nucleus at a given time point is F_{Model}. If F is between the 40^{th} and 50^{th} percentile, it has a relative error σ of ~ 0.23, according to Figure 6. The model then defines a quantity $\mu = \sigma \times F$, which determines the distribution of noise around the mean value. The model uses the MATLAB function normrnd to pick a value in a normal distribution around 0, with standard deviation μ. This value is then added to the average (the fluorescence calculated pre-noise), to model a noisy fluorescence measurement. Thus, the equation from Including Fluorescence Noise,

$$F_{Final} = F_{Model} + F_{Noise}, \tag{S5}$$

becomes

$$F_{Final} = F_{Model} + normrnd(0, \mu) \tag{S6}$$

References

Eck, E., Liu, J., Kazemzadeh-Atoufi, M., Ghoreishi, S., Blythe, S., and Garcia, H. (2020). Quantitative dissection of transcription in development yields evidence for transcription factor-driven chromatin accessibility.

Garcia, H. G., Tikhonov, M., Lin, A., and Gregor, T. (2013). Quantitative imaging of transcription in living Drosophila embryos links polymerase activity to patterning. *Current biology : CB*, 23(21):2140–2145.

Gregor, T., Tank, D. W., Wieschaus, E. F., and Bialek, W. (2007). Probing the Limits to Positional Information. *Cell*, 130(1):153–164.

Lucas, T., Ferraro, T., Roelens, B., De Las Heras Chanes, J., Walczak, A. M., Coppey, M., and Dostatni, N. (2013). Live imaging of bicoid-dependent transcription in Drosophila embryos. *Current biology : CB*, 23(21):2135–2139.