# FLUCTUATIONS AND RESPONSE IN COMPLEX BIOLOGICAL SYSTEMS: WATCHING STOCHASTIC EVOLUTIONARY AND ECOLOGICAL PATTERN DYNAMICS

BY

K. MICHAEL MARTINI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Assistant Professor Thomas Kuhlman, Chair
Professor Nigel Goldenfeld, Director of Research
Professor Karin Dahmen
Professor Kenneth Schweizer

# Abstract

My research uses computational and analytical techniques from statistical physics to examine spatial patterns and dynamics in complex biological systems. More specifically I used these techniques to analyze aspects of three different complex biological systems, namely stochastic Turing patterns, transposon and retrotransposon dynamics in live cells, and bistability in ant foraging.

In collaboration with experimentalists at MIT and UIUC, I have shown how noise can stabilize emergent behaviors such as Turing patterns in biofilms. Normally, one would think that noise destroys patterns but we found that fluctuations in the copy number of signaling molecules acting as activator and inhibitors of gene expression leads to pattern formation. Surprisingly, we can show theoretically that these fluctuations increase the range of experimental conditions in which patterns can form.

In collaboration with experimentalists at UIUC, we have observed how evolution acts on variation in time, space, and genome locus by imaging live cells with fluorescent reporters that allow us to track transposon dynamics. Transposons, also known as "jumping genes," are found in all organisms and have activity that can cause mutations and drive evolution. As part of this collaboration I developed the software for image analysis of the cells and analyzed the resulting statistics of events. We discovered that the excision rate of transposons depends on orientation of the element, spatial location of the cell, and some heritable factors.

In a follow-up experiment, I recently developed a model to explain our collaborators' observation that the number of retrotransposon transcripts, transcripts produced by a copy and paste type of mobile genetic element, produces an exponential growth dependence defect. I developed a model for the copy number dynamics of retroelements and the time it takes these elements to be lost from a population of cells depending on the observed growth rate defect, transposition rate, and inactivation rate. This model explains why Group II introns are present in about 30% of bacterial species, while retrotransposons are essentially absent. This research sheds light on the early evolution of the eukaryotic spliceosome, the cellular machinery allowing complex organisms to remove intra-gene junk DNA during gene expression.

I have extended a model for ants foraging from two food sources [1] to include indirect recruitment of ants with pheromones rather than direct recruitment by the ants themselves. This model continues to show bistable foraging for

ants when their population is below a critical population size that depends on the deposition rate and evaporation rate of pheromones.

*For everyone who believed in me.*

# Acknowledgments

I would like to first and foremost thank my advisor Nigel Goldenfeld. Nigel has been instrumental in guiding me as a young scholar. He is responsible for helping me develop many of the ideas and research directions found in this thesis. He has been a huge help in getting me unstuck when I was stuck and was a great resource for cool new ideas and interesting projects. Nigel has helped facilitate many excellent opportunities for me, including being able to participate in the $66^{\text{th}}$ Lindau Nobel Laureate meeting, teaching at the CPLC summer school for biophysics, and attending various conferences. I would like to thank Nigel for all his aid in my post-doc applications. Nigel not only cares for his students academic growth, he also cares about his students personal well-being. He runs a wonderful group; during our often six-hour weekly meeting, group members teach each other about their research interests and support one another in our respective projects.

I would like to thank all of Nigel's former and current group members with whom I have interacted, including: Vikyath Rao, Farshid Jafarpour, Chi Xue, Hong-Yan Shih, Maksim Sipos, Tommaso Biancalani, Purba Chatterjee, Minhui Zhu, and Zhiru Liu. I would especially like to thank Vikyath and Farshid for being good friends, and often eating our lunch together. In addition, I would like to thank Minhui and Zhiru for giving me the opportunity to learn how to be a mentor, and for all their hard work on other projects, including colony size scaling laws and segmenting pictures of corn roots.

I would like to thank Tom Kuhlman as an excellent collaborator and mentor, whose patient explanation of basic biology and experimental details was very helpful. Meeting with him always provided insight. Tom also provided me the opportunity to be his teaching assistant for Physics 435, and I learned a lot about teaching from this experience. In addition, I would like to thank the members of Tom's group with whom I have had a wonderful opportunity to interact and discuss ideas for our collective research, including Nicholas Sherer, Neil Kim, Gloria Lee, and Davneet Kaur. Additionally, I would like to thank Ron Weiss for the use of his experimental Turing data and his collaboration. I also thank Andrew Leakey for his collaboration and insights on the segmentation of pictures of plant roots. In addition, I would like to thank all the members of my committee; Karin Dahmen, Kenneth Schweizer, Tom Kuhlman, and Nigel Goldenfeld.

I am appreciative of the funding support that I have received. It enabled me to pursue my research goals. I have

received funding from the National Science Foundation through the Center for Physics of Living Cells (PHY 1430124) and I also received funding from Andrew Leakey's and Nigel Goldenfeld's grant on maize root structure, NSF IOS 1638507.

I would like to thank the support that I have received over the years from my family. They have kept me grounded and have been there for me whenever I needed them. They were a source of very important emotional support and over the years have provided me with countless opportunities.

I would also like to thank the professors at my undergraduate university, RIT, who helped develop me into a young researcher. I would especially like to thank George Thurston and Dawn Hollenbeck, my undergrad co-research advisors. They spent a lot of time with me, gave me opportunities to work on their research, and cared about my personal and academic growth.

I would also like to thank all the other people who have helped me over the years and that I have neglected to mention directly in this thesis. Just know that I deeply appreciate everything you have done for me.

# Table of Contents

# List of Abbreviations

| | |
|---|---|
| TE | Transposable element |
| L1 | LINE-1 |
| *E. coli* | *Ecscheria coli* |
| *B. subtilis* | *Bacillus subtilis* |
| *P. aeruginosa* | *Pseudomonas aeruginosa* |
| IPTG | isopropyl $\beta$-D-1 thiogalactopyranoside |
| GFP | Green flourescent proteins |
| RFP | Red flourescent proteins |
| 2DFT | Two-dimensional Fourier transform |
| 2D | Two-dimensional |
| ISLEAD | Imperfect palindromic sequences encoded in the leading strand |
| ISLAG | Imperfect palindromic sequences encoded in the lagging strand |
| LE IP | Left end imperfect palindromic sequence |
| RE IP | Right end imperfect palindromic sequence |
| HILO | Highly inclined and laminated optical sheet |
| AIC | Akaike Information Criterion |
| NHEJ | non-homologous end joining |
| aTc | anhydrotetracycline |
| TPRT | Target-primed reverse transcription |

# Chapter 1

# Introduction

My research uses computational and analytical techniques from statistical physics to examine spatial patterns and dynamics in complex biological systems. Using these computational and analytical techniques I examined Turing patterns in biofilms, tracked live cell transposon and retrotransposon dynamics, and bistability in ant foraging.

In Part I, Stochastic Turing Patterns, I present my research on Turing patterns in biofilms. In collaboration with experimentalists at MIT and UIUC (Professor Ron Weiss' group) I have shown how noise can stabilize emergent behaviors such as Turing patterns in biofilms. Normally, one would think that noise destroys patterns but we found that fluctuations in the copy number of signaling molecules acting as activator and inhibitors of gene expression leads to pattern formation. Surprisingly, we can show theoretically that these fluctuations increase the range of experimental conditions in which patterns can form.

In Part II, Transposable Elements, I present work performed in collaboration with experimentalists at UIUC (Professor Thomas Kuhlman's group). We have observed how evolution acts on variation in time, space, and genome locus by imaging live cells with fluorescent reporters that allow us to track transposon dynamics. Transposons, also known as "jumping genes," are found in all organisms and their activity can cause mutations and drive evolution. As part of this collaboration I developed the software for image-analysis of the cells and analyzed the resulting statistics of events. We discovered that the excision rate of transposons depends on orientation of the element, spatial location of the cell, and some heritable factors.

In a follow-up experiment, discussed in Chapter 6 of part II, I recently developed a model to explain our collaborators' observation that the number of retrotransposon transcripts, transcripts produced by a copy and paste type of mobile genetic element, gives rise to an exponential growth dependence defect. I developed a model for the copy number dynamics of retroelements and the time it takes these elements to be lost from a population of cells depending on the observed growth rate defect, transposition rate, and inactivation rate. This model explains why Group II introns are present in about 30% of bacterial species, while retrotransposons are essentially absent. This research sheds light on the early evolution of the eukaryotic spliceosome, the cellular machinery allowing complex organisms to remove intra-gene junk DNA during gene expression.

In Chapter 7 I have extended a model for ants foraging from two food sources [1] to include indirect recruitment

of ants with pheromones rather than direct recruitment by the ants themselves. This model continues to show bistable foraging for ants when their population is below a critical population size that depends on the deposition rate and evaporation rate of pheromones.

## 1.1 Stochastic Turing Patterns

In his paper, "The Chemical Basis of Morphogenesis," [2] Turing showed how a periodic pattern instability can emerge from an initially uniform activator inhibitor reaction diffusion system. In this picture the activator activates its own production and that of the inhibitor while the inhibitor inhibits its own production and that of the activator. In an initially homogeneous state the activator will amplify any small perturbation creating more activator and inhibitor locally. The inhibitor then diffuses more quickly than the activator suppressing further growth. In this way a periodic pattern will form. Note that in the traditional analysis this requires the two morphogens to have very different diffusion rates for pattern formation. On the other hand, in a stochastic model using the intrinsic noise from the birth and death processes, it turns out that the diffusion rates do not need to be so widely different [3, 4, 5].

In Chapter 3 of Part I, I analyzed the data from an experiment that was conducted at MIT where our collaborators attempted to engineer a Turing pattern. From my analysis I was able to show that the patterns they formed were actually stochastic Turing patterns rather than traditional Turing patterns. I did this by measuring the power spectrum of the pattern and showed it was consistent with theoretical predictions for the power spectrum of a stochastic Turing pattern. In addition, I analyzed and simulated a detailed model of their system and showed for the measured parameters of their system the model can only produce patterns if the stochasticity of the birth and death processes were included. Because the precise values of the model parameters are not known, I also mapped how much of parameter space would produce stochastic patterns as compared to deterministic Turing patterns, and showed that it was most probable that the experiment was in the stochastic Turing pattern regime.

## 1.2 Transposable Elements

Transposable elements (TE), more colloquially known as jumping genes, are DNA sequences that can move their position around the genome. There are two main types of transposable elements, DNA transposons which use a cut and paste mechanism of transposition and retrotransposons which use a copy and paste mechanism of transposition. Transposable elements make up a large fraction of eukaryotic genomes. For example, roughly 85% of maize's genome is TE and 46% of the human genome consists of TE. Transposable elements, through their activity, generate mutations in the genome and thus are important for understanding disease, development, and evolution.

### 1.2.1 DNA Transposons

Little is known about the dynamics of TE elements in live cells. Previous works [6, 7, 8] inferred transposition rates from bulk sampling of cells, which averages over many cells and loses information about fluctuations. Others [9, 10] attempted to measure the rates from phylogenetic comparisons, but this method suffers from the limitation that only events that have become fixed in a population can be observed. This misses events that could cause extinctions and the corresponding estimates of rates will likely underestimate the rate of transposition.

To overcome these limitations our experimental collaborators use a transposon to interrupt a promotor for the expression of mCerulean. When the transposon excises it will produce a full promotor and the cell will start to express mCerulean and glow blue. Additionally, the protein that is responsible for excision of the transposon, known as transposase, is tagged with another fluorescing protein that glows yellow. By observing the fluorescence of cells the amount of transposase can be quantified and it is possible to determine if a transposition event has occurred.

As discussed in Chapter 5, I developed image analysis software to automatically detect when and where transposition had occurred. Using this software we were able to extract rates of excision of $6.3 \times 10^{-3}$ events/cell/hr. Furthermore, we were able to test if excision was uniform in time and space (See Chapter 5). We found that the rate of transposition is growth state dependent. Initially, no events were detected until growth arrest. Events were uniform in time upon their initiation in growth arrest for 35 hours and had Poisson statistics. Furthermore, we found excision events to be clustered in space as shown by an excess in the radial correlation function within a few cell lengths as compared to simulations assuming a completely uniform event distribution. This clustering of excision events suggests that there may be a heritable change that effects excision rate. To test this hypothesis we measured the distribution of event rates from 984 colonies. We found that the resulting distribution was well fit by a two-step process: first, a heritable change can occur during exponential growth that predisposes cells to TE activity; then, upon growth arrest the cells containing this change have a probability of their TEs excising.

The growth state dependence, spatial clustering, and heritability of TE excision suggests that mutations caused from reintegration will also be heterogeneous and growth state dependent. This is potentially important since many models of mutation and evolution start with the assumptions of uniform and homogeneous mutation rate.

### 1.2.2 Retrotransposons

Retrotransposons are abundant in Eukaryotes but are rare in bacteria. In Eukaryotes, retroelements exist in high copy number, while in bacteria a simpler form of retroelement known as a group II introns can be present, but only in low copy number and only in 30% of bacterial species. For example, in humans the retrotansposon LINE-1 (L1) makes up 17% of the human genome, with about 500,000 integrants and roughly 100 active copies, whereas group II introns in bacteria typically have only 1-10 copies. To try to characterize some of the differences between bacteria

group II introns and Eukaryotic retrotransposons our experimental collaborators succeeded in transplanting a human L1 into a bacterial host *Escherichia coli* (*E. coli*). They observed that L1 expression is detrimental to the growth of *E. coli* and *Bacillus subtilis*. The growth rates of these bacteria were exponentially depressed with additional copies of L1 transcript. I modeled this in Chapter 6 using a simple binary growth model where each transcript has a certain probability of integrating and disrupting the cell's ability to grow. Thus, the probability that a cell will be able to grow is the binomial distribution with zero negative integration events. This simple model produces an exponential growth defect. Our experimental collaborators measured the growth defect for both bacteria and measured the growth defect for group II introns.

I also developed a model for the copy number of retroelements given a measured birth defect, transposition rate per transposon, inactivation rate of the retroelements, and death rate of bacteria. This more complicated model predicts that the measured growth defect of L1 in bacteria will cause the bacteria to quickly lose L1, matching results of the experiment. It also predicts that for the measured growth defect of group II introns, they will persist in low copy numbers for at least millions of generations, consistent with the observations of group II introns in bacteria. Finally, it predicts that for L1 to persist in human populations at high copy number the growth defect must be very small. This may be achieved in Eukaryotes by the spliceosome which limits genetic damage caused by integrants.

In summary, this project suggests that the spliceosome in Eukaryotes may have evolved in response to selection pressure from retroelements. In particular, it is consistent with phylogenetic evidence that shows how group II intron proteins were early predecessors of eukaryotic spliceosomal proteins, suggesting that the spliceosome was transmitted to Eukaryotes by an early horizontal gene transfer from bacteria [11, 12, 13, 14].

## 1.3 Stochastic Switching in Ant Foraging

Bistability is usually modeled using a double well potential and simple white noise. There is, however, an alternative mechanism for achieving bistability, a simple harmonic potential and multiplicative noise. The noise is greatest at the bottom of the well and vanishes at the boundaries of the well. The characteristic equation for this type of bistability is $\dot{z} = -z + s\sqrt{1-z^2}\eta$, where $\eta$ is Gaussian noise, z is the bistable quantity and $s$ controls the strength of the noise. In a recent paper it was shown that ants foraging from two food sources that directly recruit one another exhibit this type of bistability[1]. At small population sizes the ants will forage bistably from the different food sources, but as the population size is increased they will start to forage from both food sources equally. In Chapter 7 I have taken this model and extended it to include indirect recruitment via a pheromone. I investigate how the critical population size depends on the evaporation rate of the pheromones and the rate at which the pheromones are created. The conclusion is that the stochastic switch in foraging is robust to model elaboration, suggesting that these predictions could potentially

be experimentally tested.

## 1.4 Contributions and Publications

All the work done in this thesis was in close collaboration with my advisor Nigel Goldenfeld.

The work appearing in chapter 3 of Part I is the result of a collaboration between MIT and UIUC. The people involved in this collaboration included David Karig, Ting Lu, Nicholas A. DeLateure, Nigel Goldenfeld, and Ron Weiss. This work is in the process of being published in PNAS and this chapter is a modified version of that paper. The experiments and experimental design were developed by David Karig and Ting Lu. I performed the stochastic simulations and measured pattern characteristics of both the simulations and experiments, including spot size, power spectrum, and minimum distance between spots. I developed the phase diagram and sensitivity analysis of parameters for the detailed stochastic model. I also developed the phase diagram for the reduced model. For Chapter 3 of Part I, the majority of my contributions appear in sections 3.3 and 3.4.

The Work in Part II represents a collaboration with Thomas Kuhlman's lab.

Chapter 5 is a modified version of Real Time Transposable Element Activity in Individual Live Cells [15], a collaboration between Neil H. Kim, Gloria Lee, Nicholas A. Sherer, K. Michael Martini, Nigel Goldenfeld, and Thomas E. Kuhlman. For this chapter I contributed the software that detected the transposition events, calculated the radial correlation function for the experimental data, calculated the average intensity profile of an excision event, and wrote a simulation to compare to the measured $g(r)$ and event rate distribution. The majority of my contributions appear in section 5.5 of this chapter.

The work done in chapter 6 represents a collaboration between Gloria Lee, Nicholas A. Sherer, Neil H. Kim, Ema Rajic, Davneet Kaur, Niko Urriola, K. Michael Martini, Chi Xue, Nigel Goldenfeld, and Thomas E. Kuhlman. I developed the models in this section in collaboration with Chi Xue and Thomas Kuhlman. The analytic calculations appearing in 6.6, 6.7.1 are mine as are the simulations. I developed the model in 6.7.2 and conducted some of the simulations.

All of the work in chapter 7 is my own.

# Part I

# Stochastic Turing Patterns

# Chapter 2

# Introduction on Pattern Forming Systems and Stochastic Calculation Techniques

## 2.1 Linear Stability Analysis and the Turing Mechanism

Alan Turing, in 1952, made the surprising observation that in the right circumstances diffusion can act to destabilize an initially homogeneous state into a patterned state [2]. This is now known as the Turing instability. One of the simplest examples of a classical Turing pattern is an activator-inhibitor system. In this classical reaction diffusion system, one of the chemicals is a slowly diffusing activator, activating the synthesis of itself and the synthesis of the inhibitor. The other chemical is a fast diffusing inhibitor, inhibiting synthesis of the activator and itself. Initially the activator and inhibitor are distributed randomly. Areas with local concentrations of activator will autocatalytically grow, forming dense clumps of activator. Inhibitor will also be produced near these clumps of activator and will rapidly diffuse outward suppressing the further spread of activator.

In what follows I will describe the process of linear stability analysis and show that at certain wave numbers an initially homogeneous state becomes unstable. To serve as an example of an activator and inhibitor model I will use the Levin-Segel model of herbivore-plankton interaction [16].

$$\partial_t \phi = \mu \nabla^2 \phi + b\phi + e\phi^2 - p\psi\phi \tag{2.1}$$

$$\partial_t \psi = \nu \nabla^2 \psi + p\psi\phi - d\psi^2 \tag{2.2}$$

Here $\phi$ is the concentration of activator, $\psi$ is the concentration of inhibitor, $\mu$ and $\nu$ are the the diffusion constants of the activator and inhibitor, respectively. The term $p\psi\phi$ is a competition term, $e\phi^2$ is a nonlinear activation term, and $-d\psi^2$ is a nonlinear self inhibition term for the inhibitor.

In linear stability analysis, the growth or decay of an infinitesimal perturbation away from an equations' fixed points $(\phi^*, \psi^*)$ is examined. A fixed point is a place in the dynamics where the concentrations do not change. That is $\partial_t \phi^* = 0$ and $\partial_t \phi^* = 0$. Specifically, we want to consider an initially homogeneous solution for these equations. We need our initial state to be constant and not vary spatially. For the Levin-Segel model this involves solving the system of equations:

$$0 = b\phi^* + e(\phi^*)^2 - p\psi^*\phi^* \tag{2.3}$$

$$0 = p\psi^*\phi^* - d(\psi^*)^2 \tag{2.4}$$

The solution of these equations show that there is a coexistence point at $\psi^* = \frac{pb}{P^2-de}$ and $\phi^* = \frac{bd}{p^2-de}$. Since the population densities must be positive, this imposes the condition $p^2 > de$.

A fixed point is considered linearly unstable if the perturbation grows and linear stable if the perturbation shrinks. In linear stability analysis we choose a perturbation of the form $\delta\phi e^{\sigma t - ikx}$ and $\delta\psi e^{\sigma t - ikx}$. Plugging in $\phi = \phi^* + \delta\phi e^{\sigma t - ikx}$ and $\psi = \psi^* + \delta\psi e^{\sigma t - ikx}$ into eq. 2.1 and only keeping terms linear in $\delta\phi$ and in $\delta\psi$ we find that

$$\sigma\delta\phi = -\mu k^2\delta\phi + (b + 2e\phi^* - p\psi^*)\delta\phi + (-p\phi^*)\delta\psi \tag{2.5}$$

$$\sigma\delta\psi = -\nu k^2\delta\psi + (p\psi^*)\delta\phi + (p\phi^* - 2d\psi^*)\delta\phi \tag{2.6}$$

where the 0th order terms were already eliminated using 2.3. These equations can be written in matrix form and further simplified using 2.3.

$$\sigma \begin{bmatrix} \delta\phi \\ \delta\psi \end{bmatrix} = \begin{bmatrix} -\mu k^2 + e\phi^* & -p\phi^* \\ \frac{p^2\phi^*}{d} & -\nu k^2 - p\phi^* \end{bmatrix} \begin{bmatrix} \delta\phi \\ \delta\psi \end{bmatrix} \tag{2.7}$$

This is a standard eigenvalue problem for $\sigma$. Its solution can be written as $\sigma = 1/2(Tr \pm \sqrt{Tr^2 - 4Det})$ where $Tr$ is the trace and $Det$ is the determinant of the above matrix. The real part of the eigenvalue $\sigma$ determines if the perturbation grows or shrinks. If $\mathbb{R}(\sigma)$ is negative the perturbation will shrink, positive the perturbation will grow. Notice that $\sigma$ is in fact a function of wavenumber $k$. See Fig. 2.1 for a sketch of a typical situation. At most wave numbers $\mathbb{R}(\sigma)$ is negative and the homogeneous state is stable. However, at some values of $k$, $\mathbb{R}(\sigma)$ can become positive and destabilize the homogeneous state. The wavenumber with the largest positive real eigenvalue sets the characteristic scale for pattern formation. Once destabilized the nonlinearities will eventually stabilize the pattern. For this linear stability analysis to be valid, the initial state has to be stable, and therefore $p > e$.

This system exhibits a Turing instability when [16]:

$$\frac{\nu}{\mu} > \left( \frac{1}{\sqrt{p/d} - \sqrt{p/d - e/p}} \right)^2. \tag{2.8}$$

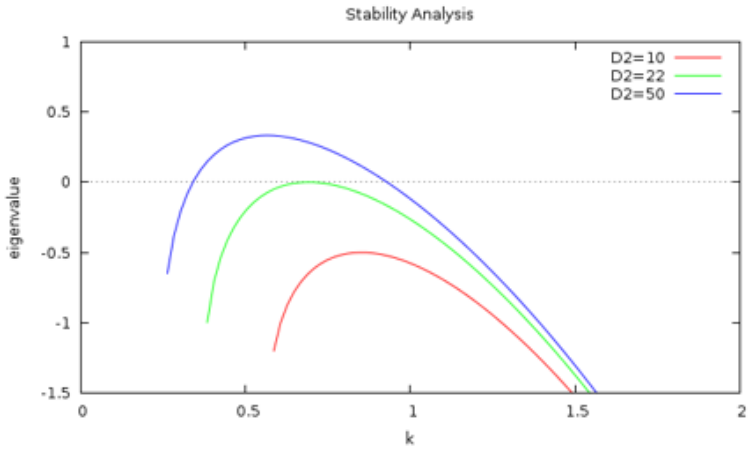


Figure 2.1: **Linear Stability Analysis.** The real part of the eigenvalue $\sigma$ is plotted as a function of wavenumber for different ratios of diffusion constants. At high enough diffusion constant ratio the eigenvalue becomes positive creating an instability in the initial stable homogeneous state.

## 2.2 Individual Level Models

There are many different types of models for physical systems, each appropriate to answering different questions. Many questions in Ecology and Chemistry deal with large numbers of organisms or chemicals, respectively. In these instances we are used to writing down continuum mean field, mass action equations that describe the dynamics of the densities of these populations. This works well when the number of entities is sufficiently large that the underlying stochasticity of the birth-death processes is sufficiently small. It turns out though that even when there are large numbers of entities, if the system is spatially extended, there can be areas where the local numbers are small enough that the stochasticity of these birth and death processes matter.

In situations where one expects these effects to matter it is appropriate to use a different level of description than the continuum modeling, known as an Individual Level Model. In an Individual Level Model, interactions and the birth and death processes of entities are modeled using chemical reaction like equations. The underlying entities, however, do not need to be chemicals and in fact, could be organisms if one were trying to model an ecological system. For example, to model the birth process of a rabbit the corresponding individual level model would be $A \xrightarrow{b} A + A$, where $b$ is the birth rate. If there are $N$ rabbits the corresponding transition rate is $T(N+1|N) = bN$. In the section on the Gillespie algorithm, 2.6, these transition rates are denoted $a_i$ where $i$ corresponds to the reaction index in consideration. A few examples of reactions and their corresponding transitions rates are as follows [17]:

Table 2.1: Reactions and their corresponding transition rates.

| Reaction | $a_i$ |
|---|---|
| $A \xrightarrow{c_1} 2A$ | $a_1 = c_1 X_1$ |
| $A \xrightarrow{c_2} B$ | $a_2 = c_2 X_1$ |
| $2A \xrightarrow{c_3} A$ | $a_3 = c_3(X_1)(X_1 - 1)/2$ |
| $A + B \xrightarrow{c_4} 2A$ | $a_4 = c_4 X_1 X_2$ |
| $A + B \xrightarrow{c_5} 2B$ | $a_5 = c_5 X_1 X_2$ |

Here $X_1$ and $X_2$ correspond to the number of entities of type A and B, respectively. Note that the reaction rates for individual level models are related to their deterministic mean field counterparts by factors of system size, depending on the order of the reactions. For first order reactions they are the same. For second order they are related by one factor of the system size.

## 2.3 Master Equation

In our stochastic models we often want to know $P(\vec{X},t)$, the probability that there will be $X_1$ .... $X_N$ molecules in volume V at time t. The time evolution of the probabilities is described by the Master Equation. The Master Equation models the time evolution of the probability of being in a given state by keeping track of the rate at which that state is being populated from other states and the rate at which that state is transitioning to other states. The transition rate $T(x|x')$ is the rate at which the system will transition from state $x'$ to state $x$ . The rate of transition into state $x$ from state $x'$ given the probability of being in state $x'$ is thus $T(x|x')P(x')$. Combining this all together produces the Master Equation.

$$\frac{\partial}{\partial t}P(x,t) = \sum_{x' \neq x} [T(x|x')P(x',t) - T(x'|x)P(x,t)] \tag{2.9}$$

where terms on the left are from states $x'$ entering state $x$ and terms on the right are for the rate at which state $x$ is leaving for other states.

Solving this multidimensional PDE is often intractable both numerically and analytically. One way to make this problem more tractable analytically is by doing various approximations. One of the most common approximation schemes is called the Van Kampen system size expansion, described in section 2.4. Another strategy taken by Gillespie [17], was to simulate trajectories that represent exact samples from the probability function corresponding to the Master Equation. His strategy was to calculate the reaction probability density function, which allows his algorithm to figure out when the next reaction is going to occur and what reaction takes place. This is discussed in section 2.6.

## 2.4 Van Kampen System Size Expansion

In the Van Kampen system size expansion the Master Equation can be expanded by making an ansatz about how the fluctuations scale with the system size. Specifically, Van Kampen made the ansatz that the copy number $X_i$ consists of a deterministic part corresponding to the scaled up concentration $\Omega\phi_i$ and random fluctuations about this that scales as $\Omega^{1/2}\xi_i$, ie. $X_i = \Omega\phi_i + \Omega^{1/2}\xi_i$ [18]. This ansatz allows for a systematic expansion of the Master Equation by equating order by order in powers of $\Omega$. The leading order of this expansion results in a Fokker-Planck equation with linear coefficients.

## 2.5 Generating Random Numbers Using Inverse Transform Method

To perform stochastic simulation, it is necessary to have the ability to generate a random number from any given probability distribution. Generally, most computer programming languages and computational software give the user the ability to generate random numbers uniformly in the interval zero to one, $x \in [0,1]$. These same software systems may also give the user the ability to generate random numbers from other common distributions such as Gaussian and Poisson, but these languages cannot have all possible probability distributions preprogrammed. In those cases that the distribution is not already programmed it is still possible to generate a random number from a given probability distribution, as long as the user is able to calculate the inverse of the cumulative distribution and generate a random number on a uniform interval.

One method of generating random numbers from a desired probability distribution is called the inverse transform sampling method. The basic steps to generate a random number $x$ from an arbitrary probability distribution $p(x)$ using this method are as follows:

1. Generate a uniform random number $r$ from the interval $[0,1]$.
2. Compute the value $x$ of the cumulative probability distribution corresponding to the distribution $p$ that gives you $r$ as an output. If the cumulative probability distribution has an analytic inverse this step is easy and can be done directly. Otherwise, it is still possible to do this step computationally by using a root finding method.
3. The value $x$ will be from the desired probability distribution.

This means that to generate a random number from our desired random distribution $p(x)$ we only need to generate one random number on a uniform interval and we can transform it directly into a random number from our desired distribution. This is an example of a direct method of generating a random number. There are other methods that require generating multiple random numbers from a uniform distribution to generate one random number from the desired probability distribution, for example, using an accept and rejection method similar to throwing darts at a dart

board. The advantage of the inverse transform method is that it requires only generating one random number; but in cases where calculating the inverse of the cumulative probability distribution is computationally expensive it can be beneficial to use other methods to generate a random number from a specific distribution.

As an example of how inverse transform sampling works consider the following examples: generating a random number from an exponential distribution and generating a number from a discrete distribution. To generate a random number from the exponential distribution $p(x) = \lambda e^{-\lambda x}$ we first generate a uniform random number $r \in Unif(0,1)$. We then calculate the value of $x$ that gives us $r$ from the cumulative probability distribution.

$$\begin{aligned} r &= \int_0^x \lambda e^{-\lambda s} ds \\ r &= [1 - e^{-\lambda x}] \end{aligned} \tag{2.10}$$

and we find

$$x = \frac{1}{\lambda} \ln\left(\frac{1}{1-r}\right) \tag{2.11}$$

Note that since $u = 1 - r$ is still a uniform random number in the interval $[0,1]$, instead of drawing $r$ draw $u$ and calculate $x$ directly as $x = \frac{1}{\lambda} \ln\left(\frac{1}{u}\right)$.

For a discrete probability distribution $p_x$ we can use the following procedure. Draw $r \in Unif(0,1)$ then pick $x$ such that

$$\sum_{i=1}^{x-1} p_i < r \leq \sum_{i=1}^{x} p_i \tag{2.12}$$

That is, compute the partial sum of probabilities until that partial sum is greater than the random number drawn. The ability to draw random numbers from the exponential distribution and from a discrete distribution are both necessary to be able to perform a Gillespie simulation.

## 2.6 Gillespie Algorithm

As mentioned in section 2.3, solving the Master Equation is often intractable, both numerically and analytically. Instead, Gillespie chose to explicity simulate each reaction, where the sequence and timing of the chemical reactions will correspond to an exact sample from the corresponding Master Equation. To be able to do this Gillespie calculated the reaction probability density function $P(\tau,\mu)d\tau$, the probability that the next reaction will be reaction $\mu$ and that it will happen in the time interval $(t+\tau, t+\tau+d\tau)$. Gillespie's main idea was that a sample from this probability

distribution will produce the time increment to the next reaction and the next reaction to perform. Upon updating the state of the system, this procedure can be repeated. A trajectory calculated in this way will correspond to a simulation of the stochastic system.

The basic steps of Gillespie's algorithm are thus:

1. Initialize.
2. Monte Carlo Step - generate the time to the next reaction and the reaction that occurs by drawing from the reaction probability density function.
3. Update the time, the number of reactants, and the transition rates.
4. Record the number of reactants and time, if the sampling time has passed.
5. Repeat.

This procedure depends on being able to calculate the reaction probability density function and sampling from it.

Gillespie was able to calculate this distribution as follows [17]. Let $a_\mu dt$ be the probability that reaction $\mu$ will occur in time dt. Then the reaction probability density function $P(\tau,\mu) = P_0(\tau)a_\mu d\tau$ consists of two parts: the probability $a_\mu dt$ that reaction $\mu$ occurs in time $d\tau$, and the probability $P_0(\tau)$ that no reaction occurs in the time interval $\tau$. To calculate $P_0(\tau)$ break up the interval $\tau$ into $K$ sub-intervals of size $\varepsilon = \tau/K$. The probability that a reaction $n$ does not occur in the sub-interval $\varepsilon$ is $(1 - a_n\varepsilon)$. The probability that no reaction occurred in the sub-interval $\varepsilon$ is therefore $\prod_{n=1}^{N}(1-a_n\varepsilon) \approx 1 - \sum_{n=1}^{N} a_n\varepsilon + O(\varepsilon^2) = 1 - a\tau/K$ where $a = \sum_{n=1}^{N} a_n$. Thus, the probability that no reaction occurred in the interval $\tau$ is the probability that no reaction occured in any of the $K$ sub-intervals of length $\varepsilon$; namely that

$$P_0(\tau) = \lim_{K\to\infty}\left(1 + \frac{-a\tau}{K}\right)^K = \exp(-a\tau) \tag{2.13}$$

Thus, we see that the reaction probabilty density function is

$$P(\tau,\mu)d\tau = a_\mu \exp(-a\tau) = \frac{a_\mu}{a}\cdot a\exp(-a\tau) \tag{2.14}$$

It is also possible to decompose the reaction probability distribution so $P(\tau,\mu) = P_1(\tau)P_2(\mu|\tau)$, where $P_1(\tau) = a\exp(-a\tau)$ is exponentially distributed and is the probability that the next reaction will occur in time $\tau$, and $P_2(\mu|\tau) = a_\mu/a$ is the probability that reaction $\mu$ will occur at time $\tau$. Thus, to do the Monte Carlo step described above one can use the inverse transform sampling method to draw two random numbers, one from the exponential distribution, and one from the discrete distribution $P_2(\mu|\tau)$. That is, calculate the time for the next reaction as $\tau = \frac{1}{a}\ln\left(\frac{1}{u}\right) \quad u \in Unif(0,1)$, and find the next reaction $x$ using $\sum_{i=1}^{x-1}\left(\frac{a_i}{a}\right) < r \leq \sum_{i=1}^{x}\left(\frac{a_i}{a}\right) \quad r \in Unif(0,1)$.

Thus in summary, the Gillespie algorithm samples from the reaction probability distribution to draw one number that is exponentially distributed which represents the time to the next reaction, and one number which represents which reaction to choose. Once these numbers have been drawn the simulation state is updated by performing the selected reaction $\mu$ and incrementing time by $\tau$. It is important to remember to update the number of molecules and transition rates $a_n$. Once the update is finished the whole process can be repeated.

By producing multiple trajectories using the Gillespie algorithm it is then possible to calculate the probability distribution of states at different times. This should obey the Master Equation. The Gillespie algorithm is a wonderful tool that is easy to implement and is a good benchmark for testing any analytic solutions that are found.

# Chapter 3

# Stochastic Turing Patterns in a Synthetic Bacterial Population

The work appearing in this chapter is the result of a collaboration with Ron Weiss's lab at MIT. The people involved in this collaboration included David Karig, Ting Lu, Nicholas A. DeLateure, Nigel Goldenfeld, and Ron Weiss. This work is in the process of being published in PNAS and this chapter is a modified version of that paper. The experiments and experimental design were developed by David Karig and Ting Lu. I performed the simulations and measured pattern characteristics of both the simulations and experiments, including spot size, power spectrum, and minimum distance between spots. I developed the phase diagram and sensitivity analysis of parameters for the detailed stochastic model. I also developed the phase diagram for the reduced model. The majority of my contributions appear in sections 3.3 and 3.4.

## 3.1 Introduction

A central question in biological systems, particularly in developmental biology, is how patterns emerge from an initially homogeneous state [19]. In his seminal 1952 paper, "The Chemical Basis of Morphogenesis," Alan Turing showed, through linear stability analysis, that stationary, periodic patterns can emerge from an initially uniform state in reaction-diffusion systems where an inhibitor morphogen diffuses sufficiently faster than an activator morphogen [2]. However, the requirements for realizing robust pattern formation according to Turing's mechanism are prohibitively difficult to realize in nature. [16, 20, 21]. Although Turing patterns were observed in a chemical system in 1990 [22], the general role of Turing instabilities in biological pattern formation has been called into question, despite a few rare examples (see for example [23, 24, 25, 26, 27, 28]).

Recently, Turing's theory was extended to include intrinsic noise arising from activator and inhibitor birth and death processes [3, 4, 5, 29]. According to the resulting stochastic Turing theory, demographic noise can induce persistent spatial pattern formation over a wide range of parameters, in particular, removing the requirement for the ratio of inhibitor-activator diffusion coefficients to be large. Moreover, stochastic Turing theory shows that the extreme sensitivity of pattern-forming systems to intrinsic noise stems from a giant amplification resulting from the non-orthogonality of eigenvectors of the linear stability operator about the spatially-uniform steady state [29]. This

amplification means that the magnitude of spatial patterns arising from intrinsic noise is not limited by the noise amplitude itself, as one might have thought naively. These developments imply that intrinsic noise can drive large-amplitude stochastic Turing patterns for a much wider range of parameters than the classical, deterministic Turing theory. In particular, it is often the case in nature that the activator and inhibitor molecules do not have widely differing diffusion coefficients; nevertheless, stochastic Turing theory predicts that even in this case, pattern formation can occur at a characteristic wavelength that has the same functional dependence on parameters as in the deterministic theory.

In order to explore how global spatial patterns emerge from local interactions in isogenic cell populations, a promising strategy that has been advocated is to develop synthetic bacterial populations whose collective interactions can be controlled and well-characterized (for an introduction to this perspective, see e.g. [30]). Synthetic systems can be forward-engineered to include relatively simple circuits that are loosely coupled to the larger natural system into which they are embedded. This makes it easier to design and control the molecular underpinnings of a biological pattern phenomenon [31, 32, 33, 34] and even front propagation phenomena [35]. Previous pattern formation efforts in synthetic biology have focused on oscillations in time [36, 37, 38, 39] , or required either an initial template [40, 32, 41, 34] , or an expanding population of cells [42], neither of which demonstrates a Turing mechanism. In short, by programming a synthetic biological system where patterning is instead driven by activator/inhibitor diffusion across an initially homogeneous lawn of cells [30], we can explore stochastic Turing patterns. The use of a synthetic population overcomes the challenges presented by natural biological Turing pattern systems – namely, that natural systems are difficult to manipulate because their chemical and genetic mechanisms are complex and not fully-understood [24, 25, 26, 27, 28, 23]. Our synthetic system helps to reveal design principles of biological patterning systems and represents a proof-of-principle for engineered biological spatial patterns stabilized by stochastic gene expression.

## 3.2 Experimental Results

### 3.2.1 Synthetic Biology of a Bacterial Community

To address the problem of Turing instability induced pattern formation alluded to above, we designed a synthetic pattern forming gene circuit that destabilizes an initially homogeneous lawn of genetically engineered bacteria. This system is subject to stochastic gene expression, and as we show below, produces stochastic patterns with a spatial scale much larger than that of a single cell. The patterns observed in our engineered cells are noisy, with power spectrum power-law tails consistent with theoretical predictions for patterns stabilized by intrinsic noise.

In our synthetic gene network design we used two artificial diffusible morphogens: the small molecule N-(3-oxododecanoyl) homoserine lactone, denoted here as $A_{3OC12HSL}$, and the small molecule N-butanoyl-L-homoserine

lactone, denoted here as $I_{C4HSL}$, from the *Pseudomonas aeruginosa las* and *rhl* quorum sensing pathways respectively in *Pseudomonas aeruginosa* [43]. $A_{3OC12HSL}$ serves as an activator of both its own synthesis and that of $I_{C4HSL}$, while $I_{C4HSL}$ serves as an inhibitor of both signals (Fig. 3.1a-b). $A_{3OC12HSL}$ activates its own synthesis and synthesis of $I_{C4HSL}$ by binding regulatory protein LasR to form a complex that activates the hybrid promoter $P_{Las-OR1}$. This promoter regulates expression of LasI, a $A_{3OC12HSL}$ synthase, and *rhlI*, a $I_{C4HSL}$ synthase. To increase the sensitivity of $A_{3OC12HSL}$ self-activation, LasR is regulated by a second copy of $P_{Las-OR1}$. $I_{C4HSL}$ inhibits synthesis of $A_{3OC12HSL}$ and itself by forming a complex with regulatory protein RhlR. This complex activates expression of lambda repressor CI which, in turn, represses transcription of LasI, RhlI, LasR and RhlR. Pattern formation in our system can be modulated by altering the concentration of isopropyl $\beta$-D-1-thiogalactopyranoside (IPTG), a small molecule inducer that binds LacI and alleviates repression of $P_{Rhl-lacO}$. Green and red fluorescent proteins (GFP and RFP) are expressed from the *rhl* and *las* hybrid promoters respectively to aide in experimental observation.

In our experimental setup, the $A_{3OC12HSL}$ activator diffuses more slowly than the $I_{C4HSL}$ inhibitor. The estimated diffusion coefficient for $A_{3OC12HSL}$ is $83\,\mu m^2/s$ and for $I_{C4HSL}$ is $1810\,\mu m^2/s$. The experimentally determined ratio of diffusion rates in our system of 21.6 is much higher than the value of 1.5 predicted by Wilke-Chang correlation in water [44], likely due to partitioning of $A_{3OC12HSL}$ in the cell membrane, which slows its diffusion from cell to cell [45]. The slower diffusion rate of $A_{3OC12HSL}$, coupled with positive feedback regulating its synthesis, allows $A_{3OC12HSL}$ to aggregate in local domains, leading to formation of visible red fluorescent spots (cellular lawn illustration shown in Fig. 3.1c). Within these red domains, both $A_{3OC12HSL}$ and $I_{C4HSL}$ are found in high concentrations, but because $A_{3OC12HSL}$ competitively binds RhlR, GFP is attenuated [46]. The faster diffusion rate of $I_{C4HSL}$ allows it to diffuse into regions outside of the red fluorescent domains. Here, $I_{C4HSL}$ is free to bind RhlR, activating GFP expression. Collectively, these processes lead to green regions between red spots.

### 3.2.2 Experimental Patterns and Controls

Cells harboring appropriate plasmids were initially grown in LB liquid media with corresponding antibiotics at 30°C until optical density at 600 nm reached $0.1-0.3$. Cells were then concentrated and re-suspended in M9 media with appropriate antibiotics[47]. 0.5 mL of concentrated cell solutions ($OD_{600} = 2.0$) were poured onto a 2% M9 agar plate ($60 \times 15$ mm Petri dish) to form a cellular lawn. Plates were incubated at 30 °C and fluorescence images were captured periodically. To examine the single cell fluorescence evolution of toggle switch cell populations, we performed flow cytometry at the beginning of the experiment (0 h) and at the end of the experiment (24 h).

To study pattern forming behavior, engineered cells harboring appropriate plasmids were initially grown in LB liquid media with corresponding antibiotics at 30°C until optical density at 600 nm reached $0.1-0.3$. Cells were then concentrated and re-suspended in M9 media with appropriate antibiotics [47]. 0.5 mL of concentrated cell solutions

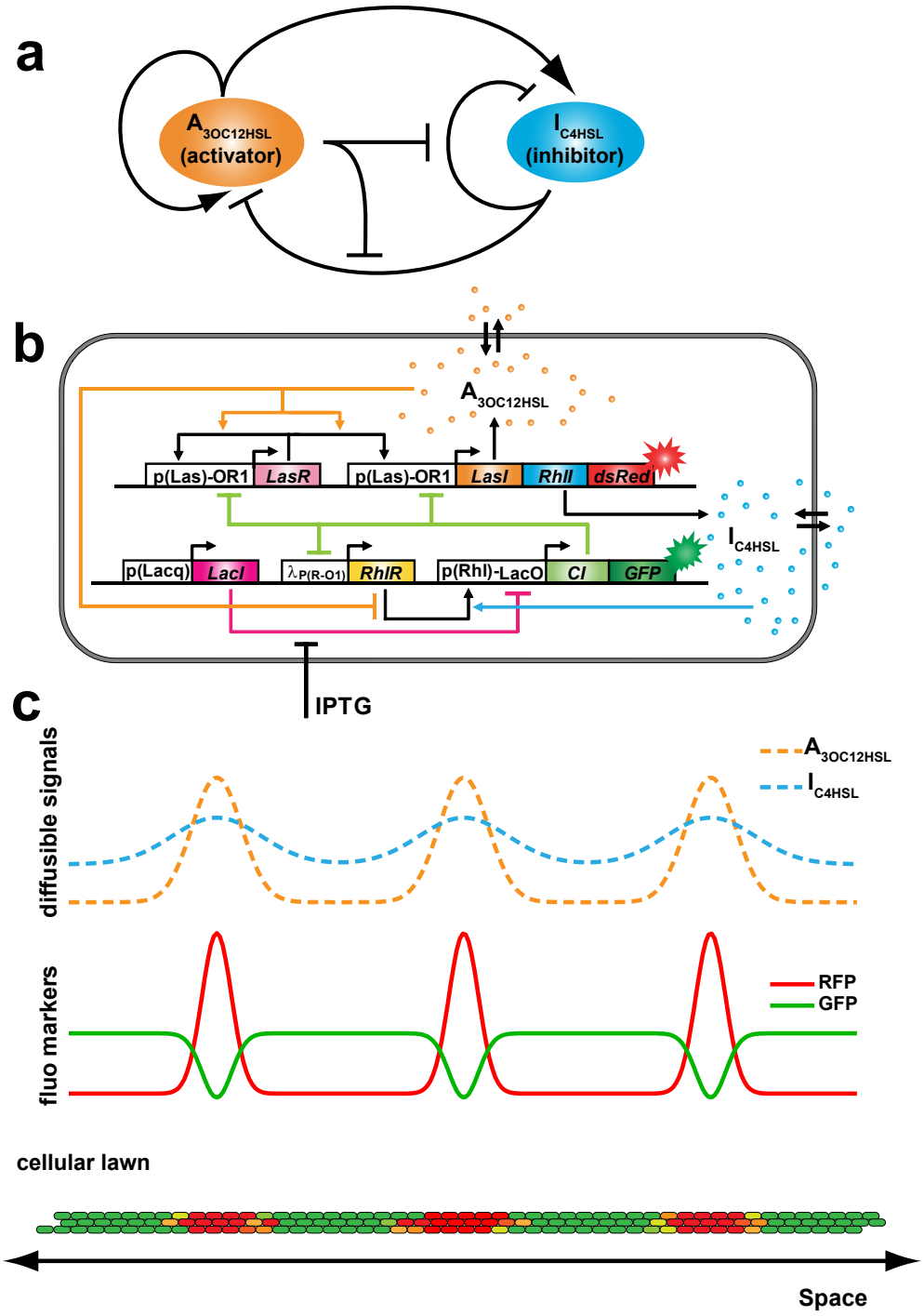


Figure 3.1: **Design of a synthetic multicellular system for emergent pattern formation.** **a**, Abstractly, the system consists of two signaling species $A_{3OC12HSL}$ and $I_{C4HSL}$: $A_{3OC12HSL}$ is an activator catalyzing synthesis of both species while $I_{C4HSL}$ is an inhibitor repressing their synthesis, with additional repression by $A_{3OC12HSL}$ via competitive binding. **b**, Genetic circuit implementation. Promoter regions are indicated by white boxes, while protein coding sequences are indicated by colored boxes. IPTG is an external inducer modulating system dynamics. **c**, Top: Illustration of signaling species concentrations in one-dimensional space. The dashed orange and blue lines correspond to $A_{3OC12HSL}$ and $I_{C4HSL}$ respectively. Middle: Spatial profiles of reporter proteins. RFP expression (red line) correlates with $A_{3OC12HSL}$ concentrations, while GFP expression (green line) roughly mirrors RFP expression. Bottom: A vertical slice of cell lawn. Cells express fluorescence proteins according to the profiles above and produce a global multicellular pattern.

($OD_{600} = 2.0$) were poured onto a 2% M9 agar plate ($60 \times 15$ mm Petri dish) to form a cellular lawn. Plates were incubated at 30 $^{o}$C and fluorescence images were captured as needed. Prior to the self-activation of the $A_{3OC12HSL}$ synthase positive feedback loop, the cell lawn exhibits no fluorescence. However, over time red fluorescent spots emerge with sizes much larger than that of a single cell (10-1000x). Simultaneously, green fluorescence develops in a pattern with dark voids positioned precisely in the locations of the intense red fluorescence (Fig. 3.7a). Time-series microscopy reveals that patterns begin to emerge after approximately 16 hours (Fig. 3.5).

In control experiments, we show that our patterns are not simply a result of the outward growth of clusters of differentially colored cells (Fig. 3.7b-c). For this we first assayed the phenotypic behavior of lawns of cells that express fluorescent proteins constitutively. As shown in Fig. 3.7b, when red and green fluorescent cells are grown separately or together, uniform fluorescent fields develop. The difference between these control experiments and the emergent patterns is illustrated clearly in the red/green fluorescence density plots (Fig. 3.7c). We further tested additional ratios of constitutively fluorescent green and red cells and again observed relatively uniform fields of fluorescence (Fig. 3.2). These experiments demonstrate that, in our experimental setup, neither cell growth nor initial spatial heterogeneity of cell density give rise to the large scale spatial patterns observed with the Turing cells.

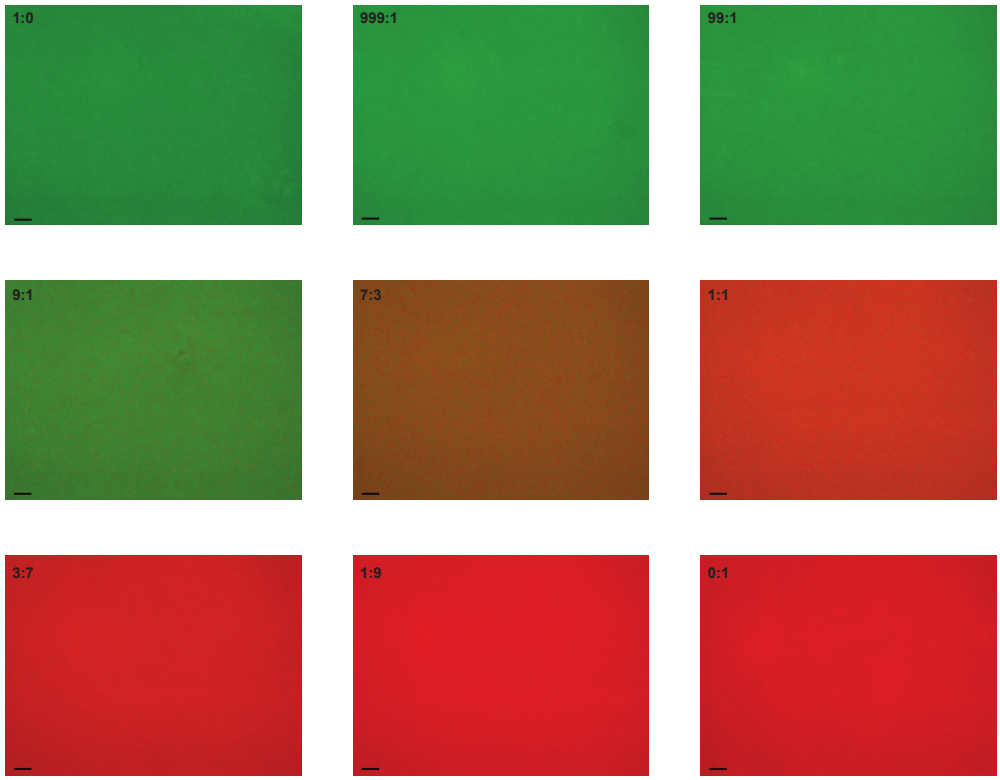


Figure 3.2: Populations of *E. coli* expressing constitutive fluorescent reporter proteins GFP or mCherry were mixed in various ratios of [Green:Red] on M9 supplemented minimal media and then imaged by microscopy. Scale bar, 200 µm.

Also, by performing an experiment with cells that harbor independent bistable green/red toggle switches, we test whether observable patterns would emerge if individual cells autonomously made cell-fate decisions at some point after plating (Fig. 3.3a) [48]. For these switches, which are essentially net positive feedback loops, IPTG induction results in expression of TetR/GFP, aTc induction results in expression of LacI/RFP, and absence of inducer results in a "memory" of the cells' most recent state (at 30°C) [48]. Co-induction with both inducers gives rise to co-expression of all proteins; subsequent simultaneous removal of the inducers causes each cell to make an independent quasi-random decision and enter one of the two stable states. To explore whether such an independent decision-making process results in global pattern formation, we induced toggle cells with 3 $\mu$M IPTG and 0.3 $\mu$M aTc in liquid culture for 5 hours. Flow cytometry analysis confirmed that after this initial incubation period, all cells in the population had roughly the same red/green fluorescence levels (Fig. 3.3b). Co-induced cells were then plated onto Petri dishes lacking inducers (bistability condition) using the same technique as the experiments above. The fluorescence fields after 24 hour incubation at 30 °C were uniform, showing no emergence of patterns (Fig. 3.3d-f). However, flow cytometry analysis of cells scraped from the plate after 24 hours revealed that the initially homogeneous cell population had bifurcated almost completely into two subpopulations, one with high GFP expression and the other with high RFP expression (Fig. 3.3c). The toggle switch cell lawn maintained spatial homogeneity but individual cells settled into one of the two states, suggesting that this autonomous quasi-random fate decision by individual cells does not lead to global spatial patterning.

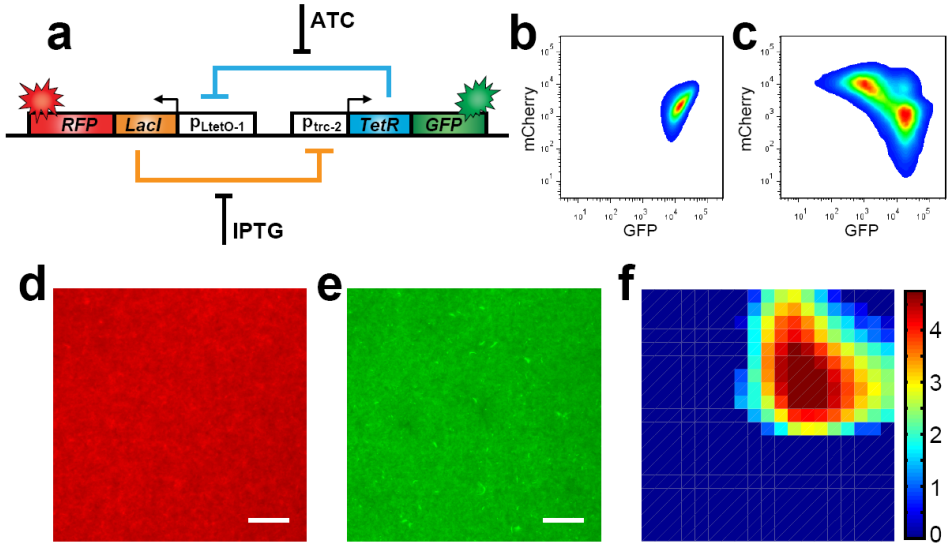


Figure 3.3: Behavior of cell populations with each cell harboring an intracellular green/red bistable toggle switch. **a**, A bistable toggle switch derived from Gardner *et al* [48]. **b**, Flow cytometry fluorescence density plot of the toggle cells at time 0. **c**, Flow cytometry fluorescence density plot of cells scraped from the cell lawn at 24 hours. **d**-**e**, Microscope images of cell lawns harboring the toggle switch circuit at 24 hours. Both RFP (**d**) and GFP (**e**) are homogeneously distributed and qualitatively different from that of a population carrying the emergent circuit in Fig. 2**a**. Scale bar, 100 µm. **f**, Fluorescence density plot of the microscope images in panels **d**-**e**.

Next, we examine how changes in the strengths of localized interactions lead to different global outcomes in our pattern forming gene circuit. In our system, IPTG can be used to modulate the inhibitory efficiency of $I_{C4HSL}$ in individual cells by affecting CI expression from $P_{Rhl-lacO}$, up to the threshold of toxicity. Our data show that mean GFP levels increase sigmoidally with inducer concentration while the overall area of red spots decreases (Fig. 3.8a-c), correlating well with the results from our mathematical model (Fig. 3.8d-g). To quantify changes in the spatial characteristics of the patterns in response to different IPTG concentrations, we define a *collectivity metric* as follows:

$$\Theta = \sum_{i,j=1}^{M} \sigma_{i,j}, \quad \text{where } \sigma_{i,j} = \begin{cases} 1 & \text{if pixels } i \text{ and } j \text{ are in the same red spot} \\ 0 & \text{otherwise} \end{cases}$$

where $M$ is the total number of pixels in the image. Figure 3.8b shows that in our experiments, the collectivity metric decreases approximately 9 fold as a function of IPTG, indicating that an increase in the inhibitory effect of $I_{C4HSL}$ in each individual cell results in reduced overall global clustering. Moran's I [49] is also plotted to illustrate how the spatial autocorrelation of an image decreases with IPTG (Fig. 3.8b, inset), qualitatively consistent with the analysis of our simulated patterns (Fig. 3.4). Additionally, spots become smaller due to IPTG induction as is visible in the microscope images in Figure 3.8a.

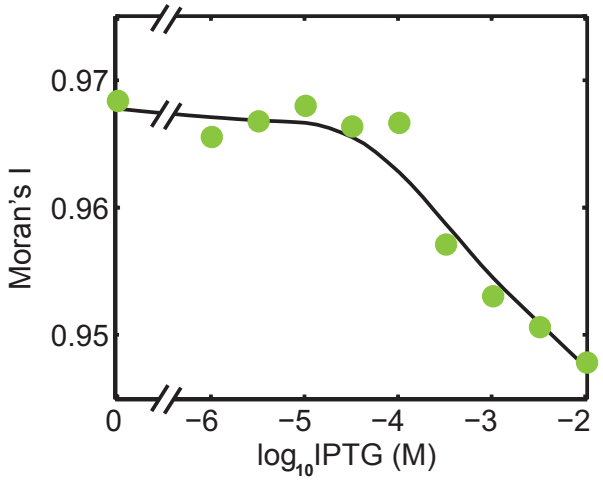


Figure 3.4: Moran's I of the patterns from our deterministic simulations.

We performed a 32-hour experiment to gain a better understanding of the dynamics of pattern emergence. A lawn of cells was prepared as described above, placed in a microscope chamber and incubated at 30 $^{o}$C. Fluorescence images of the same region were captured once every 30 minutes. Figure 3.5 shows images at 4 hour intervals (0-, 4-, 8-, 12-, 16-, 20-, 24-, 28-, and 32-hour). There is no fluorescence initially until hour 16 when tiny spots emerge. These tiny spots grow quickly and new spots continue appearing and growing during the following few hours. By hour 24,

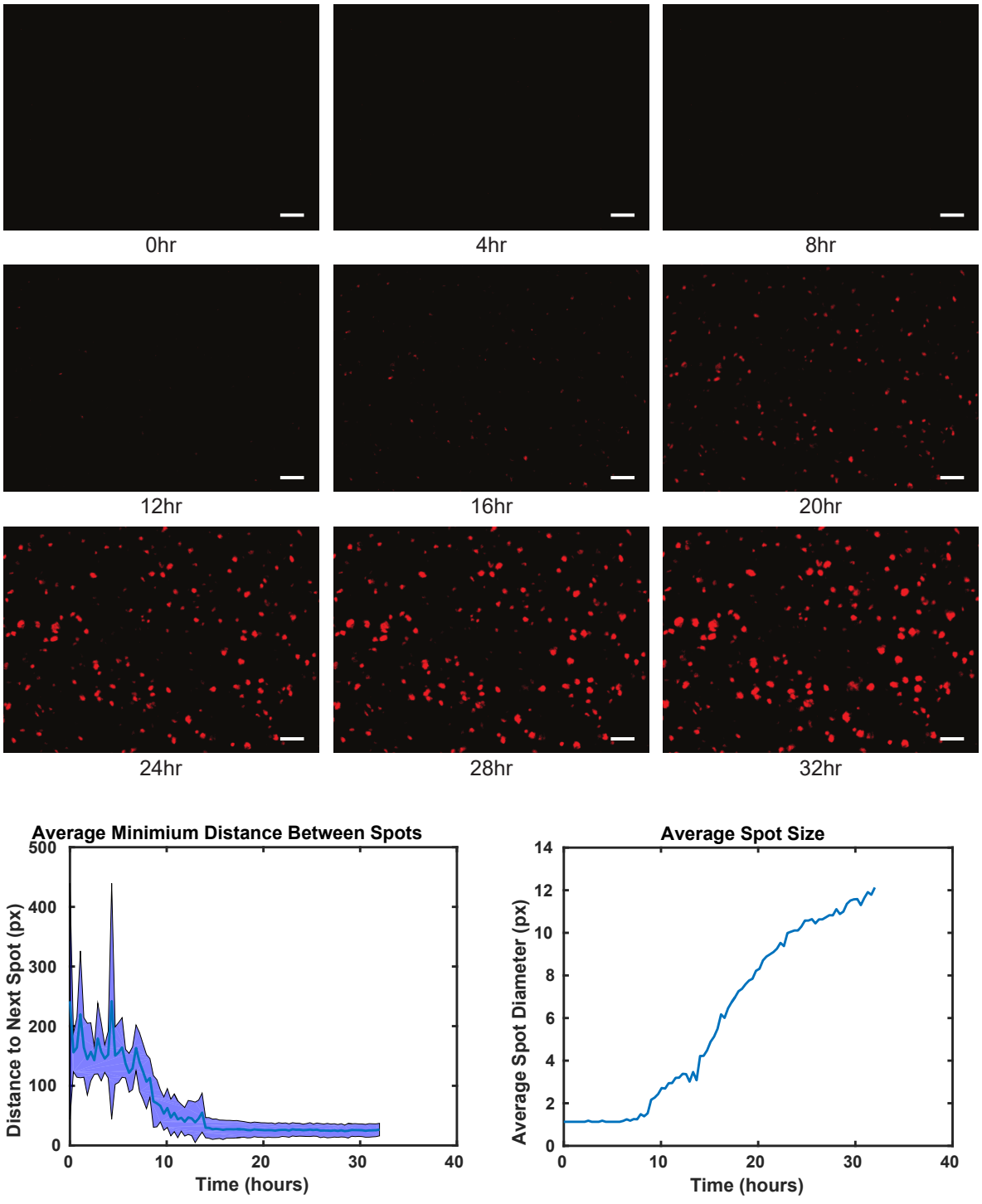


Figure 3.5: Emergence of patterns over time. Snapshots of red fluorescence were taken every 30 minutes for 32 hours. Shown are images in 4-hour intervals. Scale bar, 100 µm. Left: Average minimium distance between spots as a function of time, blue shading indicates standard deviation. Right: Average spot diameter as a function of time.

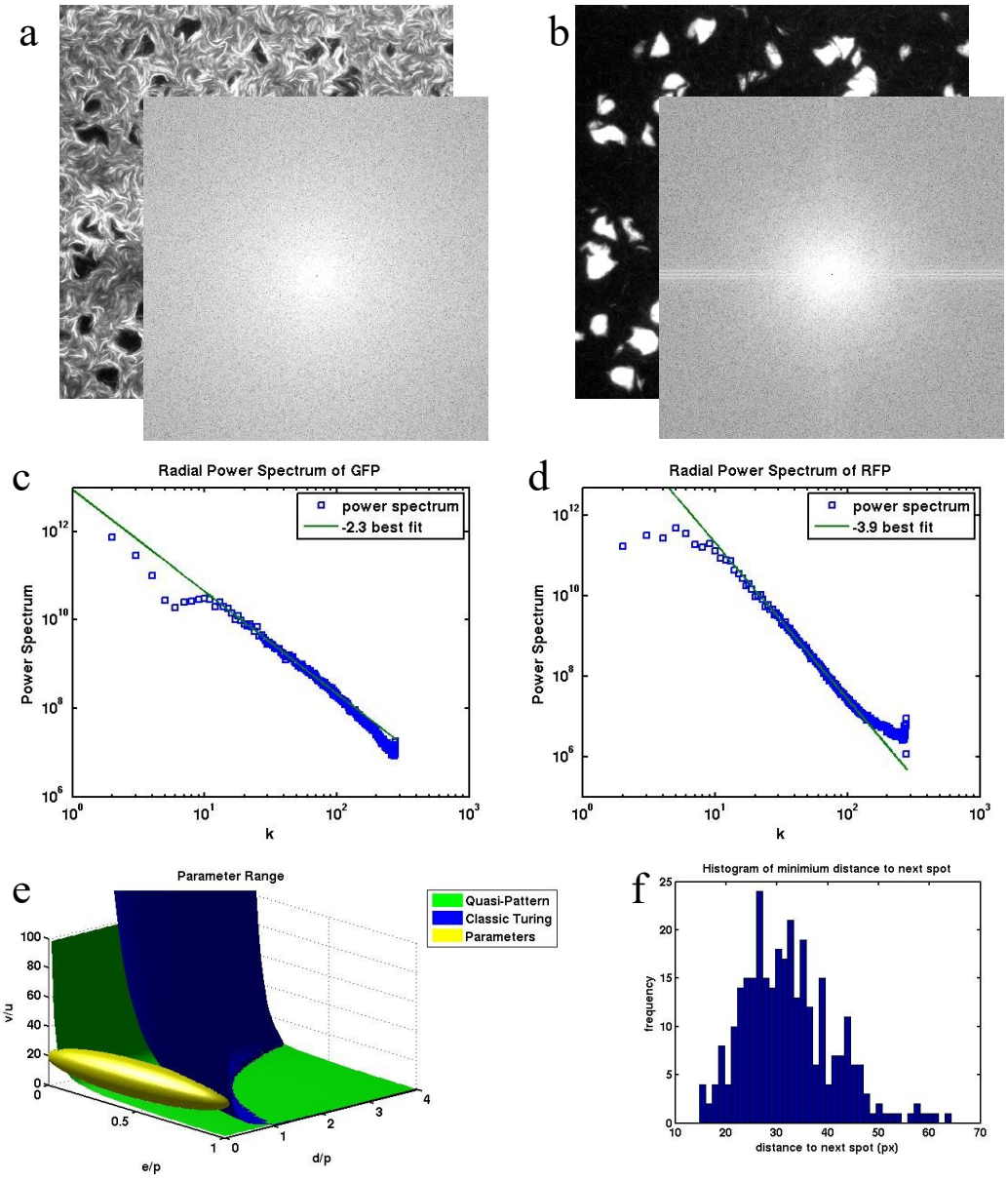


Figure 3.6: **a**, GFP image and corresponding Fourier transform.**b**, RFP image and corresponding Fourier transform. **c**, Radial power spectrum of GFP and power law fit of -2.3. **d**, Radial Power spectrum of RFP with a powerlaw tail fit of -3.9. **e**, Pattern forming regimes in parameter space and estimated parameters for our system. The parameters fall above the region where stochastic patterns form but below the region where normal Turing patterns form. **f**, Characteristic separation of spots with average separation of $32 \pm 8px$ ($45 \pm 11$ $\mu$m).

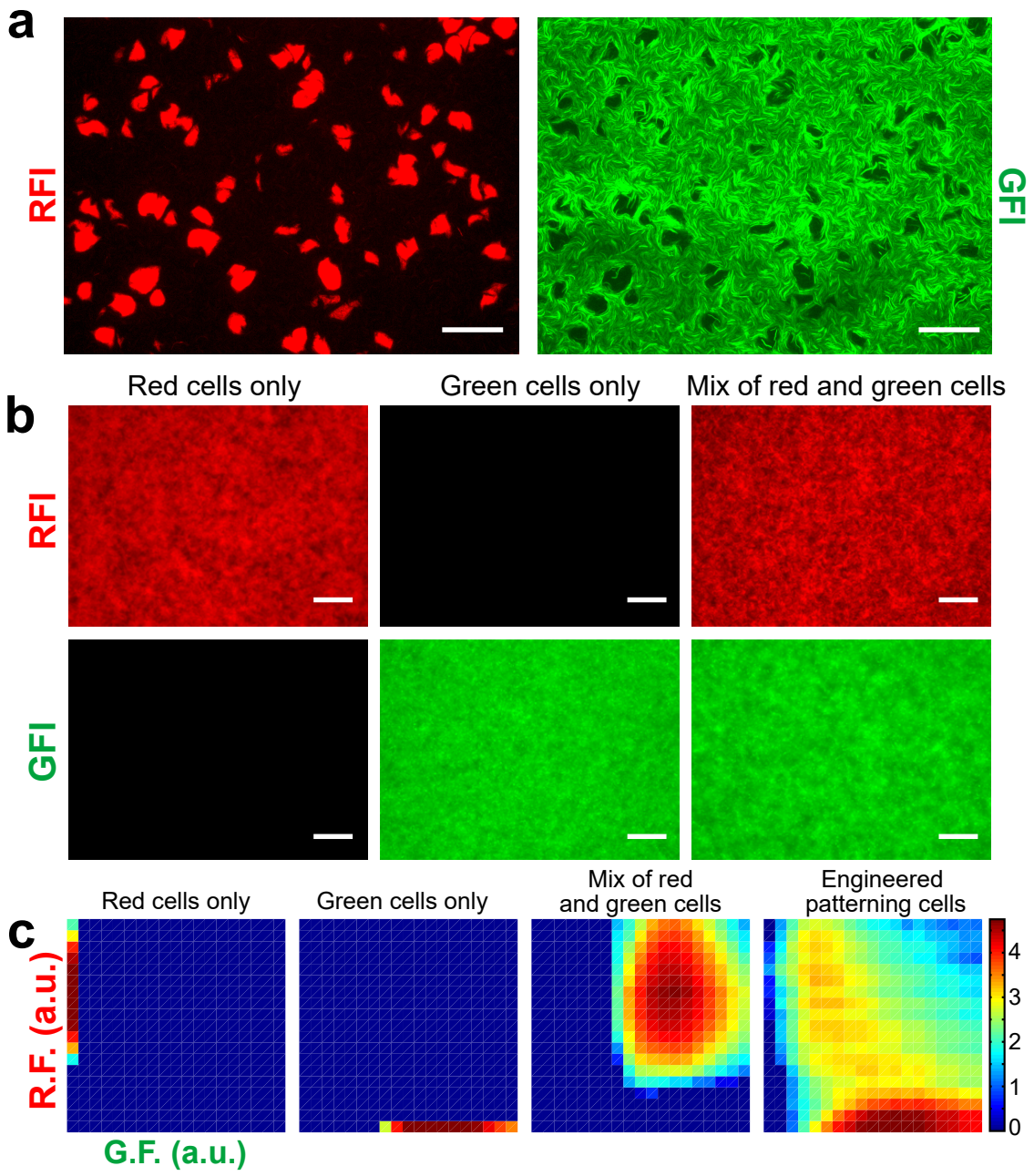


Figure 3.7: **Experimental observations of emergent pattern formation.** **a**, Representative microscope images (based on 6 technical replicates) of a typical field of view showing a fluorescent pattern formed by an initially homogeneous isogenic lawn of cells harboring the Turing circuit with no IPTG. Spots and voids appear in the red and green fluorescence channels, respectively. Scale bar, 100 µm. **b**, Microscope images of cell lawns with constitutive expression of fluorescent proteins. Left: cells expressing RFP; Middle: cells expressing GFP; Right: mixed population of red and green cells. **c**, Fluorescence density plots computed from the images above (left-to-right: red, green, red/green, and Turing). Color intensity is in log scale (a.u.).

spots have emerged with typical sizes much larger than that of a single cell. The spot pattern remains roughly the same from hour 24 to hour 32. However, as our experimental system is fundamentally a dissipative system, and we do not feed fresh nutrients, an eventual breakdown is inevitable.

We can also extract the characteristic scale of the pattern. To do this we found the centroid points of each clump of activator. We then created a histogram of distances to the nearest neighboring centroids Figure 3.6f. From this plot we found that the average separation of clumps is $45 \pm 11$ $\mu$m. Additionally, we can extract the distribution for sizes of the spots. We found the the average radius of the clump is 14 $\mu$m.

#### Moran's I

Moran's I was developed to measure spatial autocorrelation and indicates whether adjacent observations of the same phenomenon are correlated. Moran's I was proposed as follows [49]

$$I = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{3.1}$$

where $N$ is the total number of pixels, $x$ is the variable of interest (red fluorescence level here), $\bar{x}$ is the mean of $x$, and $w_{ij}$ a weight matrix of pixels. We employ a simple form of the weight matrix as follows: $w_{ij} = 1$ if two pixels are directly adjacent and $w_{ij} = 0$ otherwise. Moran's I values typically range from $+1$, representing complete positive spatial autocorrelation, to $-1$, corresponding to complete negative spatial autocorrelation.

## 3.3 Theoretical Results

Having established that our system forms emergent patterns, we proceeded to study the mechanisms driving these patterns. We formulated deterministic and stochastic models and analyzed our data to assess agreement with the theory of stochastic Turing patterns.

### 3.3.1 Deterministic Model

We first developed a detailed deterministic reaction-diffusion model. The model explicitly describes chemical reactions for the LasI and RhlI synthases, regulatory protein CI, and synthesis and diffusion of the morphogens $A_{3OC12HSL}$ and $I_{C4HSL}$. As the overall system involves a large number of reactions with rate constants that span multiple timescales, we made two commonly used simplifying assumptions. First, we assume that operator states of a promoter fluctuate much faster than protein degradation rates. Second, we assume that mRNA half-life is much shorter than protein half-life. These assumptions allow us to eliminate operator fluctuation and mRNA kinetics and model the

system at the communication signals and protein levels as follows:

$$\frac{\partial U}{\partial t} = \alpha_u I_u - \gamma_u U + D_u \nabla^2 U \tag{3.2}$$

$$\frac{\partial V}{\partial t} = \alpha_v I_v - \gamma_v V + D_v \nabla^2 V \tag{3.3}$$

$$\frac{\partial I_u}{\partial t} = \alpha_{iu} F_1(X_1, C) - \gamma_{iu} I_u \tag{3.4}$$

$$\frac{\partial I_v}{\partial t} = \alpha_{iv} F_1(X_1, C) - \gamma_{iv} I_v \tag{3.5}$$

$$\frac{\partial C}{\partial t} = \alpha_c F_2(X_2, L) - \gamma_c C \tag{3.6}$$

where $U$ and $V$ are the concentrations of the two diffusible morphogens $A_{3OC12HSL}$ and $I_{C4HSL}$, $I_u$ and $I_v$ are the concentrations of corresponding AHL synthases, and $C$ refers to CI.

We model the hybrid promoters using the following Hill functions:

$$F_1(X_1, C) = \frac{[1 + f_1(\frac{X_1}{K_{d1}})^{\theta_1}][1 + f_2^{-1}(\frac{C}{K_{d2}})^{\theta_2}]}{[1 + (\frac{X_1}{K_{d1}})^{\theta_1}][1 + (\frac{C}{K_{d2}})^{\theta_2}]} \tag{3.7}$$

$$F_2(X_2, L) = \frac{[1 + f_3(\frac{X_2}{K_{d3}})^{\theta_3}][1 + f_4^{-1}(\frac{L}{K_{d4}})^{\theta_4}]}{[1 + (\frac{X_2}{K_{d3}})^{\theta_3}][1 + (\frac{L}{K_{d4}})^{\theta_4}]} \tag{3.8}$$

where $F_1(X_1, C)$ and $F_2(X_2, L)$ are the production rates of the promoters $P_{Las-OR1}$and $P_{Rhl-lacO}$, $X_1$ and $X_2$ are the LasR-$A_{3OC12HSL}$ complex and the RhlR-$I_{C4HSL}$ complex respectively, and $L$ is the concentration of unbound LacI protein. We use the definitions

$$X_1 = R_u U \tag{3.9}$$

$$X_2 = \frac{R_v V}{(1 + U/K_{c3})} \tag{3.10}$$

$$L = \lambda_l \left( \frac{1 + f_6^{-1} (I/K_{d6})^{\theta_6}}{1 + (I/K_{d6})^{\theta_6}} \right) \tag{3.11}$$

where $I$ is the IPTG concentration, $R_u$ and $R_v$ are the regulatory proteins LasR and RhlR:

$$R_u = \lambda_u I_u \tag{3.12}$$

$$R_v = \lambda_v \left( \frac{1 + f_5^{-1} (C/K_{d5})^{\theta_5}}{1 + (C/K_{d5})^{\theta_5}} \right) \tag{3.13}$$

A summary of the variables used in our model is available in Table 3.1 and definitions of the rate constants in Tables 3.2-3.3. As the goal of producing this deterministic model was to see if we could reproduce the principal features of the observed pattern, we use order of magnitude estimates for parameters. Hill functions employed in this model

have a shared form of $Y = \frac{1+f(X/K)^{\theta}}{1+(X/K)^{\theta}}$, where $X$ and $Y$ correspond to the input and output of the function, $K$ is the dissociation constant, $\theta$ is the Hill coefficient and $f$ is the fold change of $Y$ upon full induction by $X$.

To study patterning using our model, we divide a cellular lawn into a mesoscopic $M \times M$ grid ($M = 64$ in our simulation). As is common for deterministic Turing simulations, we introduce small variation into the initial concentrations of the molecules for initial symmetry breaking. All the variables (species) were initially assigned low values (random values obeying a Gaussian distribution that has a mean of 1.0 and a variance of 0.05) to approximate the initial condition in our experimental setup. We numerically integrate the partial differential equations over time to simulate spontaneous pattern formation. We also perform numerical simulations with a range of IPTG concentrations (from $10^{-6}$ to $10^{-2}$ M) to explore modulation of pattern formation. Sizes of simulated patterns are determined in terms of relative fluorescence intensities rather than absolute values to match our image processing procedures for the experimental data.

We initially ran simulations of this model using a high diffusion rate ratio ($\frac{D_v}{D_u}$) of 100. These simulations yield patterns of red spots and green voids (Fig. 3.8d), suggesting that the underlying dynamics of our system are Turing-like, with the potential for Turing instabilities. Deterministic simulations of IPTG modulations also correlate well with the trends of the experimental results (Fig. 3.8e).

While the overall behavior of our system is reminiscent of classical Turing patterns [50], there are key differences. In particular, when we ran simulations at the measured diffusion rate ratio of $\frac{D_v}{D_u} \approx 21.6$, patterns did not arise (Fig. 3.8f). For some two-node implementations of Turing systems, this rate would be sufficient for pattern formation [51]. In addition, certain networks with more nodes can allow small or even equal morphogen diffusion rate ratios to generate Turing instabilities [52]. However, a practical biological implementation imposes certain dynamics such as delays associated with protein production that can strongly impact pattern formation [53, 54]. Indeed, our deterministic modeling results suggest that the ratio of diffusion constants for the activator and inhibitor in our system is either barely within the range required for a Turing instability, or even outside the range, depending on the precise medium in which signal diffusion is measured. In addition, whereas in the deterministic simulation, spots are identical and evenly distributed, those in the experimental systems vary in size, shape, fluorescence intensity, and the intervals between them.

### 3.3.2 Stochastic Model

The deterministic modeling results indicate that our system may be beyond the regime where classical Turing patterns are formed, but still within the regime where stochastic Turing patterns occur [3, 4, 5, 29]. Indeed, gene expression in microbes is inherently noisy due to the small volume of cells and the fact that many reactants are present in low numbers, suggesting that stochastic Turing patterns could be present in our system [55].

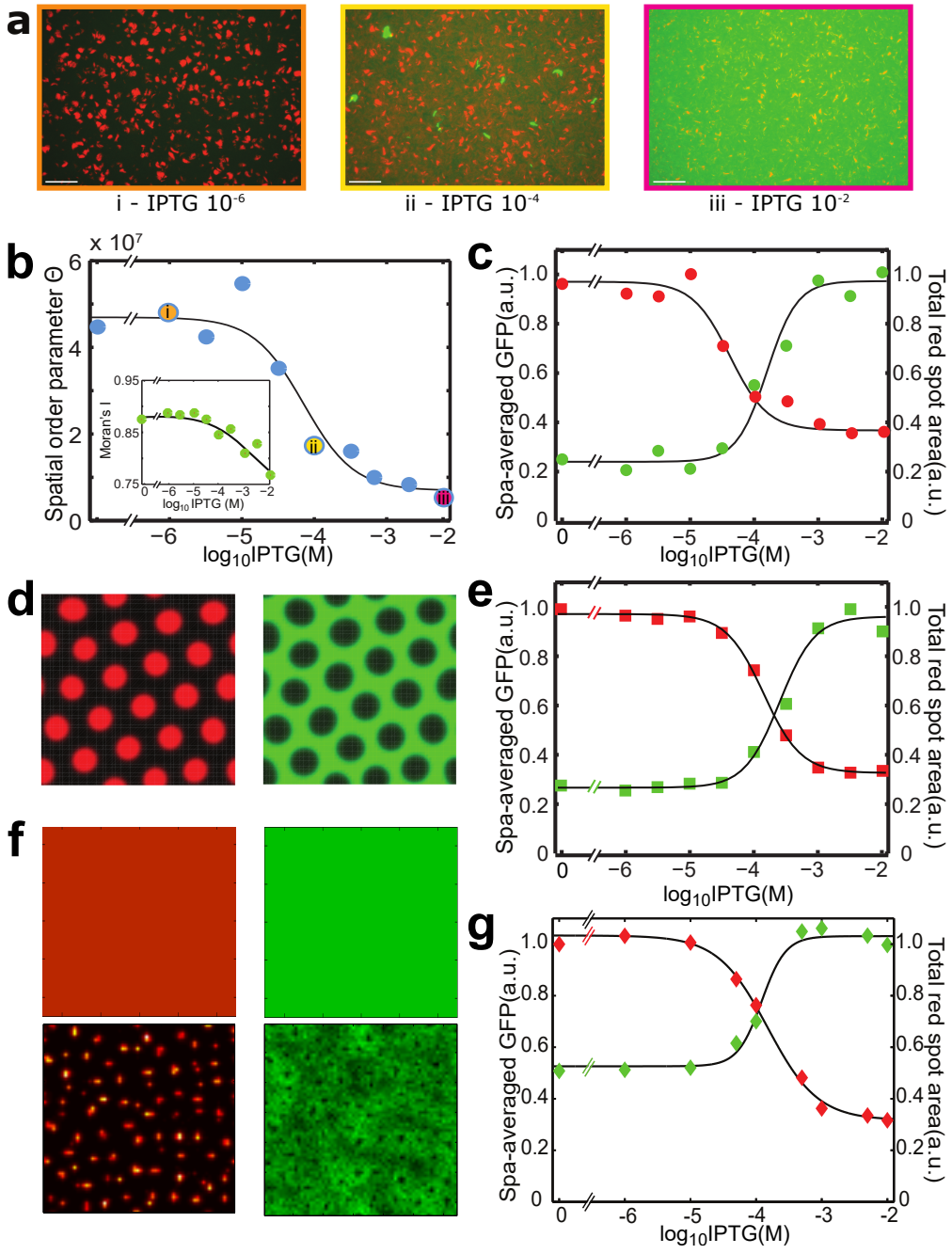


Figure 3.8: **Mathematical modeling and correlation between pattern modulation experiments and simulations**. **a**, Experimental results for IPTG modulation of pattern formation with microscopy images corresponding to specific IPTG concentrations in b. The same display mappings were used for all images in a. **b**, Collectivity metric parameter $\Theta$ is influenced by IPTG modulation. Inset: Moran's I (section 3.2.2) decreases with IPTG. **c**, Pattern statistics over IPTG modulation for experimental results. **d**, Pattern obtained from simulating a deterministic reaction-diffusion model with $D_v/D_u = 100$. **e**, Pattern statistics over IPTG modulation for deterministic modeling. **f**, Patterns obtained from simulating our deterministic model (top) and stochastic spatiotemporal model (bottom) at the measured diffusion ratio of $D_v/D_u = 21.6$. **g**, Pattern statistics over IPTG modulation for stochastic modeling.

Noise in stochastic Turing patterns expands the range of parameters in which patterns form, in contrast to the usual expectation that noise serves as a destabilizing agent. The patterns observed in stochastic Turing systems correspond to the slowest decaying mode of the fluctuations. Similar noise stabilization phenomena can be observed in other systems that are out of equilibrium. For example, in predator-prey systems, fluctuations can drive temporal oscillations of populations [56, 57]. Noise-driven stabilization has also been recently discovered in the clustering of molecules on biological membranes [58, 59] and in models that exhibit Turing-like pattern formation [5]. In particular, whereas spatial symmetry breaking and pattern formation via the original Turing design requires two morphogens with diffusion rates that differ by a large factor on the order of ten or a hundred [19], the requirements to form stochastic Turing patterns are less stringent. For example, in a pattern-forming plankton-herbivore ecosystem, the noise associated with discrete random birth and death processes reduces the required ratio of diffusion constants for pattern formation from a threshold of 27.8 for normal Turing patterns to a threshold of 2.48 for stochastic Turing patterns [3, 4, 5, 29].

To determine whether noise in the chemical reactions underlying gene expression and morphogen diffusion in our system can cause the emergence of patterns over a wider range of parameters than a deterministic model, we constructed a stochastic spatiotemporal model employing the same diffusion and rate constants used in our deterministic model. The patterning process is modeled with exactly the same biochemical reactions used in our deterministic model but simulated stochastically using an efficient tau-leaping stochastic algorithm[60, 61]. To speed up this large scale spatiotemporal simulation, we employ a hybrid technique where all intracellular chemical reactions are stochastic but signal diffusion is deterministic since the diffusion time scales are typically much faster than the intracellular reactions considered in our model. This model captures stochastic effects in the production and degradation of the proteins and morphogens in our system, yet approximates diffusion as deterministic. Simulations of the stochastic model generically produce patterns with large variability in spot size, shape, intensity, and intervals, which are similar to the patterns observed in our experiments, and different from those predicted for the deterministic model (Fig. 3.8f). We have compared the experimental patterns with stochastic simulations, in both real space and in two-dimensional Fourier transform (2DFT) space (Fig. 3.15). Neither the experimental 2DFT nor the simulated 2DFT contains pronounced peaks that would be present in a deterministic honeycomb Turing pattern. Moreover, as the IPTG concentration is increased, both experimental and simulated patterns become more regular (Fig. 3.8g).

To illustrate the behavior of each species in our pattern formation system (Table 3.1), we performed a spatial stochastic simulation using the parameters depicted in Tables 3.2-3.3. The top of Figure 3.9 shows $A_{3OC12HSL}$ and $I_{C4HSL}$ patterns produced in our stochastic simulation using the parameters given in tables Tables 3.2-3.3. The red line indicates the location of the cross-section used for all other dynamic variables. The bottom of Figure 3.9 shows cross-sectional slices of variables $U$ ($A_{3OC12HSL}$), $V$ ($I_{C4HSL}$), $I_u$ (LasI), $I_v$ (RhlI), $C$ (CI), $R_u$ (LasR), $R_v$ (RhlR), $L$ (free LacI), $X_1$ (LasR-$A_{3OC12HSL}$ complex), and $X_2$ (RhlR-$I_{C4HSL}$ complex).

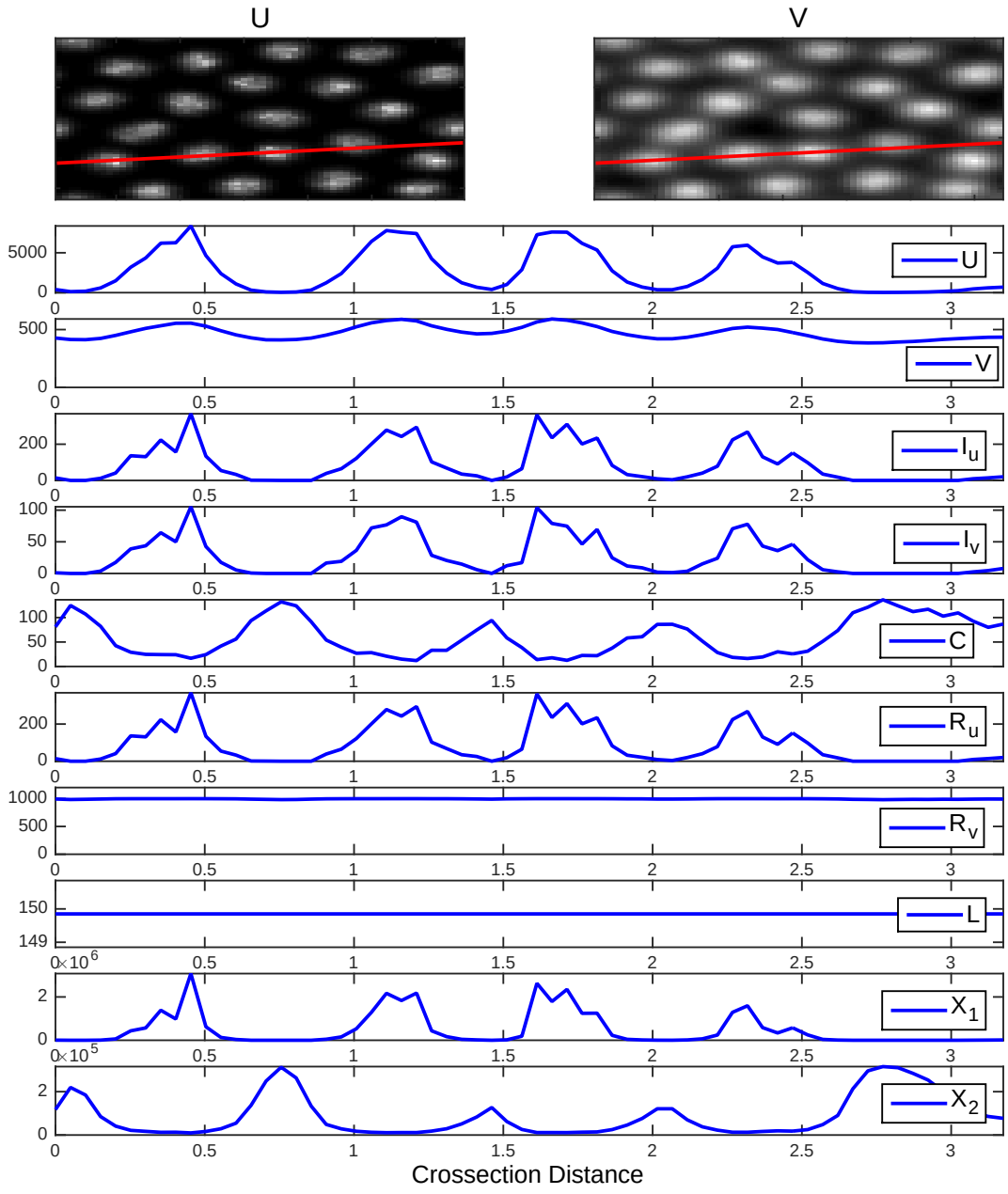


Figure 3.9: $A_{3OC12HSL}$ and $I_{C4HSL}$ patterns produced in our stochastic simulation using the parameters given in Tables 3.2-3.3. The red line indicates the location of the cross-section used for all other dynamic variables. Cross-sectional slices of variables $U$ ($A_{3OC12HSL}$), $V$ (IC4HSL), $I_u$ (LasI), $I_v$ (RhlI), $C$ (CI), $R_u$ (LasR), $R_v$ (RhlR), $L$ (free LacI), $X_1$ (LasR-$A_{3OC12HSL}$ complex), and $X_2$ (RhlR-$I_{C4HSL}$ complex).

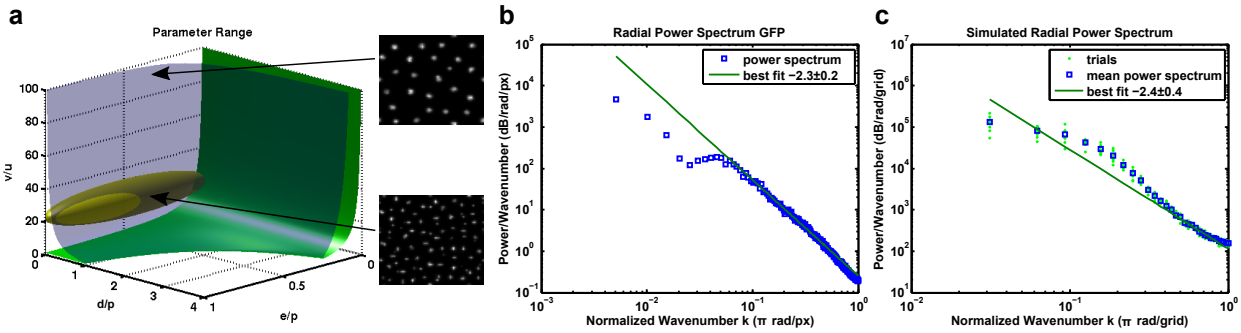


Figure 3.10: **Spectral analysis and parameter analysis**. (**a**) Pattern forming regimes in parameter space and estimated parameters for our system. Parameters above the green surface of neutral stochastic stability can form stochastic patterns and parameters above the blue surface of deterministic neutral stability can form deterministic Turing patterns. The ratio of the diffusion coefficients $\nu/\mu$, ratio of degradation rate to production rate $d/p$, and the ratio of production rates are estimated for our system by yellow ellipsoid. The parameters for our system are mostly in the regime where stochastic patterns form and outside the region where deterministic Turing patterns form. Example stochastic simulations are shown for parameters drawn from a deterministic parameter region with $D_\nu/D_\mu = 100$ (top) and a stochastic region with $D_\nu/D_\mu = 21.6$ (bottom). (**b**) Radial power spectrum of green fluorescence and best fit power law tail with an exponent of $-2.3 \pm 0.2$. (**c**) Radial power spectrum for 8 trials of our stochastic simulation, their mean, and the best fit power law tail.

As seen in Figure 3.9, the concentrations of LasI, RhlI, and LasR are proportional to the $A_{3OC12HSL}$ activator. This is due to the fact that these proteins are expressed from the $A_{3OC12HSL}$ activated $P_{Las-OR1}$ promoter. Likewise, the LasR-$A_{3OC12HSL}$ complex is directly proportional to $A_{3OC12HSL}$ concentrations. In addition, since RhlI catalyzes $I_{C4HSL}$ synthesis, $I_{C4HSL}$ is also directly proportional to $A_{3OC12HSL}$ concentrations. However, since $I_{C4HSL}$ diffuses faster than $A_{3OC12HSL}$, relatively high concentrations of $I_{C4HSL}$ are also found in between the $A_{3OC12HSL}$ activation domains.

Interestingly, the RhlR-$I_{C4HSL}$ complex is inversely proportional to $A_{3OC12HSL}$. In regions of high $A_{3OC12HSL}$, $A_{3OC12HSL}$ competitively binds RhlR, lowering the concentration of the RhlR-$I_{C4HSL}$ complex. However, as mentioned, $I_{C4HSL}$ concentrations remain relatively high outside of the $A_{3OC12HSL}$ activation domains. Thus, RhlR-$I_{C4HSL}$ is highest in between the activation domains. Collectively, this behavior results in green fluorescence (following the RhlR-$I_{C4HSL}$ complex concentration) surrounding red fluorescent activation domains (following the LasR-$A_{3OC12HSL}$ complex concentration).

To characterize the stochastic simulation we calculated the distribution of spot sizes and spacing. We binarized the simulation data shown in Figure 3.18 and determined locations of the centroids of spots and the areas of the spots. The spacing is calculated by finding the distance to the nearest neighboring centroid. The distribution of spot sizes and spacing is shown in Figure 3.19. As the IPTG concentration is increased spot sizes decrease and have more variance. Similarly the spacing between spots decreases as the IPTG concentrations is increased.

Our analysis of the stochastic Turing model predicts that stochastic patterns form over a wide range of parameters.

Indeed, our stochastic model predicts that stochastic Turing patterns are possible at the measured ratio of diffusion rates for $A_{3OC12HSL}$ and $I_{C4HSL}$ (Fig. 3.10a, Fig. 3.8f). To determine the sensitivity of the stochastic model to the parameters chosen and the range in which stochastic patterns will form, we individually varied parameters from half their nominal value to 1.5x their nominal value while keeping all other parameters fixed at their best estimated value. (Fig. 3.12) For each set of parameters we calculate the analytical power spectrum and the eigenvalues of the Jacobian (linear stability matrix) of the stochastic model evaluated at a fixed point found numerically. In this analysis, we classify each set of parameters as either producing an unstable homogeneous state at wavenumber $k = 0$, a stable homogeneous state, a stochastic Turing pattern, or a deterministic Turing pattern. Specifically, we classify a set of parameters as producing a pattern if they produce a peak in the calculated power spectrum at a nonzero wavenumber. To distinguish between stochastic Turing patterns and deterministic Turing patterns we examine the eigenvalues of the corresponding Jacobian. If the real part of all the eigenvalues is negative for all wavenumbers then the pattern must be due to stochasticity. If there is any range of wavenumbers that have corresponding positive real parts of their eigenvalues then the pattern is produced by the traditional Turing mechanism.

The results of this analysis are shown in Figure 3.11 and illustrate the significant ranges for each parameter that can lead to stochastic Turing patterns. Indeed, the estimated parameter values yield stochastic Turing patterns and variation of $D_u$, $D_v$ and IPTG, as well as several other parameters, never produce deterministic patterns. So our results are very insensitive to estimation error of these important parameters. Overall, varying the parameters one at a time, 68% of the values yield stochastic Turing patterns. In addition to calculating the phase diagram we quantified the sensitivity of the pattern to each parameter by calculating the change to the maximum eigenvalue (Fig. 3.12).

To attempt to quantify the way in which stochasticity enlarges the pattern forming regime of parameter space, we simultaneously varied all model parameters and performed the classification used above. Specifically, we used Latin hypercube sampling to randomly generate 500 parameter sets where all of the parameters were allowed to vary between half their nominal value and 1.5x their nominal value. For this analysis we found that 24.8% of parameters produced unstable fixed points, 43.2% produced stable homogeneous states, 13.2% produced stochastic Turing patterns, and 18.8% produced Turing patterns. Thus, over this arbitrarily large range of parameters, pattern formation occurs only 18.8% of the time in the absence of stochasticity, but 32% of the time when stochasticity is included. Accordingly, by including stochasticity, the range in which patterns can form has been increased by 70%.

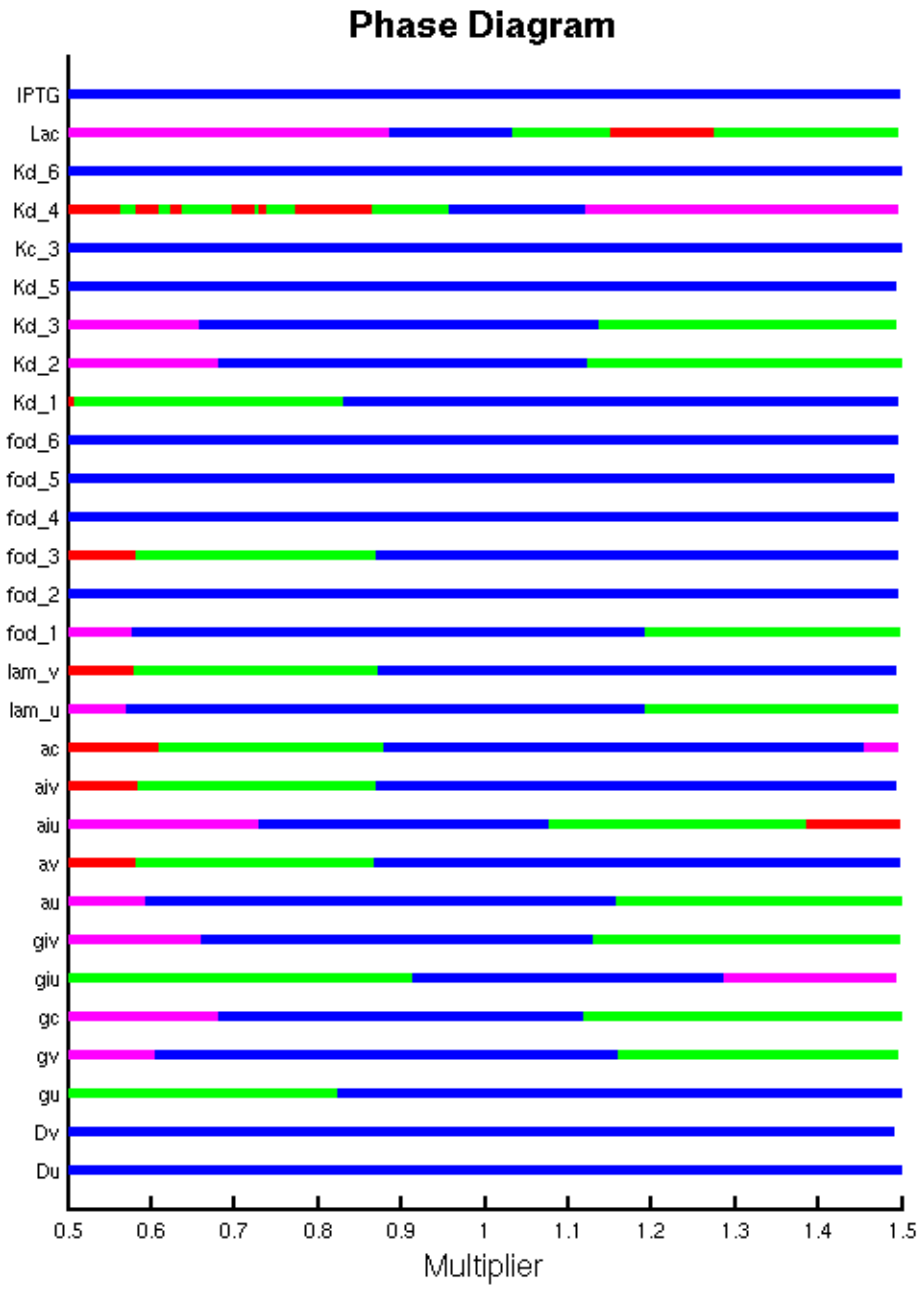


Figure 3.11: Phase Diagram showing the type of phase as each parameter is varied from half of its nominal value to 1.5x its nominal value while keeping all other parameters fixed. Red indicates an unstable fixed point, magenta a stable homogeneous state, blue a stochastic pattern, and green a deterministic Turing pattern.

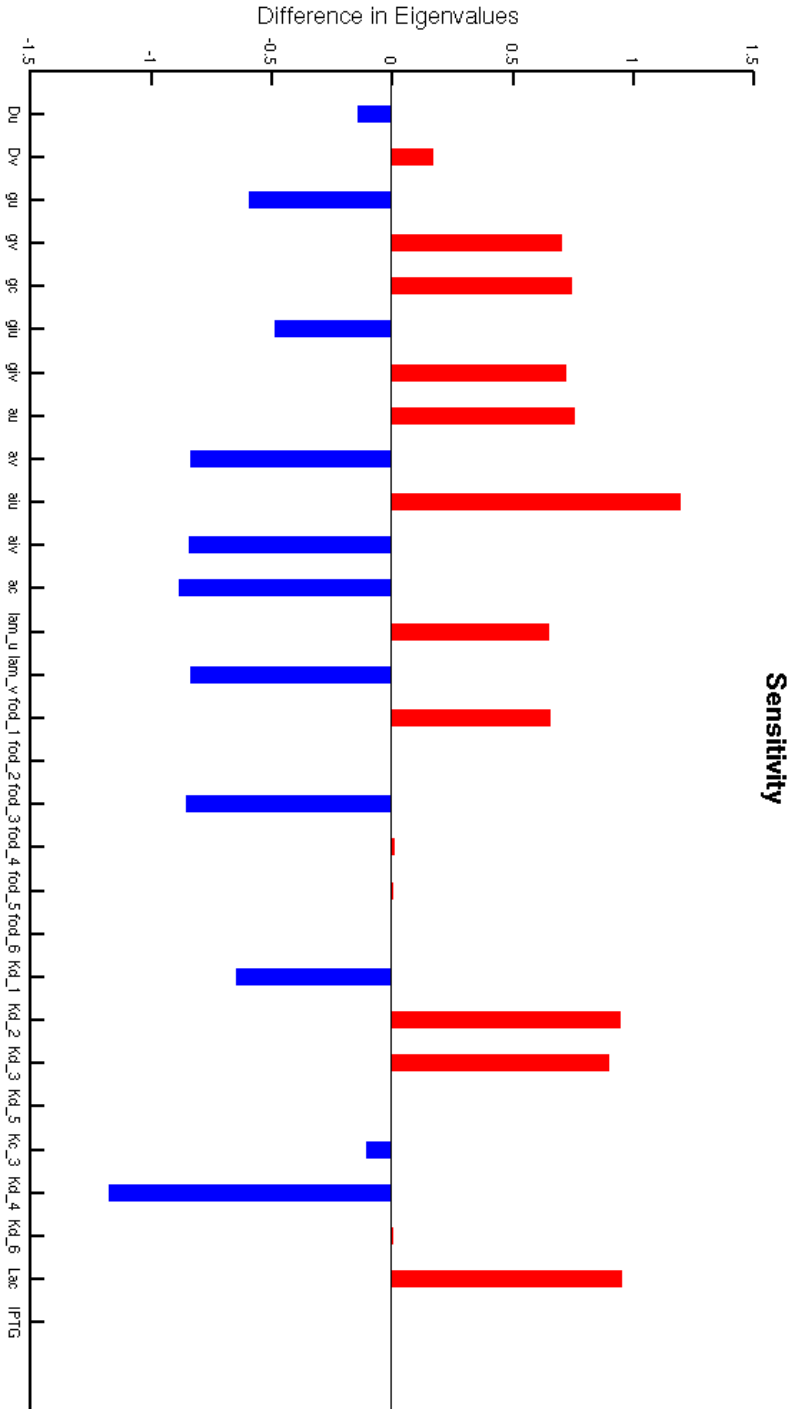


Figure 3.12: Sensitivity of a phase to a parameter is indicated by plotting the difference in eigenvalues between 1.5x the nominal value of a parameter and half of the nominal value. Red indicates a parameter that when increased promotes traditional Turing patterns and blue indicates a parameter promoting stochastic patterns.

To calculate the analytic power spectrum we used a method similar to that described in section 3.4. We wrote down the transition probabilities for the stochastic model directly from our deterministic model. For example, the transition probability for $U$ gaining a particle is $T(U \to U+1) = \alpha_u I_u$ and the transition probability for $U$ losing a particle is $T(U \to U-1) = \gamma_u U$. Using a system size expansion one can derive Langevin equations governing the fluctuations of the form

$$\partial_t x = Ax + \xi \text{ where } \left\langle \xi(t)\xi^\dagger(t') \right\rangle = B\delta(t-t'), \tag{3.14}$$

where in the case of this model

$$\begin{aligned} A &= J - \mathrm{diag}([D_u k^2, D_v k^2, 0, 0, 0]) && (3.15) \\ B &= \mathrm{diag}([\alpha_u I_u + \gamma_u U + D_u k^2 U, \alpha_v I_v + \gamma_v V + D_v k^2 V, \\ &\qquad \alpha_{iu} F_1(X_1, C) + \gamma_{iu} I_u, \alpha_{iv} F_1(X_1, C) + \gamma_{iv} I_v, \\ &\qquad \alpha_c F_2(X_2, L) + \gamma_c C]) && (3.16) \\ x^\dagger &= [\delta U, \delta V, \delta I_u, \delta I_v, \delta C], && (3.17) \end{aligned}$$

and $J$ is the Jacobian of the model evaluated at the fixed point. Using these equations the power spectrum is calculated to be $P(k, w=0) = \langle x x^\dagger \rangle = A^{-1} B (A^{-1})^\dagger$. The fixed point, Jacobian, and power spectrum are numerically calculated using a custom Matlab script. Figure 3.13 shows the calculated power spectrum corresponding to the parameters listed in Table 3.2. In this figure the full model produces power spectrum with a power law tail of $-2$ for the inhibitor and an initial power law tail of $-4$ for the activator before undergoing a crossover to a $-2$ power law. The eigenvalues of the spatially extended Jacobian, $A$, are plotted in Figure 3.14 showing that all eigenvalues are negative. This indicates that the set of parameters in Table 3.2 produces a stochastic Turing pattern.

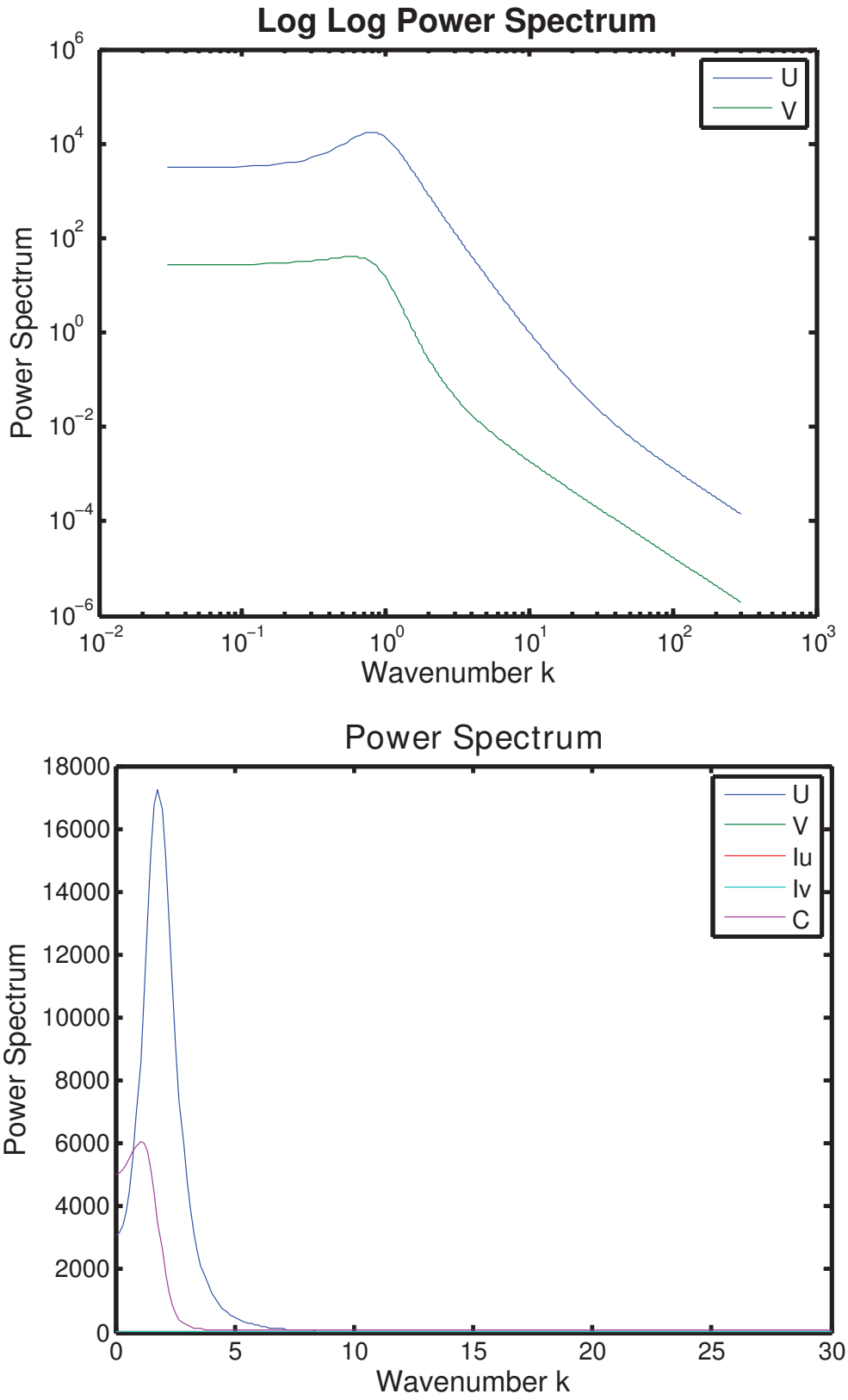


Figure 3.13: The Analytic power spectrum calculated for the parameter set given in Tables 3.2-3.3.

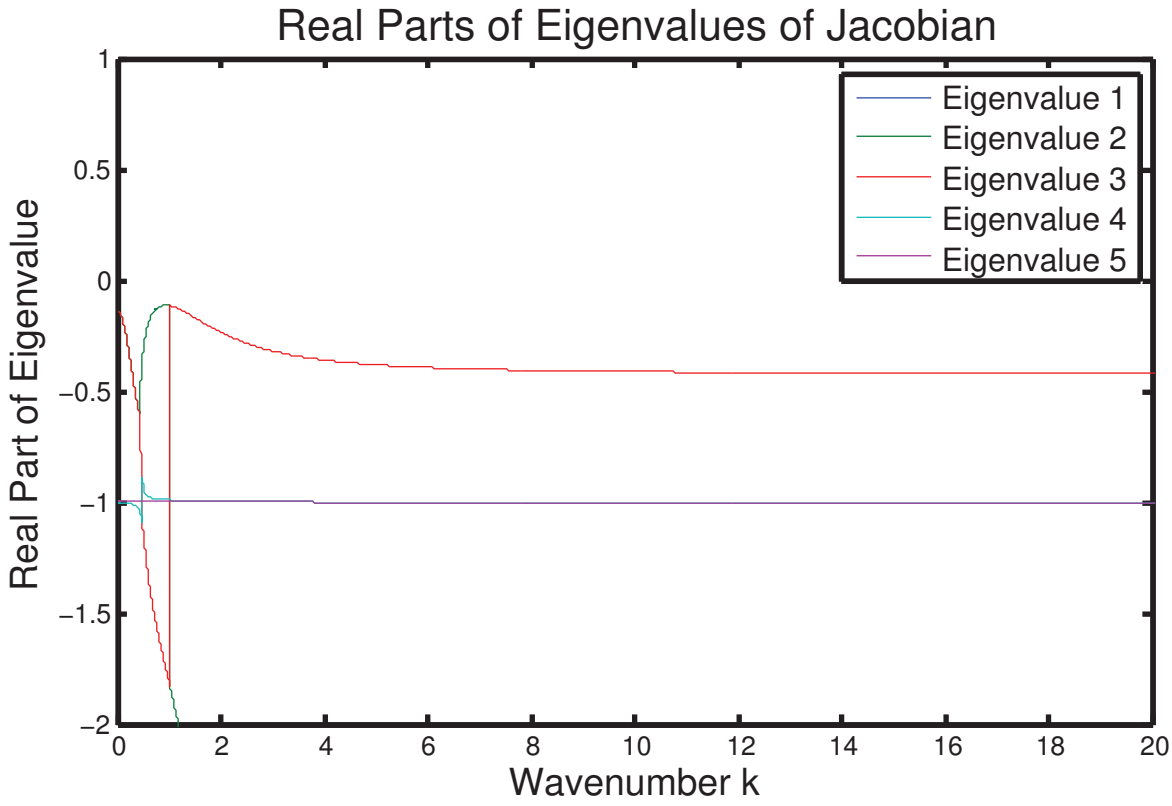


Figure 3.14: The real part of the eigenvalues of the Jacobian for the parameter set given in Tables 3.2-3.3 plotted as a funtion of k. All the eigenvalues are negative indicating that the pattern formed is stochastic.

### 3.3.3 Power Spectra Analysis of Experimental Observations

To further test the hypothesis that we are observing stochastic Turing patterns, we measured the power spectrum for both of our fluorescent reporters. Theory predicts that the power spectrum will have a power-law tail as a function of wavenumber, $k$, for large wavenumbers, with an exponent characteristic of the noise source [56, 5]. The exponent value is -2 and -4 respectively for stochastic Turing patterns and deterministic Turing patterns with additive noise, and can be interpreted simply as follows. The -2 arises because, at small frequency or wavenumber, the random variable (i.e., concentration) is simply diffusing and so follows the behavior of a random walk, whose power spectrum exhibits a -2 power law. The -4 arises because for a system that is executing deterministic damped periodic motion but driven by additive white noise, the response of the random variable is a Lorentzian, with an asymptotic behavior for the power spectrum that exhibits a -4 exponent.

We converted the pictures of red and green fluorescent proteins into gray scale images and subtracted off the mean intensity to obtain data corresponding to the fluctuations. We then conducted a discrete two-dimensional (2D) Fourier transform of the data, finding its amplitude squared. Since it is clear from the resulting Fourier transforms that the

patterns are isotropic, we perform an angular average (Figures 3.6a, 3.15). For the GFP channel, we observe a power-law tail, with an exponent of $-2.3 \pm 0.4$ (Fig. 3.10b), consistent with predictions for demographic noise. It is possible to obtain anomalous power law tails in the power spectrum due to discontinuities in the boundaries of the picture, but these artifacts are distinguishable by their lack of noise, and we are confident that such spectral leakage is not being observed in these data. For the RFP channel, we also observe a power-law tail with an exponent of $-3.9 \pm 0.4$. (Fig. 3.6b,d).

To better understand the implications of these tails, we examined our detailed stochastic model of the genetic systems and also developed a reduced stochastic model that explicitly includes only the morphogens (see section 3.4). Both models predict that our experimental parameters will produce a stochastic pattern with a power law tail of $-2$ for both the activator and inhibitor at asymptotically large wavenumbers (Fig 3.10c, and sections 3.4, 3.4.1). However, in the range of parameters likely to correspond to the experiments ( Fig. 3.10a), the detailed stochastic model predicts that the exponent of the power-law tail for the activator will be $-4$ over a large range of intermediate wave-numbers before it eventually undergoes a crossover to a power-law with an exponent $-2$ at high wave-numbers (Fig. 3.13). This behavior once again agrees with our experimental data and supports our identification of stochastic Turing patterns. In summary, spectral analysis of the patterns of activator and inhibitor is consistent with a model in which fluctuations in the amount of signalling morphogens drive stochastic Turing patterns. Radial power spectra were also calculated for other concentrations of IPTG as shown in Figures 3.16 and 3.17.

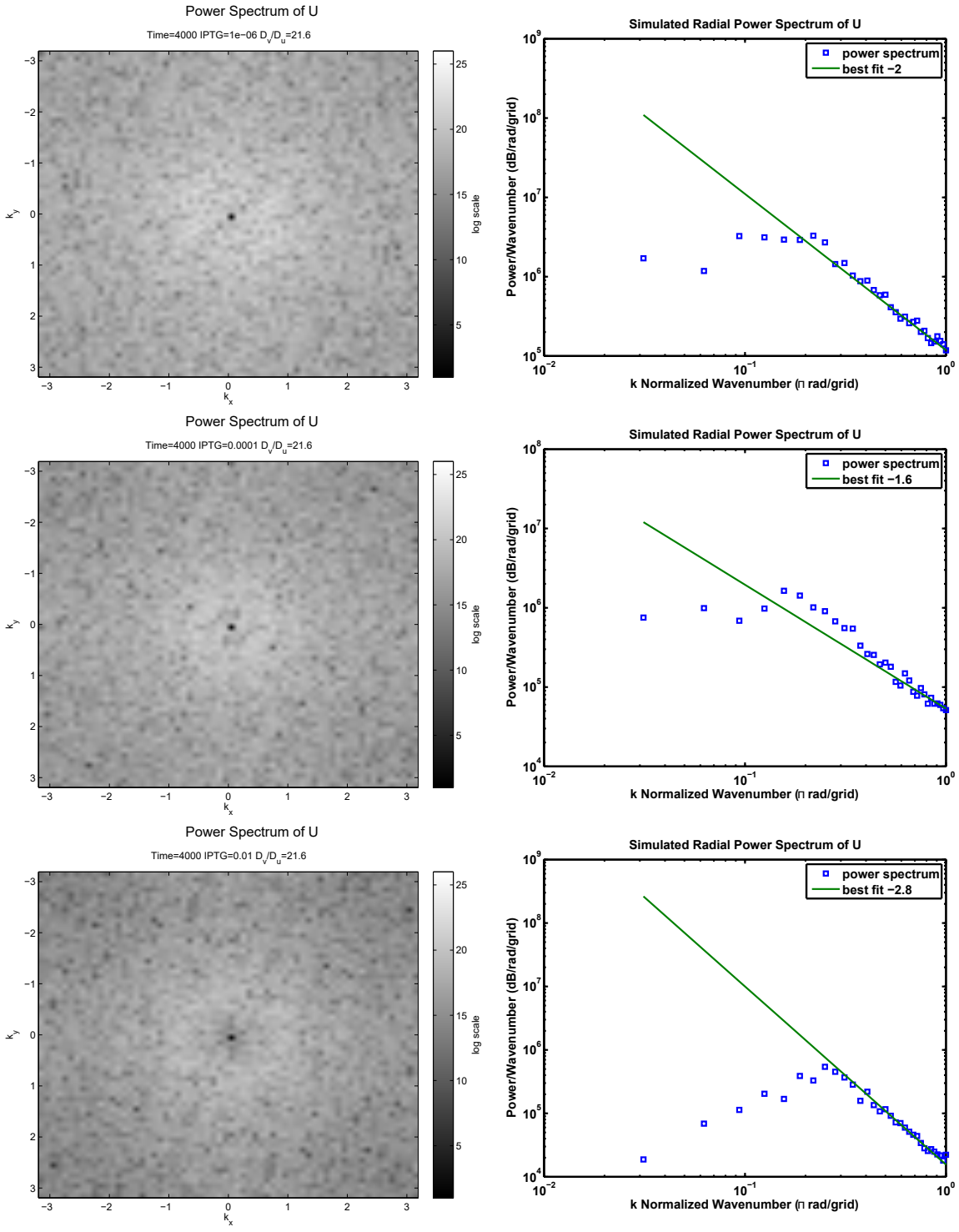


Figure 3.15: 2D power spectrum and radial power spectrum for 30C12HSL produced in our stochastic simulation using $D_v/D_u = 21.6$ for three different values of IPTG.

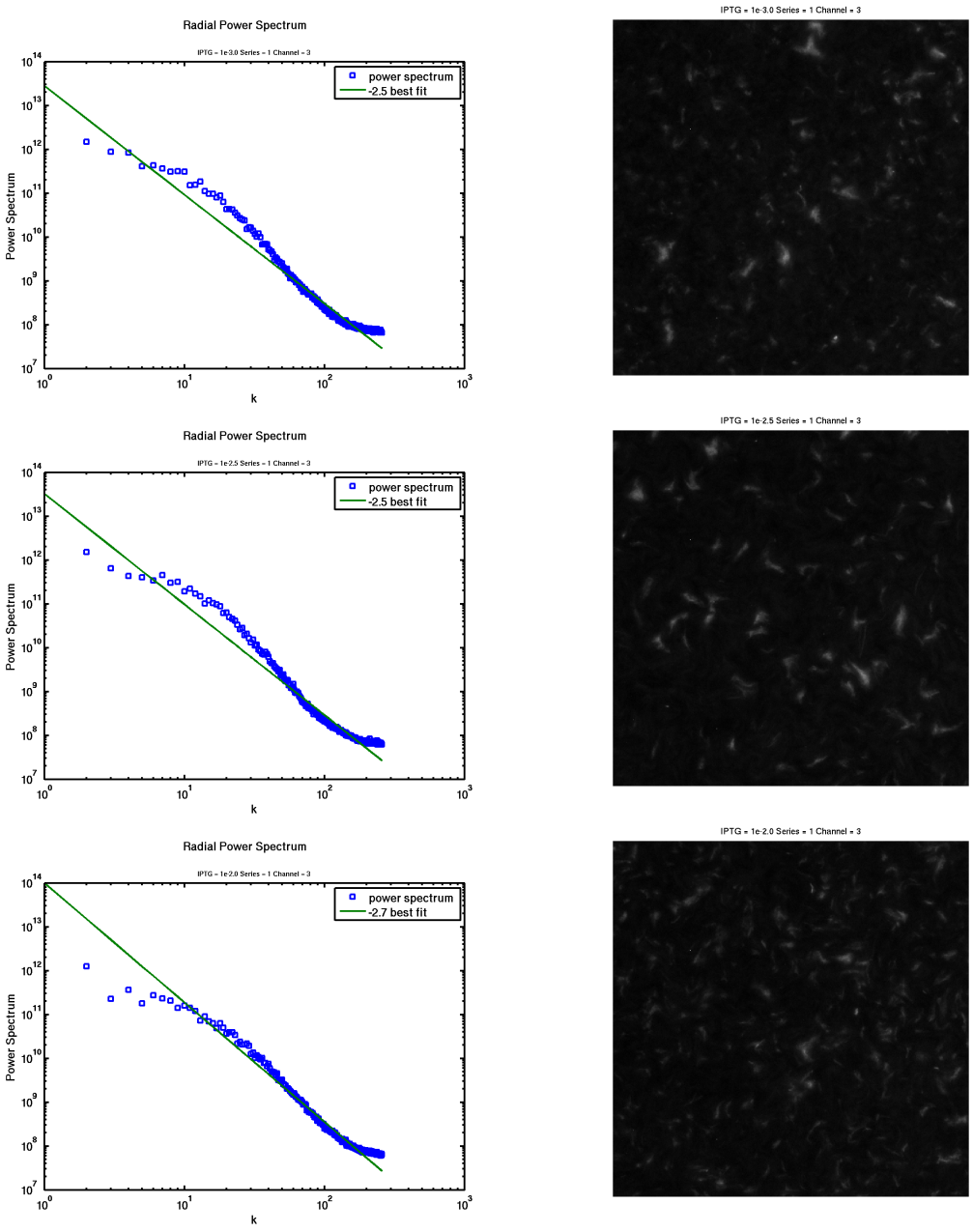


Figure 3.16: RFP images and their corresponding Radial Power spectrums with powerlaw tail fits of -2.5, -2.5, -2.9 for IPTG concentrations of $10^{-3}$ M, $10^{-2.5}$ M, and $10^{-2}$ M, respectively.

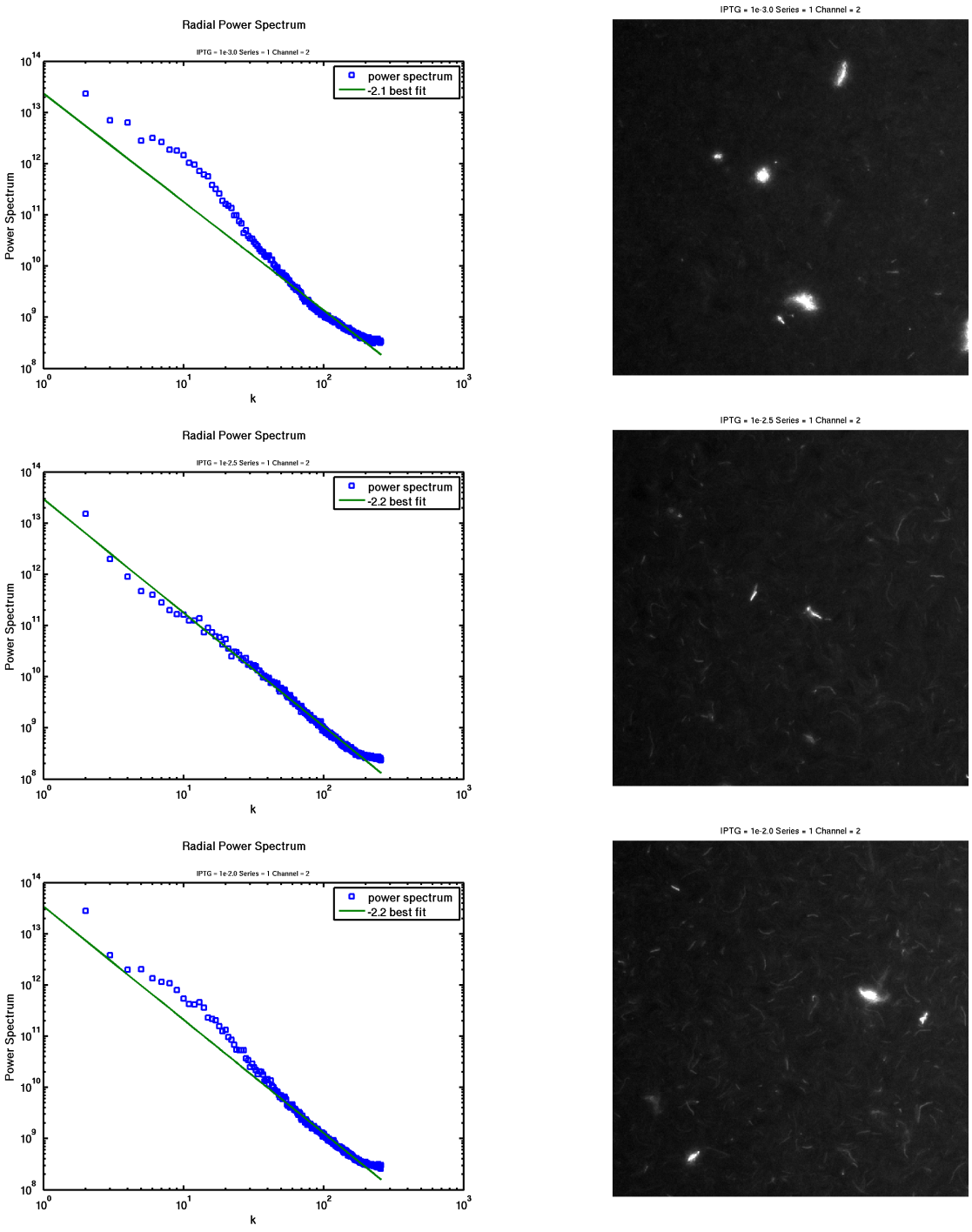


Figure 3.17: GFP images and corresponding Radial Power spectrums with powerlaw tail fits of -2.1, -2.2, -2.2 for IPTG concentrations of $10^{-3}$ M, $10^{-2.5}$ M, and $10^{-2}$ M, respectively. At IPTG concentrations smaller than $10^{-4}$ M the green channel begins to look spatially homogeneous as the signal is not strong enough to show-up beyond the background camera noise.

## 3.4 Reduced Model for Stochastic Turing Patterns

The traditional Turing mechanism usually consists of at least two chemicals. One of the chemicals is a slowly diffusing activator, activating the synthesis of itself and the inhibitor. The other chemical is a fast diffusing inhibitor, inhibiting synthesis of the activator and itself. The Turing mechanism can be explained by a simple qualitative argument consisting of three steps. Initially, activator and inhibitor are distributed randomly. Areas with local concentrations of activator will autocatalytically grow, forming dense clumps of activator. Inhibitor will also be produced near these clumps of activator. The rapidly diffusing inhibitor will suppress the spread of the clumps of activator. This simple picture of activator-inhibitor dynamics does not require large separation of diffusion rates or depend on details of rates. When this system is considered classically, however, it is found to either require fine tuning of reaction rates or have a large separation of diffusion rates. We will see shortly that the stochastic treatment solves this fine tuning problem.

Our synthetic system is designed to implement an activator-inhibitor system with $A_{3OC12HSL}$ as an activator of its own synthesis and that of $I_{C4HSL}$, while $I_{C4HSL}$ is an inhibitor of both chemicals. To develop a simplified model of the activator-inhibitor circuit, one can write down reaction diffusion equations for this system that are mathematically equivalent to the Levin-Segel model of herbivore-plankton interaction [16].

$$\partial_t \phi = \mu \nabla^2 \phi + b\phi + e\phi^2 - p\psi\phi \tag{3.18}$$

$$\partial_t \psi = \nu \nabla^2 \psi + p\psi\phi - d\psi^2 \tag{3.19}$$

Here $\phi$ is the concentration of activator ($A_{3OC12HSL}$), $\psi$ is the concentration of inhibitor ($I_{C4HSL}$), $\mu$ and $\nu$ are the the diffusion constants of the activator and inhibitor, respectively. The term $p\psi\phi$ is a competition term, $e\phi^2$ is a nonlinear activation term, and $-d\psi^2$ is a nonlinear self inhibition term for the inhibitor. These equations exhibit a Turing instability when [16]:

$$\frac{\nu}{\mu} > \left( \frac{1}{\sqrt{p/d} - \sqrt{p/d - e/p}} \right)^2. \tag{3.20}$$

To consider the effects of extrinsic noise we can simply add white noise $\xi$ to the system of equations above. Then we can construct the Fourier transformed stability matrix and solve for the power spectrum $\hat{P}(k,w)$ where $k$ is the wave-number and $w$ is frequency. When this is done it is found that the power spectrum has a power law tail, $\hat{P}(k, w=0) = \frac{k^\sigma}{\nu^2} \langle \xi\xi \rangle, k \gg k_m$ with an exponent $\sigma = -4$ [5]. In addition, the condition for pattern formation becomes $\frac{\nu}{\mu} > \frac{p}{e}$.

Intrinsic noise, represented for example by copy number fluctuations arising from stochastic gene expression, can be studied by writing down an individual level model for the activator and inhibitor [5]. The following set of chemical

reactions describe the effective activator-inhibitor system that we engineered:

$$\begin{array}{lll} A \xrightarrow{b} A+A & AI \xrightarrow{p_1/V} I & I+I \xrightarrow{d/V} I \\ A+A \xrightarrow{e/V} A+A+A & AI \xrightarrow{p_2/V} I+I & \end{array} \tag{3.21}$$

Here $A$ is the activator ($A_{3OC12HSL}$) and $I$ is the inhibitor ($I_{C4HSL}$) and $V$ is the well-mixed patch size setting the strength of the fluctuations. From these first order reactions one can derive a Master Equation governing the probability of having $m$ molecules of $I$ and $n$ molecules of $A$ at a given time:

$$\begin{aligned} \partial t P(m,n) =& b[(n-1)P(m,n-1) - nP(m,n)] + \frac{e}{V}[(n-1)(n-2)P(m,n-1) - n(n-1)P(m,n)] \\ &+ \frac{p_1}{V}[m(n+1)P(m,n+1) - mnP(m,n)] + \frac{p_2}{V}[(m-1)(n+1)P(m-1,n+1) - mnP(m,n)] \\ &+ \frac{d}{V}[(m+1)mP(m+1,n) - m(m-1)P(m,n)] \end{aligned} \tag{3.22}$$

To introduce spatial variations, the master equation can be conveniently represented as a path integral, from which mean-field equations and Langevin equations with multiplicative noise governing the fluctuations are obtained by Van Kampen expansion [5]. We calculate the power spectrum $\hat{P}(k,w)$ as a function of frequency $w$ and wave-number $k$, obtaining a power spectrum tail ($k \gg k_m$): $\hat{P}(k,w=0) \approx \frac{\psi}{\nu} k^\sigma$ (inhibitor), $\hat{P}(k,w=0) \approx \frac{\phi}{\mu} k^\sigma$ (activator) with $\sigma = -2$. Patterns arise when

$$\frac{\nu}{\mu} > \left(\frac{p}{e}\right) \frac{5+7(de/p^2)}{4+5(de/p^2)+3(de/p^2)^2}. \tag{3.23}$$

Additionally the wavelength, $\lambda$, or characteristic spacing of the pattern turns out to be the same as that for a classical Turing pattern [5] and is found to be

$$\lambda = 2\pi \sqrt{\frac{2}{\phi} \frac{\mu\nu}{e\nu - p\mu}}. \tag{3.24}$$

### 3.4.1 Stochastic Model Power Spectra Analysis

We now discuss the origin of the exponent discussed above, from the standpoint of two models that we have constructed to analyze the data. The first model is a coarse-grained phenomenological or minimal model for the morphogens. The second model is a detailed stochastic model, described in section 3.3.2. We will see that there are two factors contributing to the -2 exponent in the inhibitor channel and the -4 exponent in the activator channel, in the context of intrinsic noise. Note that extrinsic noise, if present, would lead to an exponent of -4, but in both channels. The first explanation is that the activator channel had a very sharp threshold for expressing the red fluorescence and the pictures were overexposed. This overexposure can suppress the small-scale correlations in the patterns. This idea

is supported by taking data from simulations of demographic-noise induced patterns and performing image operations on the data to simulate this effect. We ran the simulated data through an image dilation morphological operator with a disk structural element 1px in radius. The resulting power spectra have power law tails with $\sigma = -4$, even though the original data were intrinsic noise-induced, having a power law tail with $\sigma = -2$. Thus, an asymptotic power-law tail of -2 in the power spectrum appears as an effective exponent of -4 in the presence of overexposure. This is further supported by looking at the spectra in Figures 3.16 and 3.17. These images, taken at other IPTG concentrations, were not overexposed and their spectra have power law tails with exponents closer to -2.

The second explanation uses the detailed stochastic model described in section 3.3.2. This model also predicts power-law tails with $\sigma = -2$ for both the activator and inhibitor, at asymptotically large values of $k$, because, when coarse-grained, the model should be well-described by the phenomenological model for the morphogens alone. But for the range of parameters that we estimate are consistent with the experimental data, the detailed model also predicts that for a wide range of intermediate wavenumbers, there is an effective power-law with $\sigma = -4$ for the activator, before undergoing a crossover to a power law with $\sigma = -2$ at high wave-numbers. Our interpretation is that the experiment is indeed well-described by the detailed model, and that we are observing the behavior before the crossover point in the power spectrum of the RFP channel. Overexposure does not cause any additional change to the effective -4 exponent.

The measured diffusion ratio of $\frac{\nu}{\mu} = 21.6$ is too small to produce classical Turing patterns. In fact, to produce patterns qualitatively similar to the ones observed, the diffusion constants must be separated by a factor on the order of 100 in our non-stochastic simulation (see Fig. 3.8f). We can also plot the estimated range of parameters for our effective model and compare them to regimes where normal Turing patterns form and stochastic Turing patterns form (Fig. 3.6e). We see from this plot that the estimated parameters fall mainly in the regime where stochastic patterns form, but not where normal Turing patterns can form. Any parameters above the blue surface will form classical Turing patterns. Any parameters above the green surface can form stochastic Turing patterns. The yellow oval representing our estimated range of effective parameters falls mainly below the blue surface but is above the green surface, indicating most of the parameters fall within the regime of stochastic patterns. We estimated the values and ranges of the ratios $\frac{\nu}{\mu}$, $\frac{e}{p}$, and $\frac{d}{p}$ which solely control pattern formation in the reduced model. In our analysis we used the the experimentally measured ratio of diffusion constants $\frac{\nu}{\mu} = 21.6 \pm 10$ and we estimated $\frac{e}{p} < 1$ by using the knowledge that for any pattern to form, either classical or stochastic, the homogeneous state must first be stable so $p > e$. Finally, since the degradation rate is always smaller than the rate of production of our molecules we estimated $\frac{d}{p} < 1$. Even at our measured values of diffusion constants our stochastic simulation continued to produce patterns similar to the ones observed in our experiment and produced a power spectrum power law tail with $\sigma - 2.4$ consistent with our experiment.

## 3.5 Discussion

### 3.5.1 Alternative Hypotheses

Now we consider alternative hypotheses to our claim that the theory of stochastic Turing patterns explains our experimental observations. First, we consider the duration and dynamics of our pattern formation experiments. One may expect to observe early events in Turing pattern formation such as splitting of clusters or increases in inter-cluster distances. These processes may be in fact be taking place, but may be difficult to observe due to weak reporter expression in the earlier stages. In addition, we must consider the limited duration of our experiments and the possibility that, theoretically, longer observations may result in different patterns if nonlinear processes eventually began to dominate dynamics. Indeed, we do not feed fresh nutrients to sustain the system for extremely long durations. However, as confirmed by analysis of the dynamics in Fig. 3.5, cluster size growth and spacing between clusters appears to be stabilizing towards the end of the experiment. In addition, domains are neither created nor destroyed in the later time periods. Essentially, it appears that the patterns are close to stabilizing within the 32 hour observation period.

Another alternative hypothesis is that cell growth dynamics primarily drive the observed pattern formation. Our control experiments with mixtures of red and green cell populations (Fig. 3.2), along with our bistable switch control (Fig. 3.3), suggest that cell growth does not explain our patterns. Moreover, our ability to tune pattern characteristics offers support for the fact that our patterns are not a simple consequence of natural biofilm growth morphologies, but are rather driven by our genetic circuit. However, growth may indeed impact regularity and may likely explain the fact that our experimental patterns are less regular than those observed in our stochastic models (Fig. 3.8f). Indeed, future experiments to demonstrate different classes of patterns, e.g., labyrinth patterns, would offer further support, but collectively, our experiments strongly support our hypothesis that a Turing mechanism driven by our genetic circuit explains our observed patterns.

### 3.5.2 Summary of Evidence for Stochastic Turing Patterns

We summarize our evidence for having demonstrated stochastic Turing patterns and not amplification of random noise as follows. Our control experiments with mixtures of red and green cells (Fig. 3.7b, Fig. 3.2), along with a bistable switch (Fig. 3.3), did not produce the patterns that we observe with our genetic circuit (Fig. 3.7a). Our ability to tune pattern characteristics offers further support that pattern formation is driven by our genetic circuit. In addition to our experimental controls, we identify patterns in the stochastic model, but not the deterministic model of our system for the experimentally observed ratio of diffusion rates (Fig. 3.8f). These model patterns resemble the experimentally observed patterns in real space, exhibit no peaks in the 2DFT (Figs. 3.6 + 3.15), and recapitulate the observed trend with IPTG variation. Analysis of our experimental data is also in accord with the theory of stochastic Turing patterns.

The exponents in the tails of the experimental radial power spectra agree with theoretical predictions (Fig. 3.10b, Fig. 3.6). In addition, although spatial regularity is weak, we observe a radial spectral peak for our experimental patterns (Fig. 3.10b, Fig. 3.6), indicating a characteristic length scale. Furthermore, exploration of the large parameter space of the stochastic model indicates that the experimental parameters are most likely to be in the regime where only stochastic patterns can form (section 3.3.2). Collectively, this body of evidence suggests that our experiments indeed exhibit stochastic Turing pattern formation.

## 3.6 Supplement: Tables and Figures

Table 3.1: Variables used in the model.

| Symbol | Molecule |
|---|---|
| $U$ | $3OC_{12}HSL$ |
| $V$ | $C_4HSL$ |
| $I_u$ | LasI |
| $I_v$ | RhlI |
| $C$ | CI |
| $R_u$ | LasR |
| $R_v$ | RhlR |
| $L$ | free LacI |
| $X_1$ | LasR-$3OC_{12}HSL$ complex |
| $X_2$ | RhlR-$C_4HSL$ complex |

Table 3.2: Definitions and values for the rate constants used in our mathematical model.

| Parameter | Description | Value | Unit |
|---|---|---|---|
| $\alpha_u$ | $A_{3OC12HSL}$ production rate | $3.0 \times 10^1$ | $hr^{-1}$ |
| $\gamma_u$ | $A_{3OC12HSL}$ degradation rate | 1.0 | $hr^{-1}$ |
| $D_u$ | $A_{3OC12HSL}$ diffusion coefficient | $5.0 \times 10^{-1}$ | $grid^2/hr$ |
| $\alpha_v$ | $I_{C4HSL}$ production rate | $3.0 \times 10^1$ | $hr^{-1}$ |
| $\gamma_v$ | $I_{C4HSL}$ degradation rate | 1.0 | $hr^{-1}$ |
| $D_v$ | $I_{C4HSL}$ diffusion coefficient | 10.8 or 50* | $grid^2/hr$ |
| $\alpha_{iu}$ | Basal production rate of LasI | $1.0 \times 10^1$ | molecules/hr |
| $\gamma_{iu}$ | Degradation rate of LasI | 1.0 | $hr^{-1}$ |
| $\alpha_{iv}$ | Basal production rate of RhlI | 0.3 | molecules/hr |
| $\gamma_{iv}$ | Degradation rate of RhlI | 1.0 | $hr^{-1}$ |
| $\alpha_c$ | Basal production rate of CI | $1.0 \times 10^1$ | molecules/hr |
| $\gamma_c$ | CI degradation rate | 1.0 | $hr^{-1}$ |
| $\lambda_u$ | Ratio between LasR and LasI | 1.0 | ⁄ |
| $\lambda_v$ | Steady state level of RhlR by $\lambda_{P(R-O1)}$ w/o CI regulation | $1.0 \times 10^3$ | molecules |
| $\lambda_l$ | Steady state level of LacI from $p_{lacq}$ expression | $1.5 \times 10^2$ | molecules |
| $K_{c3}$ | $A_{3OC12HSL}$-RhlR dissociation constant | $1.5 \times 10^2$ | molecules |
| $I$ | IPTG concentration | $1.0 \times 10^{-6\sim-2}$ | M |

Table 3.3: Additional definitions and values for the rate constants used in our mathematical model.

| Parameter | Description | Value | Unit |
|---|---|---|---|
| $\theta_1$ | Hill coeff. for LasR-$A_{3OC12HSL}$ complex activation of $P_{Las-OR1}$ | 1.0 | / |
| $K_{d1}$ | Disso. constant of LasR-$A_{3OC12HSL}$ complex with $P_{Las-OR1}$ | $1.0 \times 10^3$ | molecules |
| $f_1$ | Fold change for full induction of $P_{Las-OR1}$ | $1.0 \times 10^3$ | / |
| $\theta_2$ | Hill coeff. for CI repression of $P_{Las-OR1}$ | 2.0 | / |
| $K_{d2}$ | Disso. constant of CI with $P_{Las-OR1}$ | $1.0 \times 10^1$ | molecules |
| $f_2$ | Fold change for full inhibition of $P_{Las-OR1}$ | $1.0 \times 10^5$ | / |
| $\theta_3$ | Hill coeff. for RhlR-$I_{C4HSL}$ complex activation of $P_{Las-OR1}$ | 1.0 | / |
| $K_{d3}$ | Disso. constant of RhlR-$I_{C4HSL}$ complex with pRhl-lacO | $1.0 \times 10^5$ | molecules |
| $f_3$ | Fold change for full induction of pRhl-lacO | $1.0 \times 10^3$ | / |
| $\theta_4$ | Hill coeff. for the LacI activation of pRhl-lacO | 4.0 | / |
| $K_{d4}$ | Disso. constant of LacI with pRhl-lacO | $1.0 \times 10^2$ | molecules |
| $f_4$ | Fold change for full inhibition of pRhl-lacO | $1.0 \times 10^3$ | / |
| $\theta_5$ | Hill coeff. for the CI activation of $\lambda_{P(R-O1)}$ | 2.0 | / |
| $K_{d5}$ | Disso. constant of CI with $\lambda_{P(R-O1)}$ | $1.0 \times 10^3$ | molecules |
| $f_5$ | Fold change for full induction of $\lambda_{P(R-O1)}$ | $1.0 \times 10^5$ | / |
| $\theta_6$ | Hill coeff. for the IPTG binding to LacI | 1.0 | / |
| $K_{d6}$ | Disso. constant of IPTG with LacI | $1.0 \times 10^{-3}$ | M |
| $f_6$ | Fold change of LacI activity for IPTG full induction | $1.0 \times 10^5$ | / |

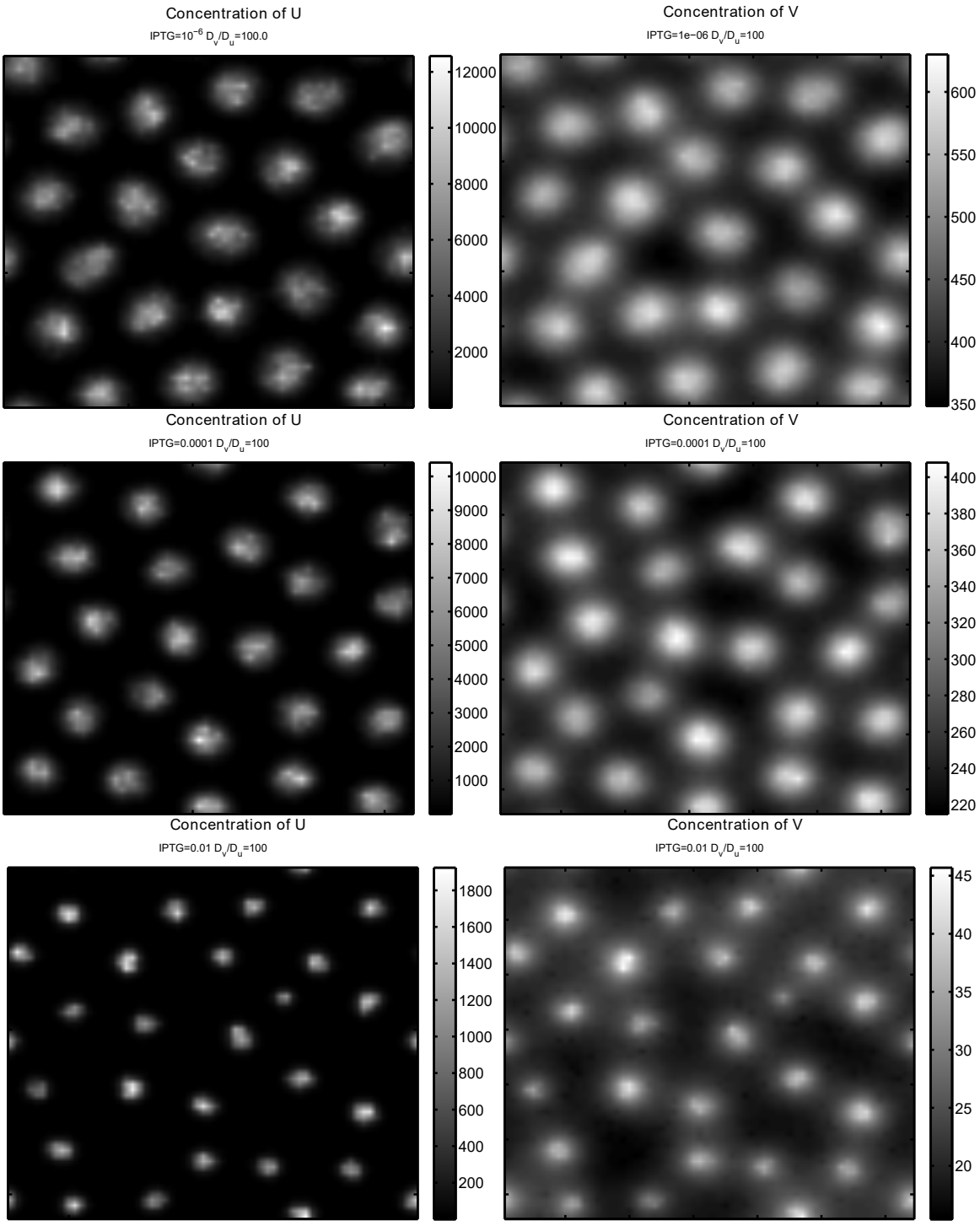


Figure 3.18: $A_{3OC12HSL}$ patterns produced in our stochastic simulation using $D_v/D_u = 100$ and the parameters given in Tables 3.2-3.3 for three different concentrations of IPTG.

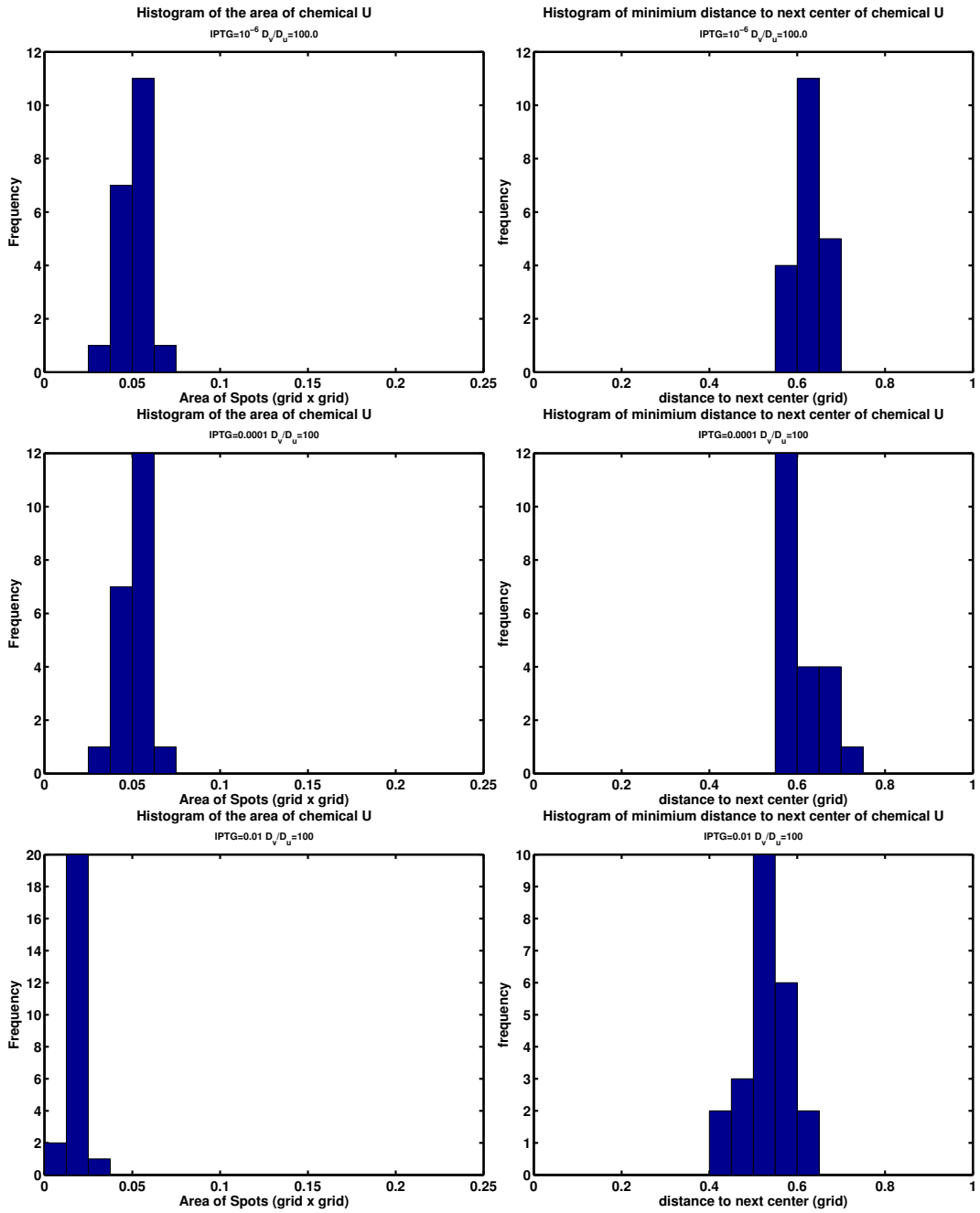


Figure 3.19: Spot size and spacing distributions for 30C12HSL produced in our stochastic simulation with $D_v/D_u = 100$ for three different values of IPTG.

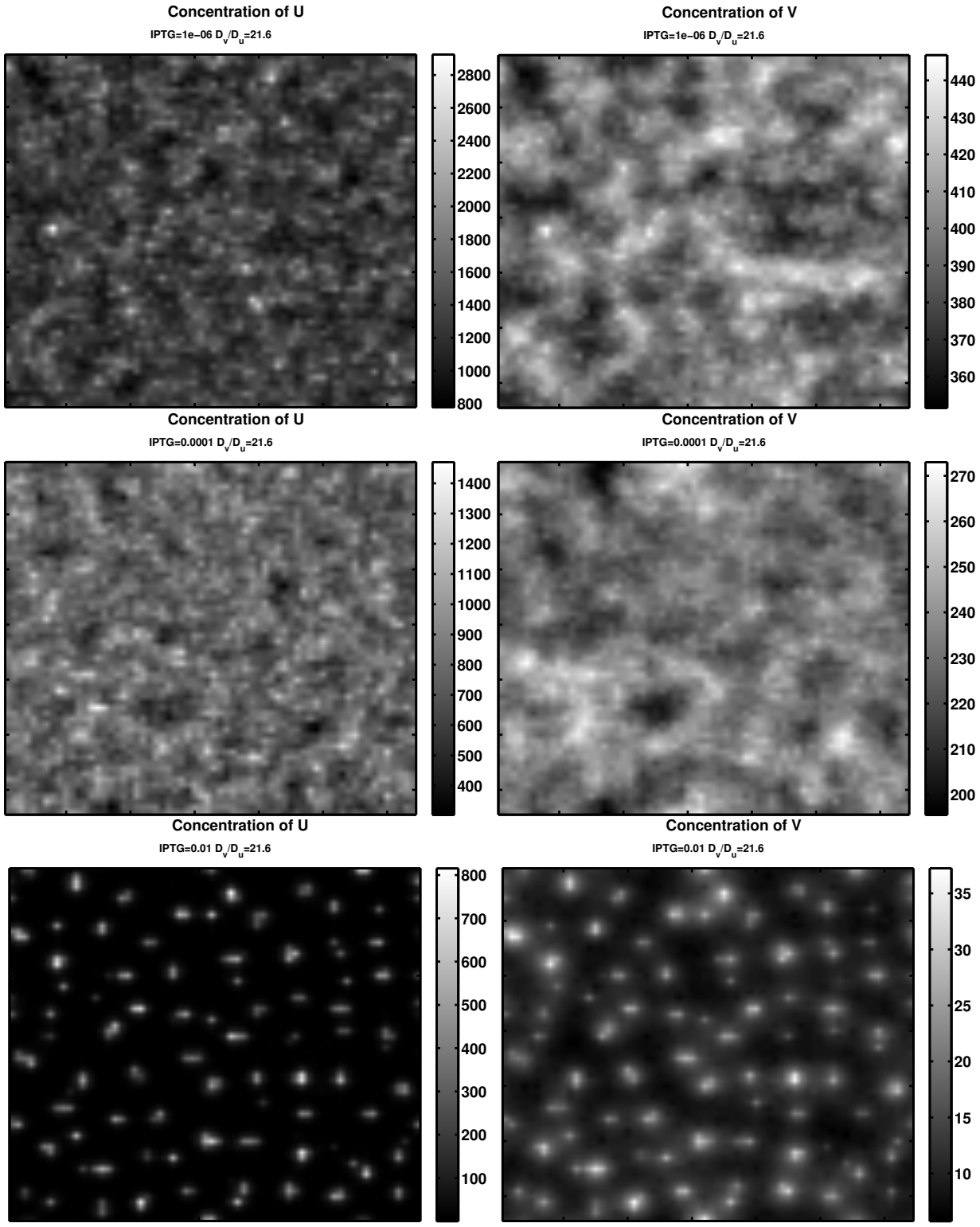


Figure 3.20: $A_{3OC12HSL}$ patterns produced in our stochastic simulation using the measured diffusion ratio of $D_v/D_u = 21.6$ and the parameters given in Tables 3.2-3.3 for three different concentrations of IPTG.

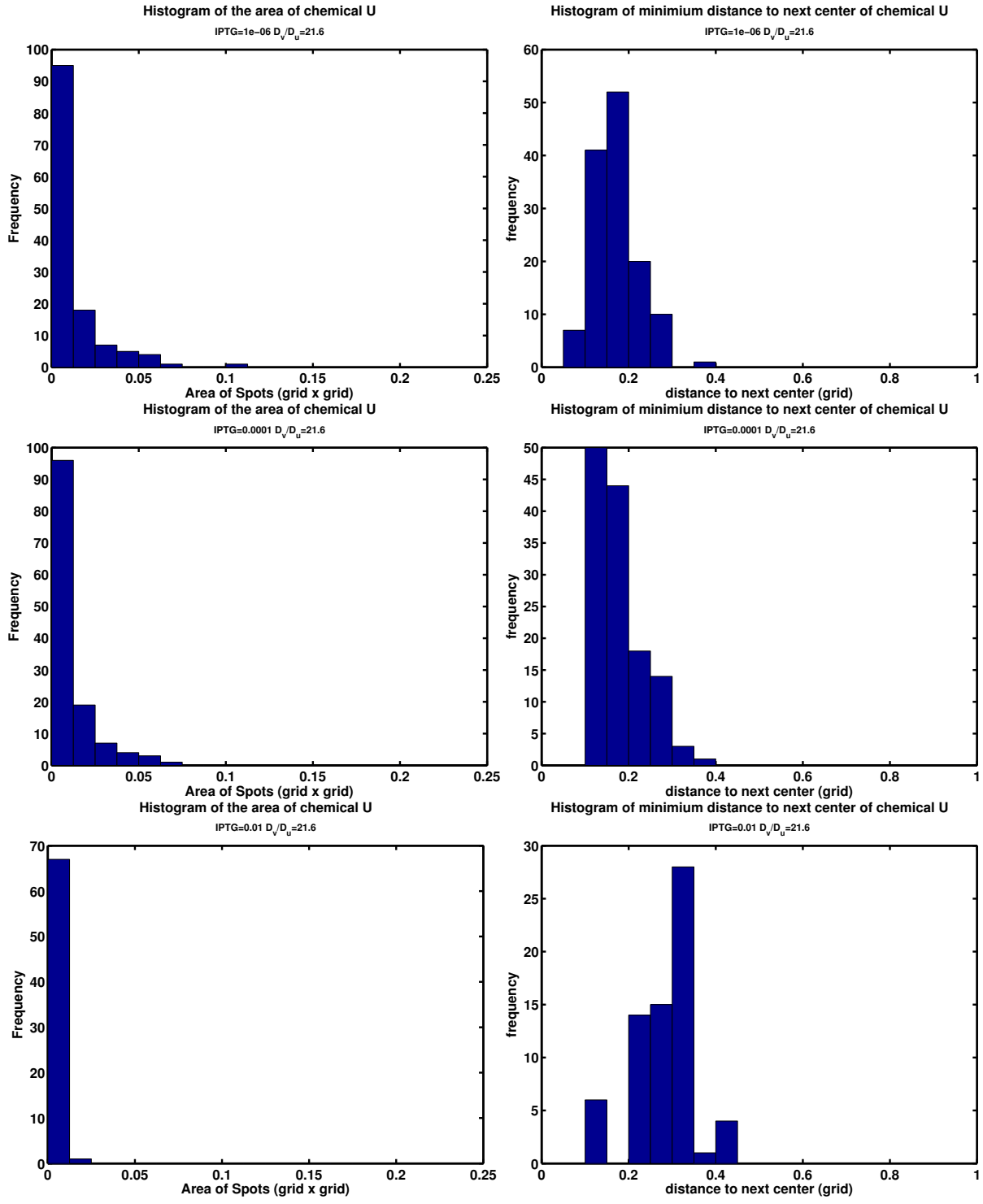


Figure 3.21: Spot size and spacing distributions for 30C12HSL produced in our stochastic simulation using $D_v/D_u = 21.6$ for three different values of IPTG.

# Part II

# Transposable Elements

# Chapter 4

# Background Chapter on Transposons and Evolution: An Introduction to DNA Transposons and Retrotransposons

The modern understanding of evolution includes three main elements: phenotypic and genomic variation, selection acting on this variation, and inheritance. Some of the key questions are: How does this variation occur? Is the variation completely random? Is the variation impacted by environmental factors? Luria and Delbruck [62] attempted to answer these questions in 1943. Their experiment was designed to test two hypotheses: (1) mutations result as a response to stress and are then passed on to future generations, or (2) mutations are pre-existing and then selected by the environment.

In the original experiment multiple clonal colonies were grown and then exposed to a phage. By counting the number of surviving bacteria it is possible to distinguish between the two hypotheses. If the mutations arose due to the stressor, then roughly the same number of bacteria should survive in each replicate. If the mutations occurred randomly in previous generations then the number of surviving bacteria depends strongly on when the mutation occurred in the past because the number of cells with a mutation grows exponentially with time after the mutation. Thus, if the distribution of bacteria that survived was heavy tailed (and followed a distribution first estimated by Luria-Delbruck), it would be good evidence for selection on pre-existing mutations [62]. The outcome of this experiment in 1943 is responsible for the widely accepted understanding that most mutations are pre-existing and are subsequently selected for by the environment.

Recently, there has been evidence for a more nuanced picture of when mutations occur. There have been reports of stress induced mutagenesis in bacteria and plants [63]. Additionally, it is known that the DNA repair mechanisms can be down-regulated due to stress on the cell [64, 65]. In addition, Luria and Delbruck's experimental data were recently reanalyzed by Nemenman's group and found to not preclude a mixed hypothesis [66]. The original experiment may not have been able to distinguish some of these effects because the stressor used was binary and very strong. The phage either kills the bacteria or the bacteria survive.

In the following chapters we examine one way in which variation arises in the genome. There are three main sources of genomic variation including point mutations, transposition, and horizontal gene transfer. Point mutations are where a single base pair can switch its identity, transposition occurs when genetic material is moved to another location within the same genome, and horizontal gene transfer is the movement of genetic material between genomes.

In this part of the thesis, we will be examining transposition. We will examine how DNA transposons and retrotransposons behave in real time in living cells. Specifically, we will show that DNA transposons have transposition rates that depend on the growth state and these transposition rates are heritable. This contributes to our understanding of the randomness of variation and how it occurs.

Transposons are colloquially known as "jumping genes." There are two main types of transposons, DNA transposons which use a "cut and paste mechanism" of transposition, and retrotransposons which utilize a "copy and paste" mechanism.

DNA transposons encode a protein called transposase which is responsible for their excision. One example of a DNA transposon is IS608. This transposable element is flanked by two palindromic imperfect repeats. While in single stranded form these form hair pin structures that serve as recognition sites for the transposase protein. The transposase that this element codes for is called TnpA. TnpA forms a homodimer that binds to the recognition sites, excises the transposon, and rejoins the single stranded DNA. The element then can reintegrate at another location [7].

Retrotransposons use a RNA intermediary to accomplish their copy and paste mechanism. They often encode their own reverse transcriptase. Retrotranposons are first transcribed from DNA to RNA. Once in RNA form, the RNA is reverse transcribed and integrated into a new location. One example of a retrotransposon is LINE1, which stands for "Long interspersed nuclear elements." Human-Line1 makes up 17% of the human genome and retroelements in general make up 45% of the human genome [67, 68]. Interestingly, the abundance of retroelements in bacteria is very small compared to that found in Eukaryotes. This will be discussed in chapter 6, where we develop a theory for their abundance based on measurements of their lethality in live cells.

# Chapter 5

# Watching Mutations and Evolutionary Dynamics in Real Time

## 5.1 DNA Transposons - Real Time Transposable Element Activity in Individual Live Cells

This chapter reports on experimental and theoretical work undertaken in collaboration with Thomas E. Kuhlman and members of his group, namely, Neil H. Kim, Gloria Lee and Nicholas Sherer. The experimental work is described for completeness and is the work of my collaborators. I, however, was deeply involved with the experimental data analysis and direction. I developed the software used to extract the excision events from the raw microscope pictures. I measured the spatial correlation function of these events and also developed a simulation for comparison. I also plotted the rate distribution of events per colony that lead us to the hypothesis of it being a Luria Delbruck-like process. The theoretical and computational analyses are primarily my work. This chapter is a modified version of our publication "Real Time Transposable Element Activity in Individual Live Cells" [15].

## 5.2 Introduction

A transposable element (TE) is a mobile genetic element that propagates within its host genome by self-catalyzed copying or excision followed by genomic reintegration [69]. TEs exist in all domains of life, and the activity of TEs necessarily generates mutations in the host genome. Consequently, TEs are major contributors to disease [70, 71, 72, 73, 74, 75, 76], development [77, 78], and evolution [6, 79]. They are also utilized as molecular tools in synthetic biology and bioengineering [80].

Despite their ubiquity and importance, surprisingly little is known about the behavior and dynamics of TE activity in living cells. TE propagation rates can be inferred from comparative phylogenetic analyses of related organisms [9, 10, 81, 82, 83, 84, 85] or endpoint analyses of TE abundance within populations [6, 7, 86, 87]. By making assumptions about the mechanisms of TE proliferation, models can be constructed to describe the distribution of TEs within genomes over evolutionary time scales, and sequenced genomes can be analyzed and fit to TE proliferation models to infer phylogeny of TE copies and estimate their rates of propagation [8]. However, most sequencing

techniques require bulk sampling of cells to provide genetic material, and sequencing is therefore generally an average over many cells. As a result, without extremely deep or single-cell sequencing techniques, most current methods are sufficient to detect only those TE events that have occurred in the germline and therefore appear in every somatic cell in the body [88].

TE rates can also be estimated by measuring relative abundances in populations that have been allowed to mutate over laboratory time scales. One of the first examples of this approach was that by Paquin and Williamson to study the effects of temperature on the rate of integration of Ty retrotransposons in *Saccharomyces cerevisiae* after growth for 6–8 generations, resulting in yeast resistant to the antibiotic antimycin A [87]; they estimated a rate of transposition of $10^{-7}$ – $10^{-10}$ insertions into a particular region of the yeast genome per cell per generation. As another example, sequencing of *Escherichia coli* (*E. coli*) at intervals in Lenski's long-term evolution experiments also provided a means to estimate transposition frequency, which they estimate to be on the order of $10^{-6}$ per cell per hour [6]. However, such measurements yield information on only the relative abundance of extant TE-affected cells in the population, and dynamic rates must again be inferred through models of population growth that may or may not be accurate.

The limitations described above mean that there is a dearth of information regarding TE behavior in individual living cells *in vivo* and the effects of TE activity on those cells. Additionally, estimation of transposition frequency from either phylogenetic comparisons or population endpoint analyses both suffer from the same serious and fundamental limitation: they are only able to detect those events that have not gone to extinction in the population, and therefore these methods almost certainly underestimate the actual rates of transposition. An analogous situation previously existed in the case of the dynamics of horizontal gene transfer: phylogenetically inferred rates of horizontal gene transfer are typically 1 per 100,000 years, whereas direct visual observation in experiments [89] has shown that the actual transfer rate is many orders of magnitude faster, about 1 per generation time.

To quantitatively study the dynamics of TE activity and its controlling factors in real time and in individual cells, we have constructed a TE system based on the bacterial TE IS*608* in *Escherichia coli.* IS*608* is a representative of the IS*200*/IS*605* family of transposable elements, which all transpose through similar mechanisms. The IS*200*/IS*605* family is widely distributed, with 153 distinct members spread over 45 genera and 61 species of eubacteria and archaea [90]. Transposition occurs by exact excision from a single DNA strand [7, 91, 92, 93, 94, 95]. Imperfect palindromic sequences flanking the ends of the TE form unique structures that are recognized by transposase protein, TnpA, which can act as a homodimer to excise the TE. The excised TE-TnpA complex can locate and integrate the TE adjacent to a short, specific sequence (TTAC). Our construct exploits the structure and regulation of the TE to allow the direct detection and quantification of TE activity in live cells using a suite of novel fluorescent reporters.

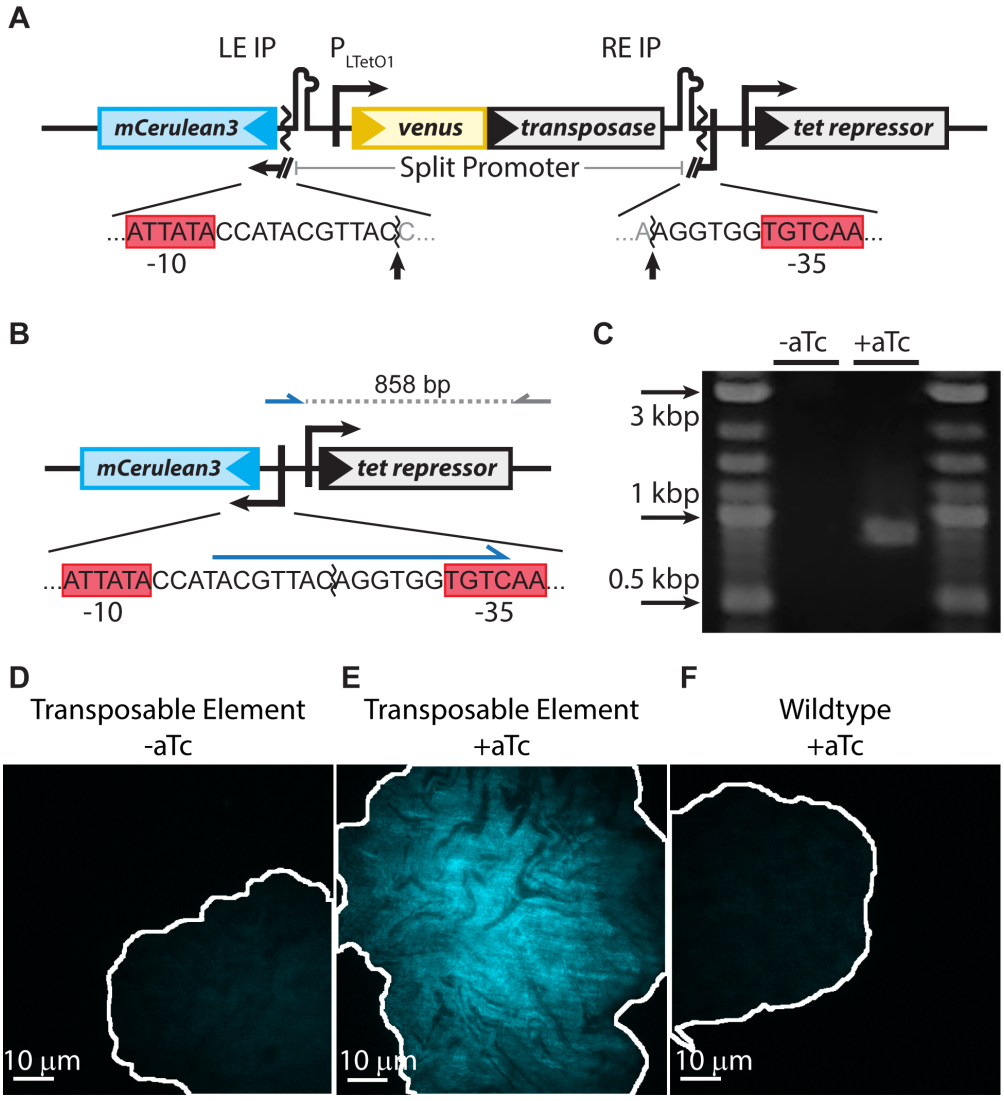


Figure 5.1: **Design and validation of the TE system.** (A) The promoter for mCerulean3 is interrupted by the transposable element, the ends of which are demarcated by left end and right end imperfect palindromic sequences (LE IP and RE IP). The transposase, tnpA (gray), is expressed from the promoter $P_{LtetO1}$, which is inducible with anhydrotetracycline (aTc). The sequences of the Promoter/TE junction and -10 and -35 sequences (red boxes) are shown below the diagram, and the sites cleaved by transposase are indicated by arrows. (B) Upon excision, the promoter for mCerulean3 is reconstituted and the cell fluoresces blue. The sequence of the reconstituted promoter is shown below the diagram. A primer designed to bind to the unique sequence formed after promoter reconstitution (blue arrow) was used to verify excision by PCR, generating an 858 bp amplicon. (C) PCR amplification using these primers only generates the 858 bp product upon induction, thus verifying excision. (D-F) Colony morphology after growth on agarose pads. Uninduced TE-carrying cells (D) and wild-type cells exposed to 20 ng/ml aTc (F) show homogeneous, low blue autofluorescence. Conversely, TE-carrying cells induced with 20 ng/ml aTc (E) show bright, inhomogeneous blue fluorescence. The brightness scale for all three images is identical. The borders of the colonies are outlined in white.

## 5.3 TE Observation System

A diagram illustrating the TE system is shown in Fig. 5.1A. The TE is composed of the transposase coding sequence, *tnpA*, flanked by a left end imperfect palindromic sequence (LE IP) and right end imperfect palindromic sequence (RE IP), which are the recognition and cleavage sites for TnpA. *tnpA* is expressed using the promoter $P_{LTetO1}$, which is repressed by tet repressor. $P_{LTetO1}$ is derived from the *E. coli* transposable element Tn*10* and titratable over a ∼100x range with anhydrotetracycline (aTc) [96]. The use of this inducible promoter allows for simple and precise control of TnpA levels within individual cells. The TE splits the -10 and -35 sequences of a strong constitutive $P_{lacIQ1}$ promoter [97] for the expression of the blue reporter mCerulean3 [98]. As shown in Fig. 5.1B, when transposase production is induced, the TE can be excised, leading to reconstitution of the promoter. The resulting cell expresses mCerulean and fluoresces blue, indicating that an excision event has occurred. The N-terminus of TnpA is translationally fused to the bright yellow reporter Venus [99], and the cells constitutively express the red reporter mCherry [100] to aid in image segmentation. Measurements of blue, yellow, and red fluorescence of controls demonstrate no crosstalk in our optical setup. The TE is hosted in the low copy number plasmid pJK14 with a pSC101 replication origin [101].

### 5.3.1 Verification of TE Observation System

Our experimental collaborators first confirmed that the TE excises upon induction of transposase production. PCR was performed using primers that bind to the unique sequence formed upon excision, and cells containing the TE and induced with aTc yielded product with amplicons of the expected length (Fig. 5.1C). Our collaborators next verified that transposase induction results in expected patterns of fluorescence corresponding to TE excision. When TE-carrying *E. coli* are grown on agarose pads with aTc, the resulting microcolonies exhibit spatially distinct bright and dark regions of blue fluorescence (Fig. 5.1E). This is expected from plasmids expressing blue fluorescent proteins after some have undergone TE excision, followed by plasmid inheritance by daughter cells. This will be discussed in more detail below. Conversely, microcolonies arising from an identically treated wild-type negative control strain carrying no plasmids and uninduced TE strains are fluorescently dim and homogeneous (Fig. 5.1 D and F).

## 5.4 Quantification of Excision Response to Transposase Concentration

Our experimental collaborators constructed two versions of the TE, one with the imperfect palindromic sequence encoded in the leading strand (ISLEAD) and the other with the imperfect palindromic sequence encoded in the lagging strand (ISLAG). Cells carrying these two constructs were grown and titrated with aTc concentrations ranging from 0-1000ng/ml. The cells were imaged after 12-13 doublings. By measuring the yellow fluorescence and the blue fluorescence for individual cells they were able to construct a TE response function of excision to transposase level

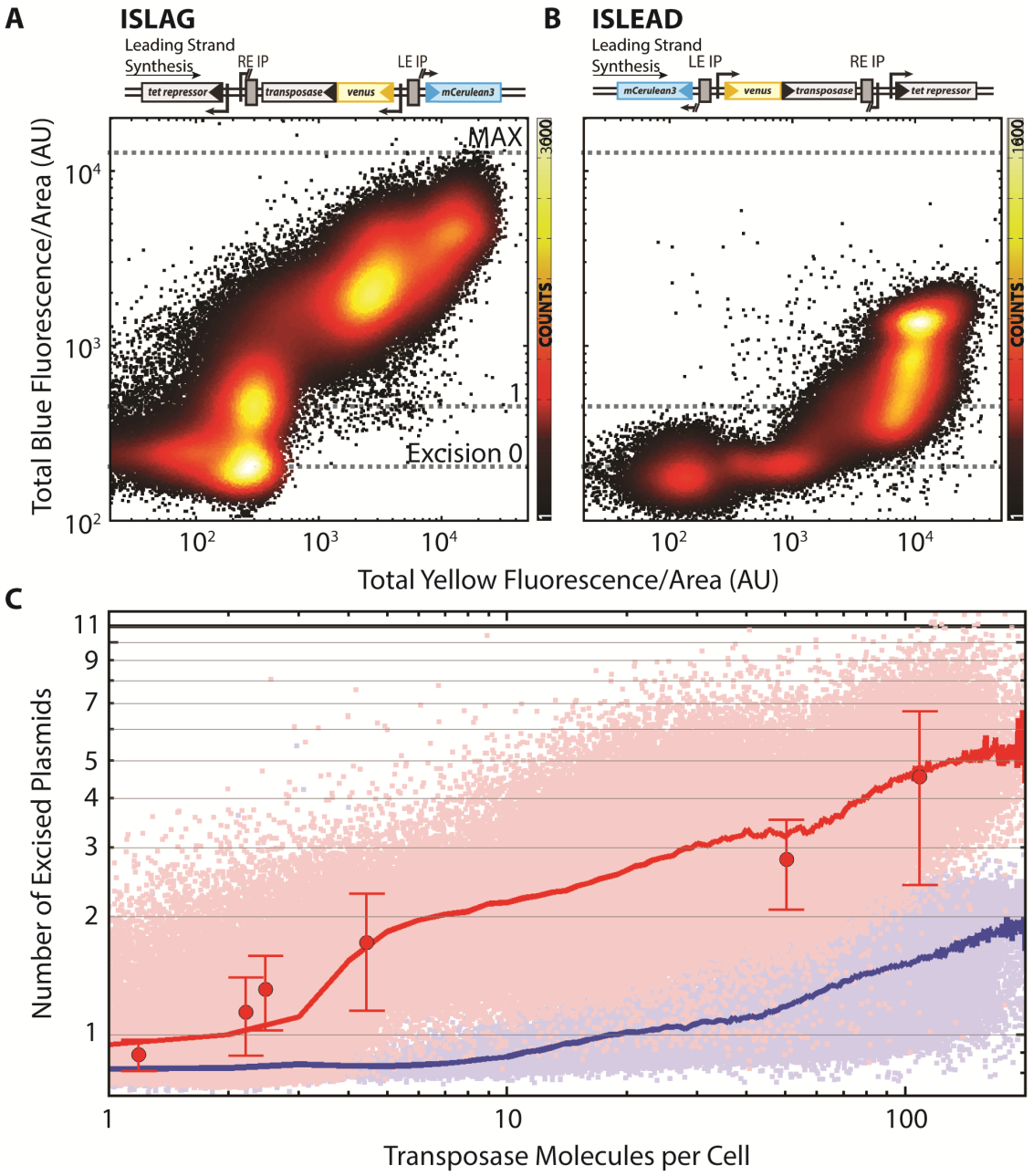


Figure 5.2: **TE excision response function.** Scatter plots of blue versus yellow total cellular fluorescence divided by cell area for TE encoded in the (A) lagging ($N_{cells}$ = 192,965) and (B) leading ($N_{cells}$ = 101,709) strand of the host plasmid. Colors indicate number of counts in each bin of a 500 x 500 grid covering the data. (C) The same data as in (A) and (B) with absolute axes. The y-axis is expressed in terms of the absolute number of excised plasmids and the x-axis is scaled to absolute number of transposase molecules per cell. Light red and blue points are lagging and leading strand data from (A) and (B) respectively. Red and blue lines are excised plasmid number averaged according to transposase molecules binned as integer quantities. Large red points indicate the number of excised plasmids as measured by qPCR; error bars are the standard error of the mean of three experimental replicates.

*in vivo.* Figure 5.2 shows scatter plots of blue fluorescence versus yellow fluorescence for individual cells carrying ISLAG (Fig. 5.2A) or ISLEAD (Fig. 5.2B). The response functions for ISLAG and ISLEAD are qualitatively different, with the ISLAG construct responding more quickly at low aTc concentrations and the ISLEAD construct responding at higher aTc concentrations. They verified that that the Venus fusion to the transposase did not affect transposase function.

They then calibrated the fluorescence intensity of the yellow and blue reporters to the numbers of transposase molecules and the number of excisions, respectively. This result is shown in Fig. 5.2C, where the response functions for ISLAG and ISLEAD are shown as the red and blue lines. The calibration for the number of transposase molecules was accomplished by measuring the bleaching kinetics of Venus-TnpA using a theoretical technique developed by Nayk and Rutenberg[102]. To calibrate the blue fluorescence intensity to excision numbers, they measured the fluorescence of a wild-type negative control with no plasmids, and a control in which every plasmid expresses mCerulean. These two reference intensities are show by the lines labeled "Excision 0" and "MAX" in Fig. 5.2. Finally, using qPCR they were able to determine the average number of plasmids per cell.

## 5.5 Observing Real Time Kinetics

By growing TE-carrying cells on agarose pads including aTc under the microscope, TE excision events can be detected in real time and their rates and statistics determined through direct observation. We find that TE activity changes as cells undergo different phases of growth, and that TE activity correlates to where cells are located within a colony. Images consisting of 40 – 80 adjacent fields of view of the TE-carrying cells were taken every 20 minutes in each of three fluorescent channels: mCherry, Venus, and mCerulean3.

I performed image analysis using custom image segmentation and analysis algorithms that I had implemented in MATLAB (MathWorks) (see Fig. 5.3). Images of the same field of view at different time points in the mCerulean channel are aligned using an image registration subroutine from Matlab's computer vision toolbox. Subsequently, the difference between images three frames apart is calculated. Only changes to the brightness show up on these difference pictures. Specifically, if an excision event occurrs, the difference picture will show a bright spot at the location of the cell that has become brighter due to an excision event. Cells that have not had an excision event will remain roughly the same brightness and will not show up on the difference image. The difference image is then thresholded to determine the exact location of the cell undergoing excision. Since cells take multiple frames to brighten and later photobleach, multiple detections of the same event will occur in consecutive processed images. It is thus important to associate only one event with all these detections. Since the cells are not moving it is sufficient to record the event location. Any subsequent detections at that location are counted as part of the original event. As long as detections are within a cell

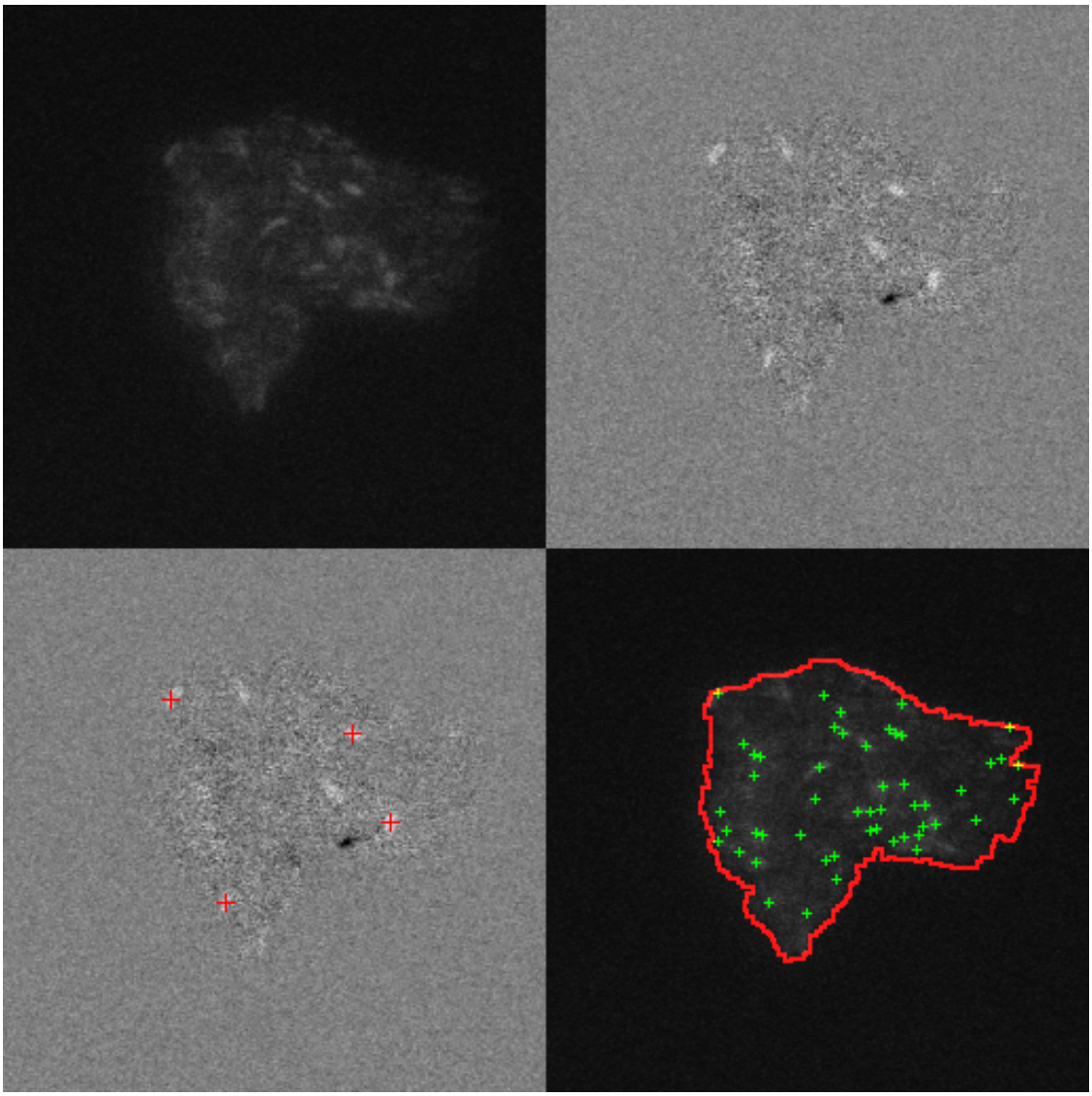

Figure 5.3: **Automatic detection of events in Matlab.** Upper left is the experimental image; upper right is the difference between the current frame and three frames previous; lower left is currently identified fashing cells; lower right displays the history of the location of prevoius events.

radius (3px) of the original detection they are counted as part of that event. Additionally, the boundaries of the colony are detected by using a different threshold on mCerulean channel. This is later used when calculating event rates for single colonies and the event densities as a function of distance from the edge of the colony.

My program recorded a time series of the intensities of the mCerulean and Venus channels at the location where events were detected by finding the average intensity for pixels involved in the event. Each individual time series is aligned by the peak of its intensity in the mCerulean channel. These time series are then averaged to produce the profile of an average event (see Fig 5.4D).

### 5.5.1 Excision Rates Depend on Growth State of Cells

At high inducer concentrations ($>$ 10 ng/ml aTc, e.g., Fig. 5.1E), a large fraction of cells immediately experiences TE events and fluoresces blue. At low inducer and transposase concentrations ($<$ 10 ng/ml aTc), we can observe individual excision events as bright flashes of blue fluorescence whose rate depends upon the growth state of the cells. As cells initially adapt to the pad, some fraction rapidly fluoresce blue, indicating TE excision. Once cells enter exponential growth, the frequency of cells becoming fluorescent drops to nearly zero; the fluorescence patterns observed in mature microcolonies at low inducer concentrations (Fig. 5.4A) arise primarily from inheritance of the initial excision events. However, upon entering final growth arrest, some cells begin to emit bright blue fluorescence (Fig. 5.4 A-C) accompanied by an increase in yellow fluorescence (Fig. 5.4D). Note in Fig. 5.4D that the excision event (blue line) is preceded by a weak increase in transposase levels (yellow fluorescence), indicating transposase-induced excision. Control strains, including a wild-type TE-less strain exposed to aTc, TE-carrying cells not exposed to aTc, and cells constitutively expressing mCerulean3, do not show similar bursts of fluorescence.

### 5.5.2 Excision Event Rate is Constant Once Initiated

Automated identification of TE fluorescence events within each colony reveals that events begin occurring with the onset of growth arrest and continue at a rate that remains approximately constant for $>$ 35 hours (Fig. 5.4E). The average event rate for this experiment, consisting of 12 colonies and $\sim$5,000 cells, was $6.3 \pm 2.6 \times 10^{-3}$ events/cell/hr. The temporal statistics are consistent with events, once initiated upon growth arrest, occurring randomly in time as described by Poisson statistics (Fig. 5.4F).

### 5.5.3 Excision Events are Spatially Correlated

Events are not uniformly random in space and are instead spatially clustered and dependent upon the location in the colony. Events are less common within $\sim$3 μm ($\sim$5 cell widths) of the colony edge compared to the center (Fig. 5.4G).

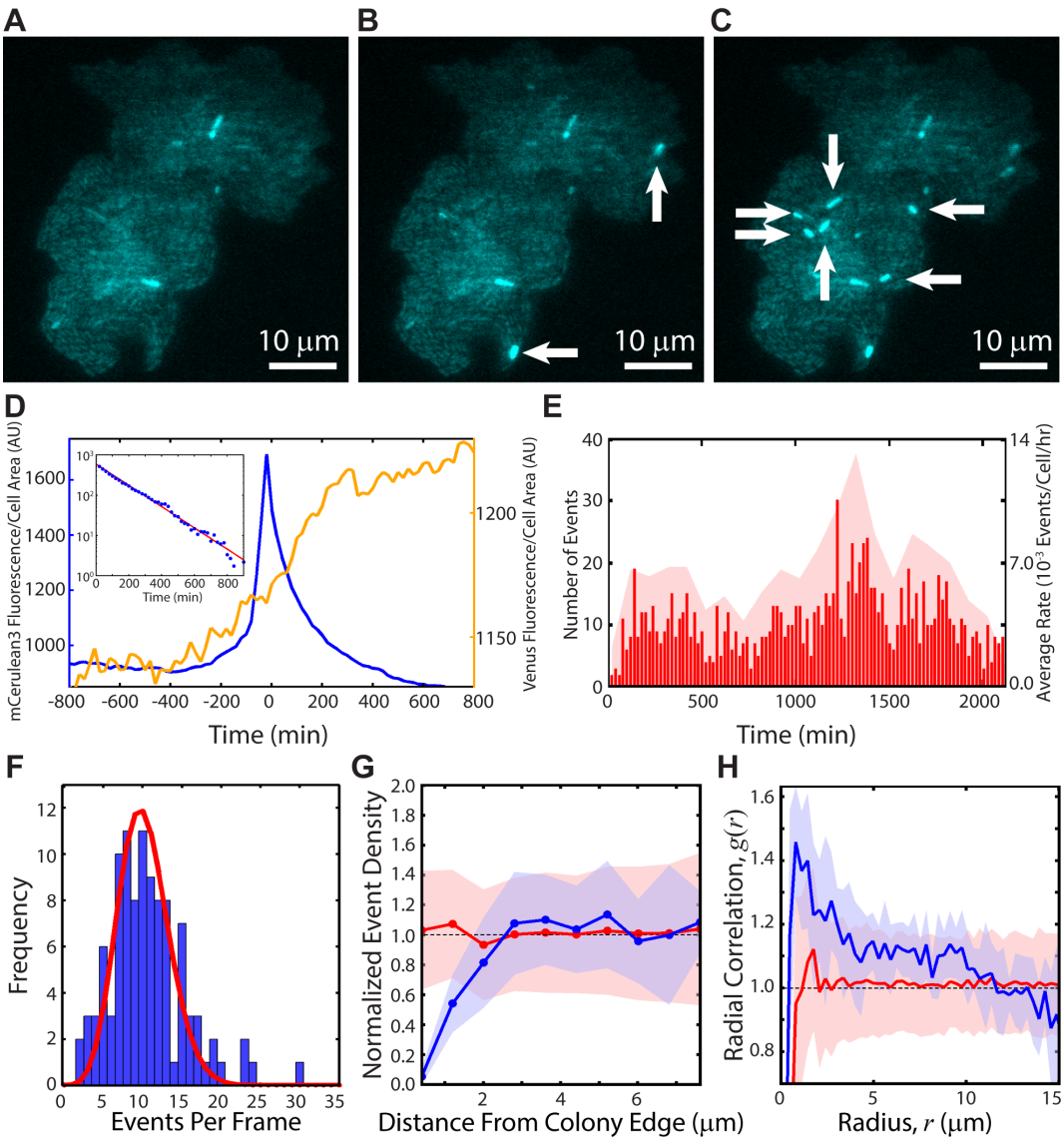


Figure 5.4: **Real time TE kinetics.** Colony induced with 5 ng/ml aTc undergoing excision events at (A) t = 0 (time of first detected events, after ∼10 hours of growth), (B) t = 40 min, and (C) t = 60 min. New events are indicated by white arrows. (D) mCerulean3 and Venus-TnpA traces for an average event. TE events were aligned with peak mCerulean3 intensity at t = 0. Shown is the mean mCerulean3 (blue, left y-axis) and Venus-TnpA (yellow, right y-axis) fluorescence/cell area as a function of time averaged over 773 events. Inset: decay of mCerulean3 fluorescence as a function of time. Red line is a fit to an exponential , with A = 589 and b = -0.006 $\text{min}^{-1}$, consistent with photobleaching. (E) Raster plot of all events in a single experiment (red lines, left y-axis) with t = 0; $N_{\text{colonies}} = 12$, $N_{\text{cells}} = 4{,}858$, $N_{\text{events}} = 1114$. The average rate was $6.3 \pm 2.6 \times 10^{-3}$ events/cell/hr. Red shaded region shows the average rates during 100 minute intervals (right y-axis). (F) Blue bars: frequency of the number of events per frame. Red line: distribution of events per frame expected from a Poisson process with an average rate of $6.3 \times 10^{-3}$ events/cell/hr. (G) Within each colony, we determine the event densities within annuli of width 0.8 μm at various distances from the colony edge. We then took an ensemble average over all colonies, where the density in each colony is normalized by the mean event density over the entire colony. Blue line: mean normalized density of events in 0.8 μm wide annuli versus the distance of the center of each annulus from the colony edge, shaded blue region is the SD. Red line: mean normalized density obtained from simulations of randomly spaced events, shaded red region is the SD. (H) Blue line: mean pair correlation function, $g(r)$, of events, shaded blue region is the SD. Red line: $g(r)$of randomly spaced events obtained from simulations, shaded red region is the SD.

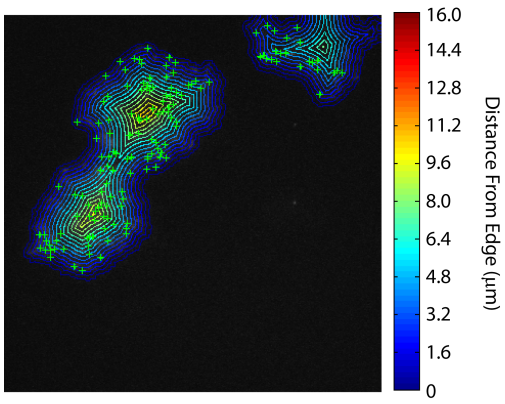


Figure 5.5: **Event density as a function of distance from colony edge.** Within each colony, I determined the number of events (green plus signs) lying within annuli of width 0.8 μm at various distances from the colony edge. The density of events in colony $i$ within an annulus at radius $r$, $\rho_i(r)$, is calculated as the number of events within that annulus, $N_i(r)$, divided by the area of the annulus, $A_i(r)$, $\rho_i(r) = N_i(r)/A_i(r)$. The data shown in Fig. 5.4G of the main text is the ensemble average over all colonies, where the density in each colony is normalized by the mean event density over the entire colony. In this image, the edges of each annulus are shown, and the color indicates the distance from the edge of the colony as given by the color bar at right.

The mean pair radial correlation, $g(r)$, also shows that events are clustered together (blue line, Fig. 5.4H; see section 5.5.4).

I performed simulations of *E. coli* growth into microcolonies combined with random distributions of TE events to determine the expected properties of $g(r)$ arising from randomly spaced TE events within an *E. coli* colony. Simulations were used to generate 200 different microcolony morphologies, each starting from a single cell and ending upon reaching a size representative of those we observe in our experiments ($\sim$300 cells with a diameter of $\sim$15 – 16 μm). After growth arrest, 15% of the cells within each colony morphology were chosen at random to undergo TE events, a rate representative of the average final number of affected cells in each colony we observe experimentally. By comparing $g(r)$ between experiment and simulation, I found that the density of events in adjacent cells in our experiment is $\sim$1.4x greater than expected compared to the simulation of events randomly distributed in space (red line, Fig. 5.4H, Fig. 5.6).

### 5.5.4 Pair correlation function, $g(r)$

The pair correlation function, $g(r)$, is a measure of the probability of finding an event (i.e., a blue fluorescent burst) at a distance $r$ away from any other event. The event density at a distance $r$ from any given reference event can be calculated as $\rho(r) = \rho g(r)$, where $\rho = N/A$ is the average event number density in an entire colony of area $A$. For a random, homogeneous distribution of ideal particles, with no hard cores, $g(r) = 1$.

To calculate $g(r)$ for a colony, each event in the colony is taken in turn as a reference particle. For each reference

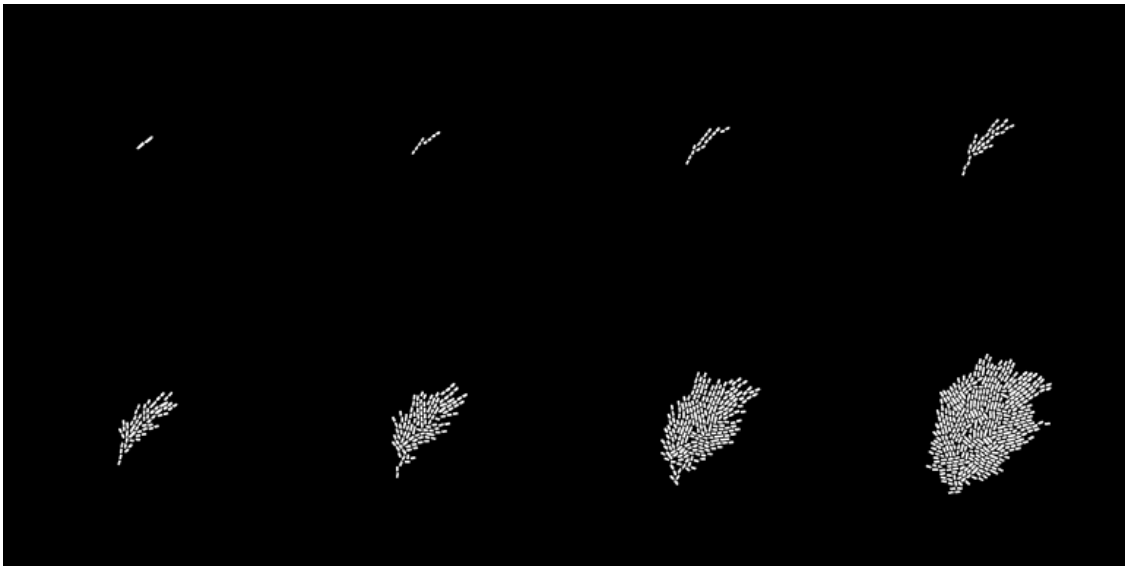

Figure 5.6: **Simulated colony growth.** Simulations of colony growth starting from a single cell. Used for calculating reference g(r) values.

particle I calculate the pair correlation function $g_i(r)$ for all $r$,

$$g_i(r) = \frac{1}{\rho} \cdot \frac{N_r}{2\pi r \cdot f_r \cdot \Delta r} \tag{5.1}$$

where $N_r$ is the number of events between two rings a distance $r$ and $r + \Delta r$ away from the reference particle. For the data presented in Fig. 5.4H, take $\Delta r = 0.32 \mu m$. To correct for edge effects, $f_r$ is the fraction of the ring area that intersects the colony. I then calculate the colony pair correlation function $g(r)$ as the ensemble average of $g_i(r)$,

$$g(r) = \langle g_i(r) \rangle . \tag{5.2}$$

### 5.5.5 Colony and $g(r)$ Simulations

Deviations of $g(r)$ from that of a random distribution of ideal particles can arise from mundane physical sources. For example, g(r) for a random distribution of hard spheres in a low-density gas shows a peak at short distances, very similar to that we observe in Fig. 5.4H purely as a consequence of entropic volume exclusion effects. However, an explicit theoretical calculation of $g(r)$ for unusually shaped particles, such as highly dense and polydisperse spherocylinders representing an *E. coli* colony, is extremely difficult. Consequently, I performed simulations of *E. coli* growth into microcolonies combined with random distributions of TE events to determine the expected properties of g(r) arising from randomly spaced TE events within an *E. coli* colony.

These simulations are a modified version of DiSCUS, an agent-based model by Goni-Moreno et al. to study horizontal gene transfer in *E. coli* [103, 104]. DiSCUS models each cell individually as a spherocylinder. The simulation is written in the python scripting language and uses the 2D physics engine pymunk as a wrapper for the physics library chipmunk (Howling Moon Software), which handles the semi-rigid body dynamics of the cells. The physics engine handles updating the forces and positions of the individual cells that arise from the environment and interactions with other cells. All the cells are non-motile but can be pushed around due to the growth of other neighboring cells. For the results discussed here, DiSCUS has been modified to remove horizontal gene transfer mechanisms and add transposable element events.

During each time-iteration the program first checks each spherocylinder to see if it is larger than a critical size. At the critical size the spherocylinder divides into two smaller spherocylinders. After the cell division step the spherocylinders are elongated. If there is too much pressure on an individual cell it will stop growing until the pressure is reduced. Finally, the physics engine resolves the forces on the cells and updates the spherocylinder positions accordingly, following standard classical mechanics.

These simulations were used to generate 200 different microcolony morphologies, each starting from a single cell and ending upon reaching a size representative of those observed in our experiments ($\sim$ 300 cells with a diameter of $\sim$15 – 16 µm). After growth arrest, 15% of the cells within each colony morphology were chosen at random to undergo TE events, a rate representative of the average final number of affected cells in each colony we observe experimentally. I calculated $g(r)$ resulting from each such random distribution of events within the 200 different colony morphologies, and repeated this process 3 times. Finally, the mean g(r) expected from a completely random distribution of TE events was calculated as the ensemble average of each such calculated $g(r)$.

To see if the cluster of events in cells was due to relatedness I performed another set of simulations using this framework. In this secondary set of simulations I allowed the cells upon cell division to have some sort of heritable change occur with some fixed probability $p_h$. Then upon growth arrest these cells with that heritable change were allowed to have an excision event with probability $p_e$. This was able to recover clustering in the $g(r)$ similar to that produced in experiment. However, I could not produce simulations of the same shape and size as we measured experimentally, which made interpretation of the simulated $g(r)$ difficult. I came up with the idea to measure distribution rates per colony and expected that if the spatial clustering was due to a heritable event that this rate would follow Luria Delbrück-like statistics and have a power law tail. Indeed upon measuring this distribution it had a power law tail (see Fig. 5.7A). Nicholas Sherer then had the idea to simulated binary trees of the same size as the colonies that were measured experimentally and was able to reproduce the same functional form as we observed.

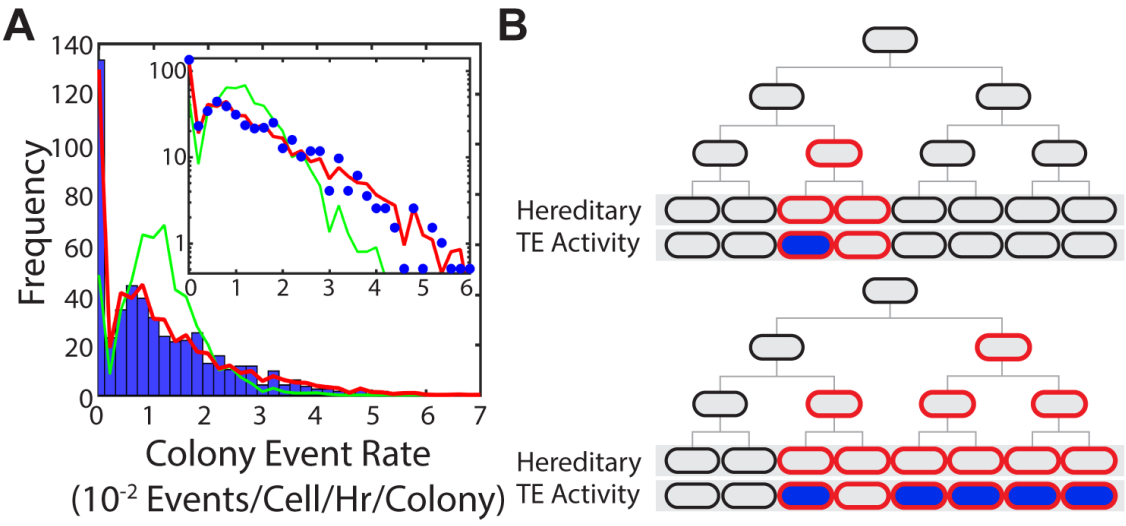


Figure 5.7: **Event Rates are Determined by a Stochastic Heritable Change.** (A) Blue bars: distribution of average event rates for individual colonies with [aTc] = 5 ng/ml; $N_{colonies}$ = 984. The mean rate of this distribution is $11.8 \pm 12 \times 10^{-3}$ events/cell/hr/colony. Red line: result of a two-step process simulated using the experimental distribution of colony sizes. Green line: result of a Poisson process with the mean rate of the experimental distribution. Inset: same data with logarithmic y-axis. (B) Cartoon picture illustrating the Luria-Delbrück process used in our simulation where some fraction of cells inherit a trait that predisposes them to TE activity (red outline), and of those cells some fraction fluoresce blue indicating TE activity (blue fill).

### 5.5.6 Distribution of Rates is Consistent with Additional Control by a Heritable Luria-Delbrück Process

As previously stated the non-uniform event distributions in space suggests that local environmental differences and/or a hereditary process are influencing TE activity. A distribution of event rates determined from 984 colonies is shown in Fig. 5.7A, with a mean rate of $11.8 \pm 12 \times 10^{-3}$ events/cell/hr/colony. This is compatible with overall number of events per frame following a Poisson distribution (Fig. 5.4F). To explain the distribution of colony event rates shown in Fig. 5.7A, we simulated a two-step process [105, 106]. First, in a Luria-Delbrück process after cells are placed on the pad, some stochastic heritable change can occur with constant probability during exponential growth that predisposes cells to TE activity (Fig. 5.7B). In the simulation, 10,000 cell colonies were simulated to grow until they reached colony sizes drawn from the colony size distribution observed in the experiment. While in growth, a heritable change occurs in a daughter cell after each division with probability $p_h$. From the affected cell, the change is inherited by all of its descendants. In growth arrest, any cell that has inherited the change can then experience a TE excision with probability pe. A good fit of event densities was found (red line, Fig. 5.7A) by searching through the two parameters with a goodness criterion (see next section). This analysis and the quality of the fit strongly suggest that the average event rate in each colony is determined by some stochastic, heritable change occurring in the lineage, for example, expression bursts, or lack thereof, of long-lived tet repressor protein.

### 5.5.7 Luria-Delbrück Modeling

To explain the data in Fig. 5.7A, we made a model of a two-step process. This model is itself stochastic and we do not attempt to solve it exactly but instead simulate the distribution it produces and compare it to the experimental distribution in Fig. 5.7A. Before we can model the excision process, it is necessary to consider the distribution of colony sizes (cells per colony). We generate a new dataset of colony sizes by sampling with replacement (bootstrapping) from the experimental colony size data.

In the first step of the model, we construct a family tree for each colony in the bootstrapped dataset. We assume each colony starts from one individual which grows and divides until it reaches the final size of the colony, and we assume that the growth rate is constant over time. Thus, the family tree for a colony is the shallowest one possible. After each division, daughter cells have a heritable change occur with probability $p_h$. All descendants of a cell with this heritable change will also have it, and two sister cells from the same mother do not affect each other. The model makes no assumptions about the biophysical nature of this heritable change.

In the experiment, the blue flashes indicating transposon excision occur after growth stops. Therefore, in the second step of the model, after growth and division, any cells with the heritable change have excisions occur with a probability $p_e$.

To find the values of the parameters of the model that best explain the experiment we simulated the model for parameter values in a small grid. The summary statistics of the standard error of the mean and standard error of the standard deviation were used to compare the mean event densities and standard deviation of event densities between theory and data; the fit was considered good if the theory was within the standard error of both the mean and the standard deviation, and acceptable if within two times the standard error for both. The results shown as the red line in Fig. 5.7A are simulations of the model with$p_h = 0.045$ per division and $p_e = 0.65$.

Notice that if $p_h = 1$, the model collapses to the simplest model you might consider, which is that all events are independent of each other in space and time and obey a Poisson process. The distribution of event rates is not that of a Poisson process for a fixed colony size, however, since colony size varies. The distribution of event rates for a Poisson process with a rate equal to the experimental rate is shown in Fig. 5.7A in green. This distribution is transparently incorrect.

One can also consider the possibility that $p_e = 1$, which would correspond to a regular Luria-Delbrück process. Although visually this fit is less accurate than the two step model (not shown), it is not obviously incorrect. Consequently, we compare the two models using the Akaike Information Criterion (AIC). We approximate the probability of the data given a model from a histogram generated by the model using 20 bins of the same size over the range of values generated. Note that since we are simulating the models, the histograms will vary slightly from simulation to simulation. 100,000 colonies were simulated to generate the histograms. Comparing the two models' fits, we

find $\Delta AIC = -80$ which corresponds to a probability of the two step model being the correct choice over a purely Luria-Delbrück process equal to $\sim 1$ (within floating point error).

## 5.6 Discussion

Our goal is to begin the quantitative understanding of how TEs fundamentally function and behave in single live cells before understanding more complex systems. Placing the TE under an inducible promoter allows us to precisely control and determine how TE excisions respond to transposase concentration. Examining the bleaching kinetics of Venus-TnpA allows us to estimate absolute numbers of transposase proteins within individual cells, which improves upon previous studies that could only infer mean TnpA levels from the applied inducer concentrations. While we use a synthetic *tet* promoter derived from an *E. coli* TE to express TnpA instead of the natural tnpA promoter, the transposase levels in any wild-type system will still sample from the same response function. That even this simple system exhibits complex dynamic behavior illustrates the necessity of using real time single-cell measurements rather than population and time-averaged estimates of TE kinetics, a parallel to the way in which real time single-molecule measurements have revolutionized our understanding of the rich dynamics hidden by population-averaged ensemble measurements [107]. This quantification of genome plasticity in real time permits the development of a precise narrative of the role of TE activity in evolution and even epidemiology.

The single-cell response curves shown in Fig. 5.2 are consistent with existing molecular models of how TnpA binds to and excises the TE from the host DNA molecule [7]. The response function displays qualitatively distinct behavior in the leading versus lagging strand. Because the lagging strand of DNA is discontinuously replicated, the lagging strand leaves single-stranded DNA exposed while synthesis of Okazaki fragments is completed. Hence, it is more energetically favorable for the folded imperfect palindromic sequences recognized by TnpA to form in the lagging strand than the leading strand, where the energetically favored state is canonically base-paired double-stranded DNA [7]. Consequently, the TE in the lagging strand is extremely sensitive to TnpA, with the first excisions occurring in the presence of only 1 – 2 TnpA dimers. Conversely, $\sim$10x higher TnpA numbers are required to initiate excision from the leading strand.

Real-time imaging allows us to track how TE activity varies from one cell to another within different colonies over time. We found that upon growth arrest, excision events are distributed non-uniformly within each colony. This non-uniformity can be described with a Luria-Delbrück process, suggesting that some stochastic, heritable trait predisposes a fraction of cells to TE activity. Additionally, the relative lack of excision activities observed near the edges of colonies may arise from local environmental variation, such as nutrient availability, between the edge and center of a colony. Together, these results demonstrate that the rate of TE excision is highly dynamic and depends

upon the amount of transposase in the cell, the TE's orientation within the genome, the growth state and life history of the host cell, and the cell's local environment.

While here we focus solely on excision, we note that since excision of a TE is required before reintegration, it is likely that integrations and the mutations they generate will occur with a rate that is dependent upon the excision rate measured here. Previous studies detecting transposition *in vivo* using time-averaged population-level methods have estimated the convolved transposition rates, i.e., the combined rates of both excision and integration, as a result of experimental or conceptual limitations in separating the two processes. Mating-out assays, for example, detect TE integration only into a conjugative plasmid which is then transferred to a virgin recipient strain for detection [108]. These methods therefore only measure the combined rate of excision, integration to the plasmid, and conjugation of the plasmid merged together. From a mechanistic standpoint, excision and integration are two separate processes that should be understood independently. It is necessary to know excision rates independently of reintegration to understand how stable transposable elements are in the genome. Furthermore, an excision itself is a mutation carrying biological significance. Any genes carried by the TE will be lost, and if the TE has silenced a gene by interrupting it, then excision may restore its function.

One of the primary results of this work is the observed heterogeneity of TE activity rates in both space and time. In a sense, this is surprising; the design of the synthetic TE employed here is extremely simple, and yet it shows complex spatial and temporal dynamics. Furthermore, since the fundamental experiments of Luria and Delbrück, the uniform randomness and homogeneity of mutation rates is frequently taken as a starting point for descriptions and models of mutation and evolution. However, as shown in Fig. 5.2, the activity of the TE is a direct function of the intracellular numbers of TnpA protein. Since it is well known that intracellular protein levels are strongly influenced by the cellular growth state [109], cell-to-cell and temporal heterogeneity in intracellular TnpA amounts and the resulting TE activity levels should perhaps be anticipated. Similar arguments can readily be made about any other mutational process that relies upon the activity of an expressed protein for its generation or repair, for example, the repair of nascent point mutations by the proteinaceous Mismatch Repair System [110].

It is difficult to draw direct and meaningful comparisons between our measurements of TE excision rates and previous measurements. Previously measured transposition rates (i.e., excision followed by reintegration) are on the order of $10^{-6}$ – $10^{-10}$ transpositions/cell/doubling [87] [or transpositions/cell/hour [6]], while the excision rates that we measure are several orders of magnitude greater. A variety of hypotheses can be proposed to reconcile these results. For example, it is possible that reintegration is extremely inefficient and only successful for a small fraction of excisions. However, we have observed that expression levels of Venus-TnpA in these and other longer time-scale measurements do not decrease over time, which suggests this is not the case (data not shown). It is also possible that previous experiments underestimate TE activity rates as a result of insufficiently deep sampling, or the deleterious

physiological effects of the TE leading to extinction of affected cells within the population. The reason for this discrepancy remains unclear and is a subject for future work.

# Chapter 6

# Characterizing Evolutionary Pressures of Retrotransposons

## 6.1 Role of Non-homologous End-joining in the Proliferation of LINE-1 Retrotransposons and Group II Introns in Bacteria

The work described in this chapter was done in collaboration with Gloria Lee, Nicholas A. Sherer, Neil H. Kim, Ema Rajic, Davneet Kaur, Niko Urriola, Chi Xue, Nigel Goldenfeld, and Thomas E. Kuhlman. The experimental work is included for completeness and was conducted by members of Thomas Kuhlman's lab. I created the model for the exponential birth defect, the stochastic models based on the Moran model to estimate extinction time of cells containing retrotransposons, and I helped develop and simulate the model for more detailed dynamics of transposon copy number in a population of cells. These models are discussed in detail in sections 6.6, 6.7.1, and 6.7.2. This chapter is a modified version of a paper that is about to be submitted to PNAS. I have modified it to concentrate on my contributions.

## 6.2 Introduction

In Eukaryotes, such as humans, retrotransposons are common. For example, retrotransposons together with introns make up $\sim 45\%$ of the human genome and constitute the majority of so-called junk DNA [67, 68]. The human retroelement LINE-1 (or L1) alone makes up $\sim 17\%$ of the genome, with $\sim 500,000$ total integrants and $\sim 80 - 100$ complete and active copies per individual [111, 112]. In contrast to retroelements in Eukaryotes, retroelements found in bacteria and archaea, known as group II introns, are rare. Group II introns are found in only $\sim 30\%$ of sequenced bacterial species and are generally present in low copy numbers of $\sim 1 - 10$ per individual [13]. The primary question of our collaboration was what limits retroelement propagation in bacteria and archaea, but allows the vast number of retrotransposons found in eukaryotes?

To answer this question our experimental collaborators created genetic constructs allowing for the controllable expression of the human retrotransposon LINE-1 and the bacterial group II intron L1.LtrB in *Escherichia coli* (*E. coli*) and *Bacillus subtilis* (*B. subtilis*). They found that the retroelements successfully integrate into both species and

that the expression of the retroelements is detrimental to growth. Surprisingly, when the human L1 was introduced to *B. subtilis* it was lethal. One of the main differences been *B. subtlis* and *E. coli* is the presence of nonhomologous end joining (NHEJ) repair of DNA double strand breaks. This suggested that the NHEJ system increased retrotransposon integration efficacy. Our experimental collaborators tested this with knockouts of NHEJ in *B. subtilis* and other experiments. Our collaborators also measured the growth rate of the bacteria at different expression levels of the retroelement and quantified the fitness cost of a retroelement in bacteria. I developed a model for the growth rate of bacteria as a function of the expression level of retroelements. This successfully explained the measured functional form of growth rate vs L1 RNA. I also used the measured fitness cost in the modeling section to determine how long it would take bacteria containing these retroelements to lose the retroelements. This was to check whether this time scale was consistent with the non-observation of retrotransposons in bacteria.

The model that I developed shows that at low copy number and at the measured fitness cost retroelements can persist. This corresponds to group II introns. The model also shows that retroelements can persist at high copy number and low fitness cost, presumably corresponding to retrotransposons in Eukaryotes. This low fitness cost is considerably smaller than what we measured in bacteria. So Eukaryotes must have other methods of reducing the fitness cost of having retroelements. One way Euykaryotes can reduce this pressure is via the spliceosome. The spliceosome removes introns. Introns are intervening sequences that disrupt the coding regions of eukaryotic genes and make up $\sim 24 - 37\%$ of the human genome [68].

There are many similarities between bacterial group II introns, the spliceosome, eukaryotic spliceosome introns, and retrotransposons. These structural and and mechanic similarities have led Belfort and others [13, 14] to hypothesize that an invasion of group II introns originating from an endosymbiotic bacterial organelle contributed to the proliferation of introns within eukaryotic genomes prior to the last eukaryotic common ancestor. They hypothesized that the spliceosome allowed the proliferation of retroelements by limiting the retroelements fitness cost impact.

The following experimental and theoretical models measure the impact that retroelements have on bacteria as a way of testing part of the proliferation of group II intron hypothesis.

## 6.3 Description of Mechanism

Human L1 and bacterial Ll.LtrB are both target-primed retroelements. L1 encodes two primary ORFs: *ORF1*, which encodes a protein (ORF1p) that binds to L1 RNA to reduce degradation, and *ORF2*, which encodes a protein (ORF2p) with endonuclease and reverse transcriptase domains [113]. Most L1H elements include a $\sim$ 100 bp DNA-encoded 3' poly(A) tract that enhances retrotransposition efficiency [114]. After transcription and translation, ORF1p and ORF2p bind *in cis* to their encoding RNA. The resulting ribonucleoprotein particle can then bind to and cut a target

DNA molecule using the ORF2p endonuclease domain. The L1 RNA 3' end hybridizes with the cut DNA, which ORF2p reverse transcriptase uses as a primer for target-primed reverse transcription (TPRT). This generates a new cDNA copy of L1 in the genome, starting from the 3' end. Reverse transcription typically aborts prior to completion, and most L1 integrations result in 5' end truncations [115]. LINE-1 contains multiple RNA splicing signals such that retrotransposition and integration into native genes leads to exonization and novel alternative splicing variants [116, 117, 118]. The function of a third ORF, *ORF0* [119], and the mechanisms driving nuclear import/export, LINE-1 truncation, second strand target-site DNA cleavage, and second strand cDNA synthesis remain poorly understood [120, 112].

Similarly, the bacterial group II intron Ll.LtrB, originally from *Lactococcus lactis*, encodes a protein, LtrA, with maturase, endonuclease, and reverse transcriptase domains. After transcription, LtrA binds *in cis* to its encoding RNA, with the maturase domain stabilizing RNA secondary structures necessary for enhancement of intron splicing from the host RNA [14]. The spliced Ll.LtrB-LtrA complex can then reintegrate into the genome with $\sim 100\%$ efficiency through retrohoming, wherein the LtrA endonuclease domain binds to and cleaves a specific sequence within the *ltrB* gene of the *L. lactis* genome and Ll.LtrB is reverse transcribed from the 3' end via TPRT [121]. Alternatively, in organisms lacking *ltrB* and the specific retrohoming target site, such as *E. coli*, Ll.LtrB can nonspecifically retrotranspose via TPRT into ectopic sites with marginal homology to the retrohoming site at low efficiencies of approximately one retrotransposition event per 109 exposed cells [121, 122, 123].

## 6.4 Description of Constructs

To construct a controllable, bacteria-expressible L1 retroelement, Thomas Kuhlman amplified from his genome a previously identified highly active L1H element [#4-35 ([111])] and modified this amplicon by PCR to employ a bacterial T7*lac* promoter [124] with a consensus Shine-Dalgarno ribosomal binding site (RBS) driving *ORF1* expression. The resulting bacteria-expressible L1 element (TL1H) was then ligated into the medium copy number plasmid pTKIP-neo [125, 126]. The strength of expression is tunable by titration with isopropyl -D-1-thiogalactopyranoside (IPTG) in bacterial strains engineered to express T7 polymerase.

TL1H exhibits characteristics that may limit its activity in bacteria. Its human codon bias may influence its expression efficiency. Additionally its *ORF2* has no bacterial RBS and TL1H has no 3' poly(A) tract. For these reasons our experimental collaborators synthesized a bacterial "optimized" L1 (GENEWIZ), based on the sequence of L1 #4-35 ([111]), they called EL1H. The synthesized EL1H is codon-optimized for *E. coli*, includes RBS sequences for both ORFs, and has the same poly(A) tract as L1 #4-35 (Fig. 6.7A, bottom). EL1H is driven by the same T7*lac* promoter and was cloned into the same pTKIP vector as TL1H. Our experimental collaborators also subcloned EL1H

into the *B. subtilis* vector pHCMC05 under the IPTG-inducible hyper-spank promoter [127].

For experiments with Ll.LtrB, our experimental collaborators employed the construct pET-TORF/RIG (Fig. 6.8A), the kind gift of the Marlene Belfort lab [128, 121]. The pET-TORF/RIG plasmid uses the same pBR322 plasmid backbone as pTKIP, and Ll.LtrB is expressed from the same T7*lac* promoter as we employed for L1 expression. Hence, expression levels of both L1 and Ll.LtrB are directly comparable between experiments. For experiments in *B. subtilis*, they subcloned TORF/RIG into the same sites of pHCMC05 as EL1H under control of the hyper-spank promoter.

## 6.5 Effects of Retroelement Expression on Growth

A brief summary of our collaborator's experiments follows. For a more detailed description see appendix 6.9.

Our experimental collaborators transformed the constructs described above into *E. coli*. They found a decrease in growth rate in response to L1 expression as the cultures were titrated with IPTG. This was measured by taking periodic measurements of the optical density in a variety of growth media. Furthermore, they conducted calibration experiments to find out how many L1 RNA are produced for a given dose of IPTG. Combining the results from both these experiments they were able to produce graphs showing the normalized growth rate of *E. coli* as a function of L1 RNAs per cell (see figure 6.1A). By fitting an exponential function to this graph they found that, on average, each L1 transcript yields a decrease in *E. coli's* growth rate of $\sim 0.83 \pm 0.06\%$ (TL1H) or $1.9 \pm 0.6\%$ (EL1H).

Our experimental collaborators conducted the same type of titration experiments in *E. coli* for the bacterial group II intron L1.LtrB, see Fig 6.1B. They measured the growth defect resulting from Ll.LtrB to be weaker than that from L1, with each Ll.LtrB transcript reducing growth rate by $0.11 \pm 0.02\%$.

We hypothesized that *B. subtlis* would be more resistant to L1 and better able to survive cleavage of DNA than *E. coli*. We thought this because *B. subtilis* is able to repair DNA double strand breaks through nonhomologous end joining (NHEJ) in a manner similar to eukaryotes. The experiments, however, showed the opposite. Wildtype *B. subtilis* cannot survive transformation with pHCMC05-EL1H. But when they knocked out the function of NHEJ and tried again they were successful. They next cloned and expressed a subset of the NHEJ repair mechanism into *E. coli* and observed the same enhanced lethality of L1. Similarly, they observed an enhanced lethality of L1.LtrB when they expressed L1.LtrB at the same time as NHEJ enzymes.

Our experimental collaborators also conducted experiments that strongly suggest that L1 successfully integrates into *E. coli's* chromosome and NHEJ enhances the retrotranposition efficiency of L1 and L1.LtrB (see appendix 6.9).

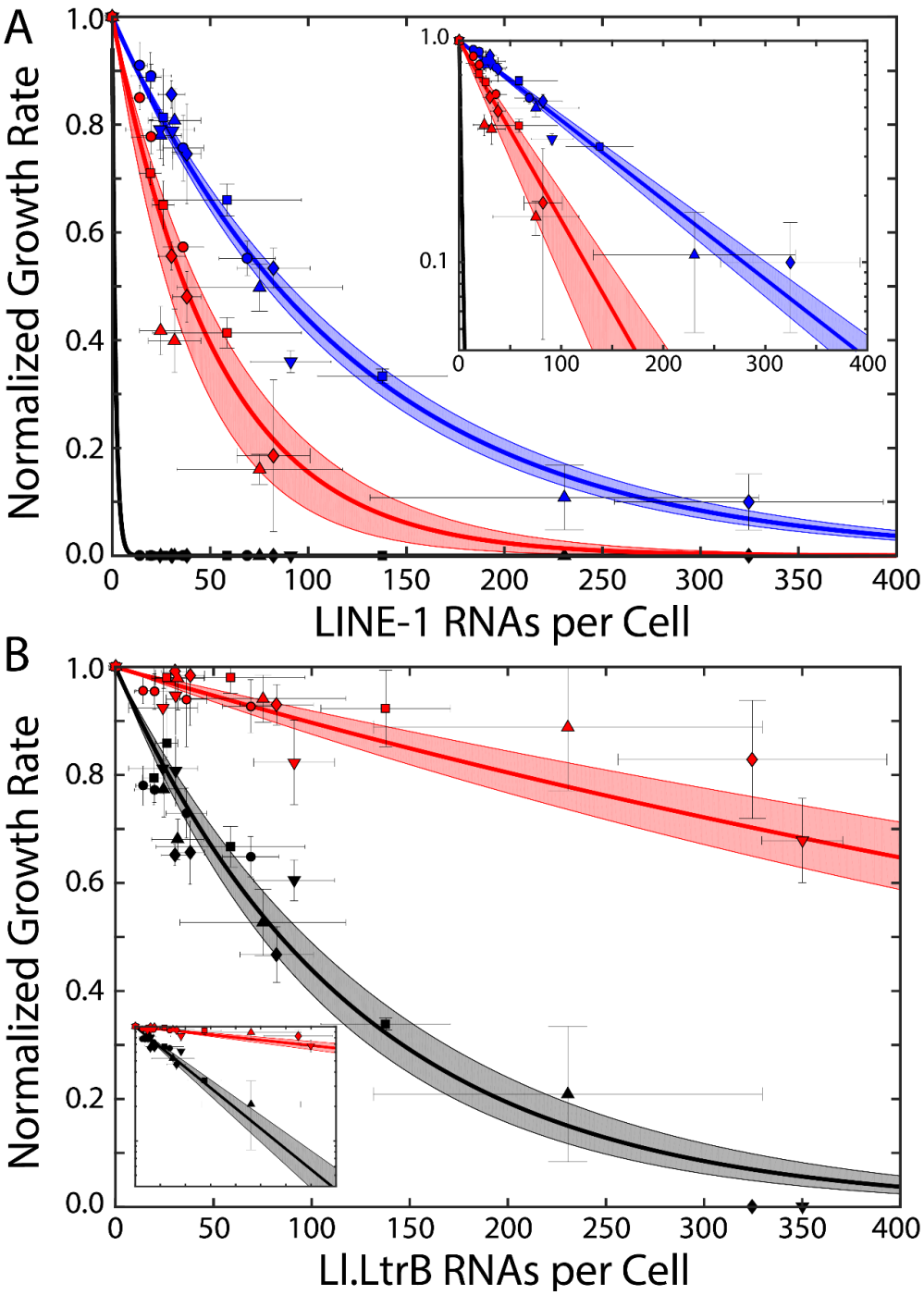


Figure 6.1: **Quantification of physiological effects of retroelement expression.** (A) Normalized growth rate as a function of LINE-1 expression on *E. coli* growth in a variety of media. •: RDM glucose; ■: RDM glycerol; ♢: cAA glucose; ▲: M63 glucose; ▼: M63 glycerol. Blue points: TL1H; red points: EL1H; black points: EL1H and TL1H + NHEJ. Each point corresponds to the mean of three growth and four qRT-PCR measurements; error bars: SEM. Solid lines: fits to $\exp[-b \cdot L]$ , yielding $b = 0.0083 \pm 0.0006$ (TL1H), $b = 0.019 \pm 0.006$ (EL1H), and $b = 0.600 \pm 0.031$ (TL1H and EL1H +NHEJ). Fit errors are 95% CI (shaded regions). Inset: same, with log y-axis. (B) Same as (A), quantifying effects of pET-TORF/RIG pZA31-tetR (red) and pET-TORF/RIG pZA31-NHEJ (black). Inset scales are identical to (A). Exponential fits yield $b = 0.0011 \pm 0.0002$ (-NHEJ), $b = 0.0082 \pm 0.0011$ (+NHEJ).

## 6.6 Modeling of Physiological Effects

I developed the following model in collaboration with Chi Xue and Nicholas Sherer. I did the final calculations.

The observed exponential decay in normalized growth rate can be explained by a simple model where we consider the effect that integrations will have by disrupting essential chromosomal genes and thus cell function. In the simplest model of this kind, we consider that there are two sub-populations of cells: those that grow normally, and those with retroelement integrations disrupting all growth. In this binary model, there are $L$ transcripts, each with a probability $w$ of integrating and disrupting growth, and the probability $q$ of a cell having no integrations affecting growth during a cell cycle given by a binomial distribution evaluated at zero:

$$q = \binom{L}{0} w^0 (1-w)^L = \exp\left[-\ln\left(\frac{1}{1-w}\right) L\right]. \tag{6.1}$$

In our growth experiments, cells are continuously growing in exponential phase. An individual cell, in the absence of integrations, will produce $g_0 dt$ new individuals in a time interval $dt$. This leads to a simple model of exponential growth of the form $\frac{dx}{dt} = g_0 x$. If we consider a binary model with a population $x$ of normal cells and a population $y$ of cells with no growth due to integrations, an individual of $x$ will still produce $g_0 dt$ new individuals but only a fraction $q$ of these will be able to grow. This leads to the population model [129]:

$$\frac{dx}{dt} = q g_0 x, \quad \frac{dy}{dt} = (1-q) g_0 x \tag{6.2}$$

The total population of cells in this model grows as $x_0 + y_0 + \frac{x_0}{y_0}[\exp(q g_0 t) - 1]$ . Thus the growth rate measured in a plate reader would be $q g_0$ and the normalized growth rate is just $q$. We fit eq. 6.1 to the form $\exp[-bL]$ and make the identification $b = -\ln[1-w]$ , which means $b \approx w$ for $w \ll 1$ . That is, $b$ is approximately equal to the probability of a retroelement transcript integrating and disrupting growth. In summary, this simple binary model recapitulates the exponential dependence of the growth rate on the number of transcripts.

More complex models of the impact of transposable element integration can be developed, with more than two sub-populations and more nuanced assumptions about the physiological effects. But we find that the dynamics of these models reduce to that of the two rate model presented above, with renormalized parameters. An example of one such model is as follows. Let the number of cells with no chromosomal integrations harming their growth be $N_0$, the population of cells with one integrant be $N_1$, and so forth. Then a set of differential equations describing the population dynamics in exponential growth with growth rate $g_0$ is

$$\frac{dN_x}{dt} = g_0 f(x)(1-\mu) N_x + g_0 f(x-1) \mu N_{x-1}, \tag{6.3}$$

where $f(x)$ is a monotonically decreasing function describing the inhibition of cell growth due to gene disruption by integrations, $\mu$ is the mutation rate, and the index $x$ runs from 0 to some integer $x_{max}$ where the number of integrants is so high the cell cannot function and dies. Making the substitution $(1-\mu) = q$ ,

$$\frac{dN_x}{dt} = g_0 f(x) q N_x + g_0 f(x-1)(1-q) N_{x-1} \tag{6.4}$$

This is a lower triangular system of equations whose eigenvalues are the diagonals. After many generations, the largest eigenvalue will dominate and correspond approximately to the measured growth rate. Since $f(x)$ is a monotonically decreasing function, this means the growth rate is $g_0 f(0) q$. $f(0) = 1$ and thus the growth rate is $q g_0$ and the normalized growth rate is $q$. This is the same result as the binary model discussed above.

## 6.7 Modeling of Retrotransposon Dynamics

### 6.7.1 Moran Model of Extinction of Transposons

The following models and calculations were all developed and performed by me.

We calculated the extinction time for cells containing retrotransposons to determine if the timescale for extinction was short enough to explain the limited number of retrotransposable elements and types found in bacteria. We calculated the extinction time for cells containing retrotransposons using the fitness cost we measured in experiment. We modeled two distinct possibilities. In the first possibility, we modeled how long it would take for a single cell without a retrotransposon to become fixed in a population of cells that initially have retrotransposons. This would correspond to a situation where there is a direct competition between cells. The second situation we modeled was how long it would take the retrotransposon to go extinct if we started with a population of cells all containing retrotransposons. In this situation we model how long it takes for a random mutation to knock out the function of the retrotransposon and then become fixed in the population.

For the first situation, we used a Moran model [130] with a population A of retrotransposons which grow with rate $q g_0$ and a population $B$ of cells without the retrotransposon which grow with a rate $g_0$. Where $q$ is the normalized growth rate measured in our experiments. The Moran model requires that the population size, $N$, remain fixed. To meet this requirement for every birth at least one other cell must die. This process can be characterized by the following set of reactions:

$$A + B \xrightarrow{q g_0} 2A, A + B \xrightarrow{g_0} 2B \tag{6.5}$$

This set of reactions is typically written with $q g_0$ set to 1 to measure everything in terms of generations of the first species. Additionally, the second reaction rate is usually written as $1+s$ where $s$ is fitness advantage population $B$ has

over $A$. We follow this convention and in our case $s$ thus becomes $s = 1/q - 1$:

$$A + B \xrightarrow{1} 2A, A + B \xrightarrow{1+s=1/q} 2B \tag{6.6}$$

From this set of reactions we can write down the corresponding Master Equation for probability, $P(x_1, x_2)$ , of having number density $x_1$ of $A$ and number density $x_2$ of $B$. The transfer rates can be written as $T_1\left(x_1 + \frac{1}{N}, x_2 - \frac{1}{N} | x_1, x_2\right) = x_1 x_2$ and $T_2\left(x_1 - \frac{1}{N}, x_2 + \frac{1}{N} | x_1, x_2\right) = (1+s) x_1 x_2$ . We define the operators $\varepsilon^+$ and $\varepsilon^-$ so that $\varepsilon^+ f(x) = f(x + 1/N)$ and $\varepsilon^- f(x) = f(x - 1/N)$ . Using these operators we can write the Master Equation as

$$\frac{\partial P(x_1, x_2)}{\partial t} = N(\left(\varepsilon_{x_2}^+ \varepsilon_{x_1}^- - 1\right) T_1 P + \left(\varepsilon_{x_1}^+ \varepsilon_{x_2}^- - 1\right) T_2 P) \tag{6.7}$$

We can now perform a Kramers-Moyal system size expansion in $1/N$ and truncate the expansion after second order [131, 132, 133]. Notice that the operators as defined above follow the Taylor expansion: $\varepsilon_x^+ f(x) = f\left(x + \frac{1}{N}\right) = \left(1 + \frac{1}{N}\frac{\partial}{\partial x} + \frac{1}{2N^2}\frac{\partial^2}{\partial x^2}\right) f(x)$ . This will produce the following Fokker-Planck equation for the probability:

$$\frac{\partial P(x_1, x_2)}{\partial t} = (-\frac{\partial}{\partial x_1}(T_1 - T_2) - \frac{\partial}{\partial x_2}(T_2 - T_1) + \frac{1}{2N}\left(\frac{\partial}{\partial x_1} - \frac{\partial}{\partial x_2}\right)^2 (T_1 + T_2)) P \tag{6.8}$$

If we make the change of variables $p = x_2$ and $c = x_1 + x_2$ we obtain an equation only depending on $p$ since $c$ is the total number density, which is a constant and set equal to unity:

$$\frac{\partial P(p,t)}{\partial t} = (-\frac{\partial}{\partial p}(s(1-p)p) + \frac{1}{2N}\frac{\partial^2}{\partial p^2}(2+s)(1-p)p) P(p,t) \tag{6.9}$$

From this Forward Fokker-Planck equation we see that the mean rate of change in frequency of $B$ per generation is $M(p) = s(1-p)p$ and the variance is $V(p) = (2+s)(1-p)p/N$. Note that this result for the variance is different from the formula quoted by Kimura and Ohta to find mean fixation time in genic selection [134]. Our variance includes a dependence on the selection coefficient $s$ and describes a haploid population as opposed to the diploid population Kimura and Ohta modeled. Using our results for the mean and variance, we can use Kimura and Ohta's general solution[134] for the mean fixation time and probability of fixation of population $B$.

The probability of fixation $u(p)$ is given by

$$u(p) = \int_0^p G(x) dx \Big/ \int_0^1 G(x) dx \tag{6.10}$$

where

$$G(x) = \exp\left[-\int \frac{2M(x)}{V(x)} dx\right]. \tag{6.11}$$

For our system the fixation probability is

$$u(p) = \frac{1 - e^{-\frac{2sNp}{2+s}}}{1 - e^{-\frac{2sN}{2+s}}} \tag{6.12}$$

and the fixation probability of one cell without a retrotransposon is

$$p_f = u\left(\frac{1}{N}\right) = \frac{1 - e^{-\frac{2s}{2+s}}}{1 - e^{-\frac{2sN}{2+s}}} \tag{6.13}$$

The fixation time is

$$t_f(p) = \int_p^1 \Psi(x) u(x) (1 - u(x)) dx + \frac{1 - u(p)}{u(p)} \int_0^p \Psi(x) u^2(x) dx, \tag{6.14}$$

where

$$\Psi(x) = 2 \int_0^1 G(z) dz / V(x) G(x) \tag{6.15}$$

We are interested in the fixation time when only one individual initially has no retrotransposon, i.e., $p = 1/N$. We can numerically evaluate the above expression to obtain the fixation time $t_f(1/N)$ .

We can also derive the scaling behavior $t_f(1/N)$ with $N$. Since $N$ is usually large, $t_f(1/N)$ can be approximated by $t_f(0)$ , which is explicitly given by the following expression

$$t_f(0) = \frac{1}{s\left(1 - e^{-\frac{2sN}{2+s}}\right)} \int_0^1 \frac{\left(1 - e^{-\frac{2sN}{2+s}x}\right)\left(1 - e^{-\frac{2sN}{2+s}(1-x)}\right)}{x(1-x)} dx. \tag{6.16}$$

The integral in the above equation can be calculated on three intervals, with $a \equiv \frac{2sN}{2+s}$ ,

$$I_1(a, \theta) \equiv \int_0^\theta \frac{(1 - e^{-ax})\left(1 - e^{-a(1-x)}\right)}{x(1-x)} dx,$$

$$I_2(a, \theta) \equiv \int_\theta^{1-\theta} \frac{(1 - e^{-ax})\left(1 - e^{-a(1-x)}\right)}{x(1-x)} dx,$$

$$I_3(a, \theta) \equiv \int_{1-\theta}^1 \frac{(1 - e^{-ax})\left(1 - e^{-a(1-x)}\right)}{x(1-x)} dx,$$

So $t_1(0) = \frac{1}{s\left(1 - e^{-\frac{2sN}{2+s}}\right)} (I_1 + I_2 + I_3)$ .

For small $\theta$, the factor $\frac{1 - e^{-a(1-x)}}{1-x}$ in the integrand of $I_1(a, \theta)$ can be approximated as $(1 - e^{-a})$ , so that

$$I_1(a,\ \theta) \approx \left(1 - e^{-a}\right) \int_0^\theta \frac{1 - e^{-ax}}{x} dx \equiv \left(1 - e^{-a}\right) h_1(a, \theta).$$

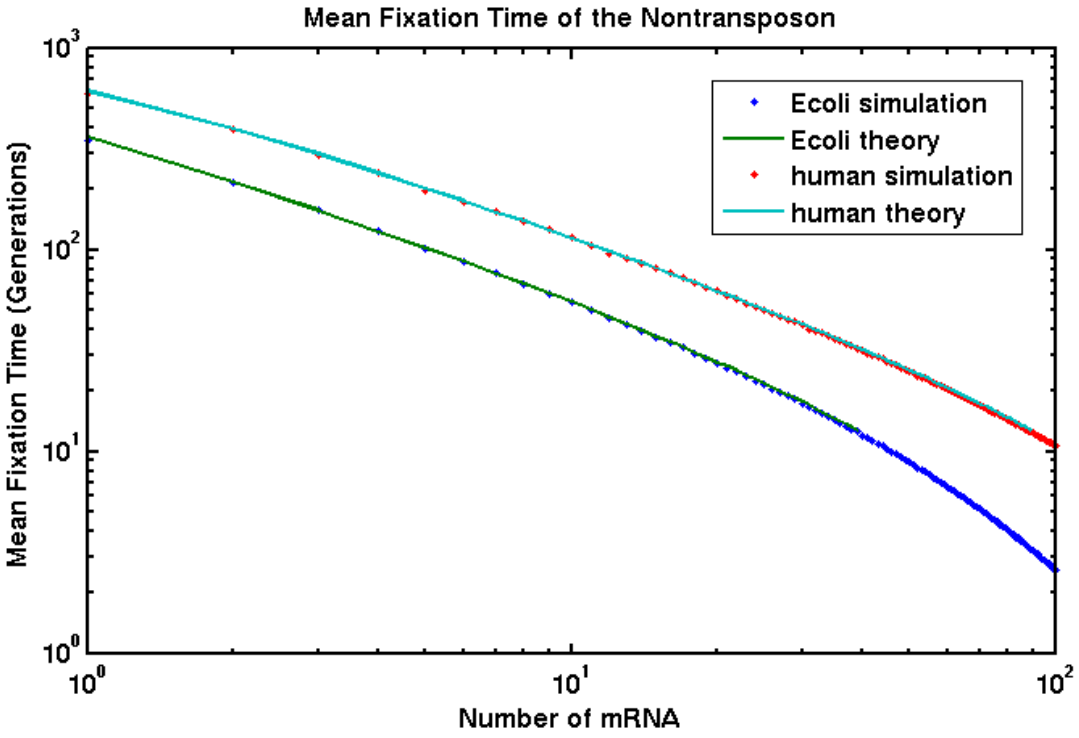


Figure 6.2: Moran model of the fixation time of a single cell without the retrotransposon in population of 999 cells with retrotransposons. The cell without the retrotransposon have selective advantage $s = \exp(0.019L) - 1$ for the *E. coli* optimized retrotransposons and $s = \exp(0.0083L) - 1$ for human retrotransposons where $L$ is number of mRNA. Blue dots correspond to Gillespie simulations of how long it takes the *E. coli* optimized retrotransposon to go extinct and the green line is the corresponding theory. Red dots correspond to Gillespie simulation of how long it takes the human L1 to go extinct in *E. coli.*

where $h_1(a,\theta) = \int_0^\theta \frac{1-e^{-ax}}{x} dx$, and $\partial_a h(a,\theta) = \int_0^\theta \partial_a \left(\frac{1-e^{-ax}}{x}\right) dx = \int_0^\theta e^{-ax} dx = \frac{1}{a}\left(1-e^{-a\theta}\right)$ . At large $N$, $a = \frac{2sN}{2+s}$ is also large and the leading term in $\partial_a h(a,\theta)$ is $\frac{1}{a}$. Therefore $h(a,\theta) \sim \ln(a)$ , and we further have $I_1(a,\theta) \sim \ln(a)$ .

Observe that $I_3(a,\theta) = I_1(a,\theta)$ , and that $I_2(a,\theta)$ does not contribute to the asymptotic leading term. We obtain the asymptotic behavior of $t_f(0)$ at large $N$, as follows,

$$t_f(0) \sim \frac{1}{s} 2\ln(a) + C = \left(\frac{2}{s}\right) \ln\left(\frac{2Ns}{2+s}\right) + C,$$

where $C$ stands for higher order terms. The fixation time scales as $2\ln(N)/s$ up to a higher order difference.

We ran a stochastic simulation of the Moran model as defined above using Gillespie's algorithm [17] . We found the results of the simulation were in excellent agreement with the analytic approximation for fixation time. We ran simulations to see how the fixation time depended on number cells and on number of mRNA. For our system, we have $s = \exp(0.019L) - 1$ for the *E. coli* optimized retrotransposon EL1H and $s = exp(0.0083L) - 1$ for the human-derived

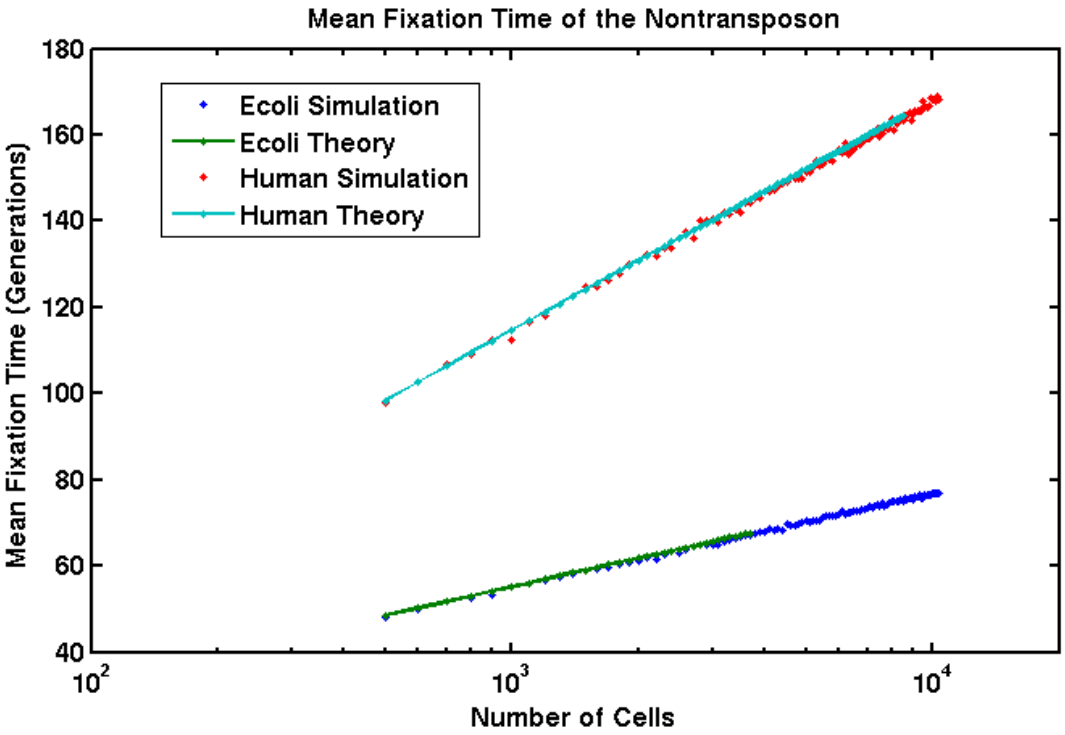


Figure 6.3: Moran model of the fixation time of a single cell without a retrotransposon in a population of $N-1$ cells with retrotransposons. Simulations and theory are calculated for cells with 10mRNA corresponding to a selective advantage $s = 0.21$ of the nontransposon over the *E. coli* optimized retrotransposons and $s = 0.087$ over the human L1 in *E. coli.* Blue dots correspond to Gillespie simulations of *E. coli* optimized retrotransposons and green dots human L1 in *E. coli*. Solid lines correspond to the analytic theory.

retrotransposon TL1H, where $L$ is number of mRNA. Figure 6.2 shows how the fixation time scales with mRNA at fixed population size of 1000. Figure 6.3 shows how the fixation time scales with population size at fixed number of mRNA. We assumed that 10 mRNA was present on average in each cell, which corresponds to what we expect if the cells are not being induced with IPTG and is the amount produced by our leaky promotor. The fixation amount does indeed scale as $\ln(N)$ and $1/s$ , agreeing with the theoretical asymptotic behavior.

To model the fixation time starting with a population of all retrotransposons, we modify our original Moran model slightly by adding one additional reaction.

$$AB \xrightarrow{1} 2A, \qquad AB \xrightarrow{1+s=1/q} 2B, \qquad A \xrightarrow{\Delta} B$$

We have added the reaction, $A \xrightarrow{\Delta} B$ to model a random mutation knocking out the function of the retrotransposon. We use an estimate of $\Delta = 10^{-6}$ /generation/cell. As before, we can find the mean and variance of the rate of change in frequency of $B$ per generation by using a system size expansion of the Master Equation. We find $M = s(1-x)x + \Delta(1-x)$ and $V = [(2+s)(1-x)x + \Delta(1-x)]/N$ . We can again use the formal solution provided by Ohta and Kimura[134] to find the fixation time.

To get a better sense of how the solution should scale with $N$ we can assume that there is a separation in time scales between the time needed to wait for a random mutation to knockout the function of the retrotransposon and the time it takes for that mutation to be fixed in the population. Define $T$ as the average time it takes to get one cell that has the retrotransposon knocked out. Then the population will have the same dynamics as the Moran model and will go to fixation with probability $p_f = u\left(\frac{1}{N}\right) = \frac{1-e^{-\frac{2s}{2+s}}}{1-e^{-\frac{2sN}{2+s}}}$ and in a time much shorter than $T$. If the cell doesn't go to fixation we will need to wait on average another period of $T$ and again have a probability $p_f$ of fixation. We can thus write down an infinite series for the average time to fixation as follows:

$$\langle t_{f0} \rangle = p_f T + p_f (1-p_f) 2T + p_f (1-p_f)^2 3T + \cdots = \sum_{n=0}^{\infty} p_f (1-p_f)^n (n+1) T = \frac{T}{p_f}$$

We can make the further assumption that the average time it takes one cell in a population of size $N$ to have the retrotransposon knocked out should scale as $T = 1/(\Delta N) + D$ where $D$ is the average time it takes to go extinct and $1/(N\Delta)$ is the average time it takes to have at least one cell knock out the retrotransposon. We can see from Figure 6.4 that when $N << 1/\Delta = 10^6$ this approximation works very well. So the average fixation time scales as $(1/(\Delta N) + D)/p_f$.

When the population size is on the same order as $1/\Delta$ or larger, then the timescale $T$ is of order unity or smaller. This means that there is no longer a separation in timescales between the time needed to wait for a mutation to knock out the function of a retrotransposon and the time it takes for that mutation to go to fixation. In this case we can

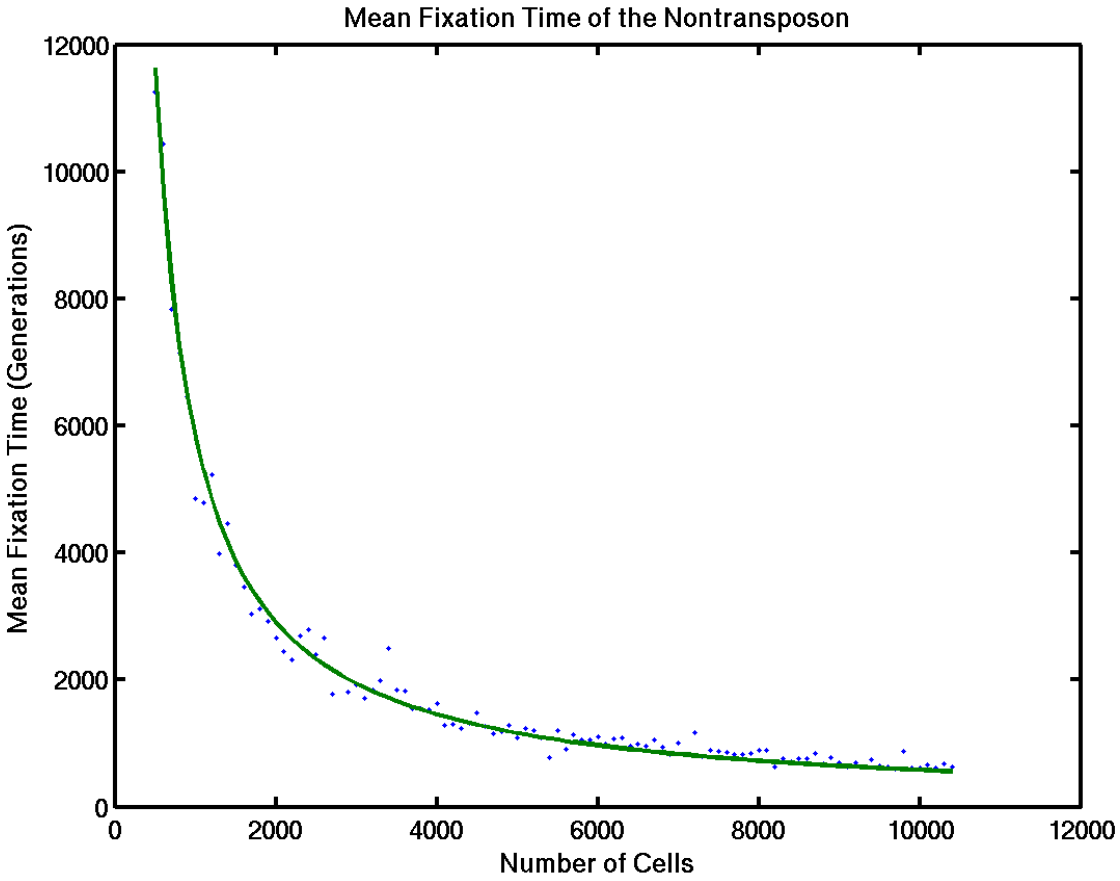


Figure 6.4: Simulation of the extended Moran model. Initially all cells are retrotransposons and have rate $\Delta = 10^{-6}$ /generation/cell to lose functionality. The nontransposon has a selective advantage $s = 0.087$ over the human L1 in *E. coli*. The green curve corresponds to the theory curve $(1/(\Delta N))/p_s = 106/N/0.172$.

arrive at an upper bound for the fixation time by using our estimate for the Moran model without the mutation and starting with a population fraction of $1/N$ of cells with transpsosons knocked out. The additional reaction of mutation strictly makes the fixation time faster and the time scale for the first cell with its retrotransposon knocked out for this initial condition is order unity or smaller. Thus, for larger population size the upper bound for fixation time scales as $s\ln(N)/s$.

Using the conservative estimate from literature for $\Delta = 10^{-8}$ [135] and assuming a population size of 10 million bacteria, a population size comparable to the number of bacteria found in the stomach, our scaling theory for small populations predicts that it will take approximately $\frac{1/N\Delta}{p_f} = \frac{1/(10^{-8}\cdot 10^7)}{0.172} = 58$ generations for the extinction of all bacteria containing retrotransposons with fitness costs similar to EL1H. Similarly, using a larger values of $\Delta$ will relax the population size needed for short extinction time scales. If we consider larger population sizes such as the $10^{14}$ bacteria found in the colon, we can use our scaling argument for the upper bound of fixation time for large populations and find $\frac{2\ln(N)}{s} = \frac{2\cdot 10^{14}}{0.21} = 310$ generations till extinction for bacteria containing retrotransposons with a similar fitness cost to EL1H. A calculation for bacteria containing retrotransposons with fitness costs similar to TL1H gives 745 generations till extinction. Most likely the extinction time would be even faster as subpopulations would form in a spatial environment leading to lower effective population sizes. Even if the populations are much larger the extinction time would be short on evolutionary time scales since at worst the extinction time for bacteria containing retrotransposons scales as the logarithm of the population size. These results show that the timescale for the extinction of retrotransposons in bacteria, with fitness costs close to those measured in experiment, is short enough on evolutionary timescales to explain the limited number of retrotransposable elements found in bacteria.

### 6.7.2 Mean Field Models Containing More Dynamics

I developed the model presented in this section in collaboration with Thomas Kuhlman.

In the previous section we only considered the simple dynamics between two populations of cells, those containing retroelements and those that do not. In this section we consider a more detailed dynamical model that tracks the population of cells with multiple copies of retrotransposons. This model more accurately reflects the reality that cells can contain multiple copies of retroelements. We use this more detailed mean field model to understand how retrotransposons will proliferate within a host genome with the experimentally measured integration rates and growth defects. We construct a model of retroelement activity and analyze its dynamics. We find that retrohoming generally will lead to low but stable numbers of retroelements, while the parameters with which retrotransposition occurs must be finely tuned in order to get long-lived states with significant proliferation of retrotransposons in the host.

First, to introduce direct competition for resources such that extinction is a possible outcome, we construct the model with a limited system size $\Omega$. Within the system, we place $N_x$ cells carrying $x$ copies of the retrotransposon,

leaving $E$ empty space. Normalizing by $\Omega$, the mean behavior of the system is described by the equations

$$
\begin{aligned}
1 &= \varepsilon + \sum_{x=0}^{\infty} \Psi_x, \qquad \varepsilon = \frac{E}{\Omega}, \qquad \Psi_x = \frac{N_x}{\Omega} \\
\frac{\partial \Psi_x}{\partial \tau} &= \varepsilon e^{-bx} \Psi_x - \beta(1-\varepsilon)\Psi_x + \mu(x-1)\Psi_{x-1} - \mu x \Psi_x + \Delta(x+1)\Psi_{x+1} - \Delta x \Psi_x \\
\frac{\partial \varepsilon}{\partial \tau} &= \beta(1-\varepsilon) \sum_{x=0}^{\infty} \Psi_x - \varepsilon \sum_{x=0}^{\infty} e^{-bx} \Psi_x
\end{aligned}
\tag{6.17}
$$

where $\tau$ is the generation time, $\beta$ is the death rate per generation [$\sim 10^{-2} - 10^{-3}$ cell-1 generation-1 [136]], $\delta$ is the mutation rate per retrotransposon per cell per generation resulting in inactivation of a copy of the retroelement [$\sim 10^{-8}$ retrotransposon$^{-1}$ cell$^{-1}$ generation$^{-1}$ [135]], $b$ is the growth defect, and $\mu$ is the transposition rate per retrotransposon per cell per generation. As we have demonstrated above, the values of $\mu$ and $b$ will depend on the retroelement in question and the presence or absence of NHEJ, with $\mu \sim 10^{-2} - 1$ and $b \sim 10^{-2} - 0.6$ for LINE-1, and $\mu \sim 10^{-9} - 10^{-6}$ and $b \sim 10^{-3} - 10^{-2}$ for Ll.LtrB.

To determine non-trivial stationary states, we set time derivatives to zero, and the $\Psi_x$ equations yield a set of recursion relations:

$$
\Psi_x^* = \frac{\beta + (\mu + \Delta)(x-1) - \varepsilon^*(\beta + e^{-b(x-1)})}{\Delta x} \Psi_{x-1}^* - \frac{\mu}{\Delta} \frac{x-2}{x} \Psi_{x-2}^* \tag{6.18}
$$

For example,

$$
\Psi_1^* = \frac{\beta - \varepsilon^*(\beta+1)}{\Delta} \Psi_0^*, \tag{6.19}
$$

which is only non-negative when

$$
\varepsilon^* \leq \frac{\beta}{\beta+1} \tag{6.20}
$$

Inspecting the equation for $\varepsilon$, we find

$$
\varepsilon^* = \frac{\beta}{\beta + \sum_{x=0}^{\infty} e^{-bx} \Psi_x^* / \sum_{x=0}^{\infty} \Psi_x^*} \geq \frac{\beta}{\beta+1}, \tag{6.21}
$$

with equivalence only for $b = 0$. Hence, the only internally consistent nontrivial stationary state is

$$
\begin{aligned}
\varepsilon^* &= \frac{\beta}{\beta+1} \\
\Psi_0^* &= 1 - \frac{\beta}{\beta+1} \qquad (6.22) \\
\Psi_x^* &= 0 \qquad \forall x > 0, \qquad (6.23)
\end{aligned}
$$

i.e., extinction of the retrotransposon. It should be noted that extinction as the sole stationary state is a consequence of the absorbing nature of the wildtype state, $\Psi_0$ . Once cells lose all retrotransposons and enter $\Psi_0$ , there is no way to leave. One possible way to avoid this is by including the possibility of horizontal transfer. However, because the cells in our experiments do not undergo horizontal transfer and the rates of horizontal transfer in the wild are poorly quantified, we do not include this possibility in our modeling.

It is possible there exist interesting non-stationary states or other states that, while not truly stationary, are extremely long lived. We therefore simulated the model equations (6.17) to determine the phase portrait of possible states as a function of $b$ and $\mu$ for the initial conditions beginning with one copy of retrotransposon per cell ($\Psi_1 = 0.1$ and $\varepsilon = 0.9$ ). For the simulations to be tractable, we set a boundary at some maximum number of retrotransposons per cell, $x_{max}$. We consider setting such a boundary in two ways. First, we set a small fixed number of available insertion sites; once occupied, no further insertions are possible (i.e., reflecting boundary conditions). We suggest that such conditions would correspond to the retrohoming of group II introns. Next, from our experimental data, [Figs. 6.7, 6.8, 6.1], we find that when the growth rate has decreased below $\sim 10\%$ of the nominal value, cells cannot survive and cease to grow. Hence, as a second approach in our simulations we set a dynamic boundary by $x_{max} = -\ln(0.1)/b$, where insertions beyond this maximum number result in cell death (i.e., absorbing boundary conditions). We suggest that these conditions would correspond to the nonspecific retrotransposition and amplification of retroelements.

Phase diagrams of simulations with populations of cells allowed to evolve over 10,000 generations are shown in Fig. 6.5A and Fig. 6.5B for reflecting and absorbing boundary conditions, respectively. For both conditions, the majority of parameter values quickly lead to extinction. With reflecting boundary conditions, Fig. 6.5A, a high insertion rate allows saturation of all available integration locations. This corresponds to retrohoming, where insertion rates correspond to $\sim 1$ per intron per cell per generation [121], but with low growth defect. As we now demonstrate, this saturated regime is approximately stable and will persist for extremely long times. With a boundary set at $x_{max}$, the model becomes

$$
\begin{aligned}
&1 = \varepsilon + \sum_{x=0}^{\infty} \Psi_x, \qquad \varepsilon = \frac{E}{\Omega}, \qquad \Psi_x = \frac{N_x}{\Omega} \\
&\frac{\partial \Psi_{x<x_{max}}}{\partial \tau} = \varepsilon e^{-bx} \Psi_x - \beta(1-\varepsilon)\Psi_x + \mu(x-1)\Psi_{x-1} - \mu x \Psi_x + \Delta(x+1)\Psi_{x+1} - \Delta x \Psi_x \\
&\frac{\partial \Psi_{x_{max}}}{\partial \tau} = \varepsilon e^{-bx_{max}} \Psi_{x_{max}} - \beta(1-\varepsilon)\Psi_{x_{max}} + \mu(x_{max}-1)\Psi_{x_{max}-1} - \Delta x_{max} \Psi_{x_{max}} - [\mu x_{max} \Psi_{x_{max}}] \\
&\frac{\partial \varepsilon}{\partial \tau} = \beta(1-\varepsilon) \sum_{x=0}^{\infty} \Psi_x - \varepsilon \sum_{x=0}^{\infty} e^{-bx} \Psi_x + [\mu x_{max} \Psi_{x_{max}}]
\end{aligned}
\tag{6.24}
$$

with the terms in square brackets present only for absorbing boundary conditions. In this case, the $\Psi_x$ equations can

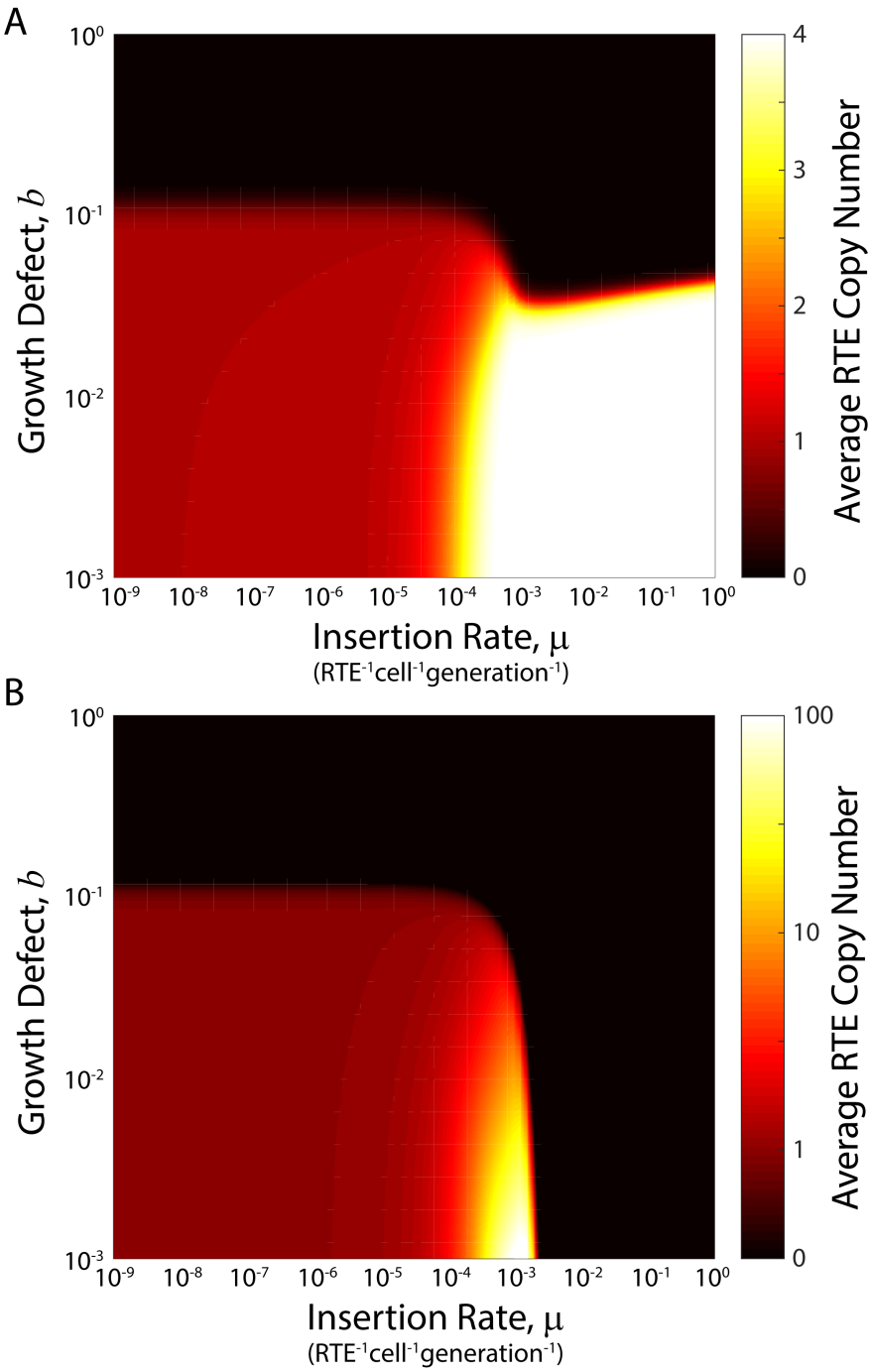


Figure 6.5: **Phase diagram of retrotransposon dynamics**. We simulated the model of retrotransposon dynamics, equations 6.24, using a total system size of $\Omega = 10^9$, with an initial population of $\Psi_1 = 0.1$ and all other states empty. This initial state was allowed to evolve for 10,000 generations with $\Delta = 10^{-8}$ retrotransposon$^{-1}$ cell$^{-1}$ generation$^{-1}$ and $\beta = 10^{-2}$cell$^{-1}$ generation$^{-1}$, at the conclusion of which we calculated the average number of retrotransposons per cell over the extant population. Results are shown for (A) reflecting boundary conditions with $x_{max} = 4$ and (B) absorbing boundary conditions with $x_{max} = -ln(0.1)/b$.

be manipulated to yield recursion relations in terms of the $\Psi_{x_{max}}$ state. In particular, for reflecting boundary conditions we find

$$\Psi^*_{x_{max}-1} = \frac{\beta - \varepsilon^* \left(\beta + e^{-bx_{max}}\right) + \Delta x_{max}}{\mu(x_{max}-1)} \Psi^*_{x_{max}}, \tag{6.25}$$

similar to the condition eq. 6.19 above. To avoid populating lower states and again running afoul of the conditions eqs. 6.20 and 6.21, we demand eq. 6.25 equal zero:

$$\varepsilon^* = \frac{\beta + \Delta x_{max}}{\beta + e^{-bx_{max}}} \approx \frac{\beta}{\beta + e^{-bx_{max}}} \tag{6.26}$$

since $\Delta x_{max}$ is small and approximately negligible. Therefore, only if $\Delta x_{max}$ is negligible, a meta-stable and extremely long-lived state similar to eq. 6.22 and consistent with eq. 6.21 is possible,

$$\varepsilon^* = \frac{\beta}{\beta + e^{-bx_{max}}} \tag{6.27}$$

$$\Psi^*_{x_{max}} = 1 - \frac{\beta}{\beta + e^{-bx_{max}}} \tag{6.28}$$

$$\Psi^*_x = 0 \qquad \forall x < x_{max}, \tag{6.29}$$

This demonstrates that the retrohoming strategy allows for low numbers of retrotransposons that are approximately stable and can persist for extremely long times. For absorbing boundary conditions, the appropriate recursion relation relative to the state with the maximum number of retrotransposons is

$$\Psi^*_{x_{max}-1} = \frac{\beta - \varepsilon^* \left(\beta + e^{-bx_{max}}\right) + (\Delta + \mu)x_{max}}{\mu(x_{max}-1)} \Psi^*_{x_{max}}. \tag{6.30}$$

In contrast with the argument for retrohoming, the non-negligible factor of $\mu x_{max}$ in the numerator renders the $\Psi_{x_{max}}$ state and other states with large retrotransposon numbers unstable. Hence, while the phase portrait Fig. 6.5B shows that there exists a small set of parameter values ($b < 0.01$ and $\mu \sim 10^{-3}$ retrotransposon$^{-1}$ cell$^{-1}$ generation$^{-1}$) where the retroelement is able to proliferate to high numbers, these states will eventually go extinct. Thus, the phase portrait with absorbing boundary conditions rapidly changes with time, and the result shown in Fig. 6.5B depends upon the interval over which the simulation is allowed to run. To determine the lifetime of these states, we performed simulations using absorbing boundary conditions for a variety of values of $b$ and $\mu$, where we recorded the number of generations required for the retrotransposon to go extinct. The result is shown in Fig. 6.6. From this analysis, we see that the time required for a retrotransposon to go extinct can vary over $\sim 7$ orders of magnitude, depending upon

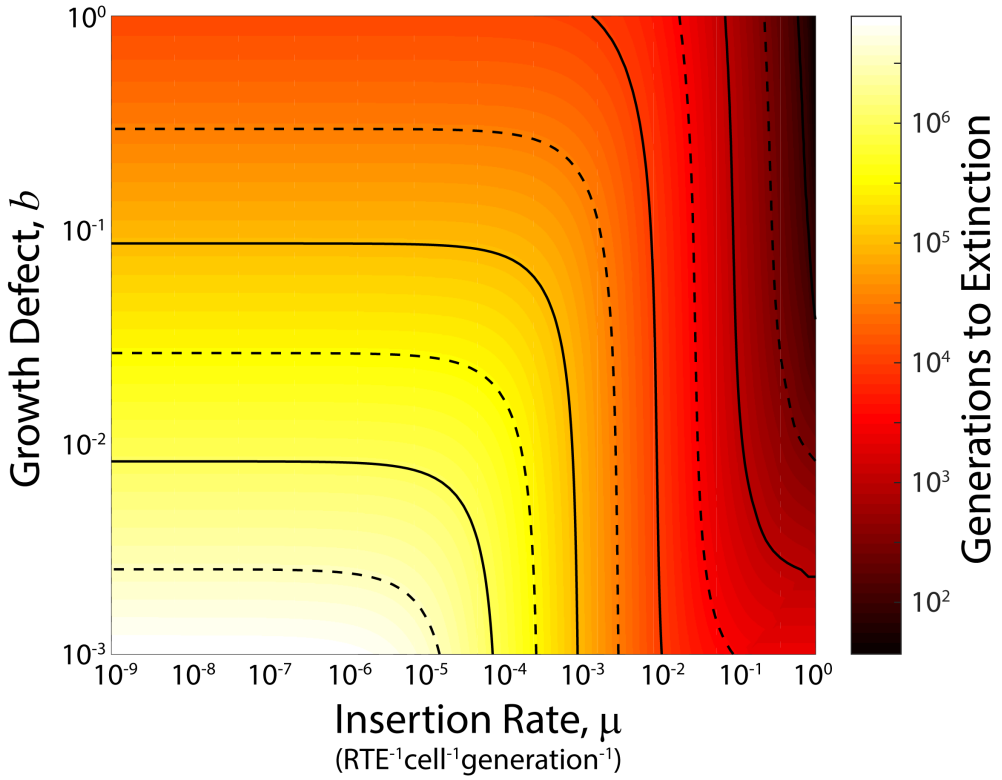


Figure 6.6: **Time to extinction of retrotransposons in a bacterial population**. Simulations of the model eqs. 6.24, with absorbing boundary condition at $x_{max} = -\ln(0.1)/b$ , system size of $\Omega = 10^{30}$, $\Delta = 10^{-8}$ retrotransposon$^{-1}$ cell$^{-1}$ generation$^{-1}$, $\beta = 10^{-2}$ cell$^{-1}$ generation$^{-1}$ and initial population of $\Psi_1 = 0.1$ with all other states empty. Color indicates the number of generations required for the average number of retrotransposons per cell to drop below $\frac{1}{\Omega}$. Solid contour lines indicate major decade divisions, dashed contour lines indicate half-decade divisions.

its dynamics and effects. For those parameter regimes corresponding to the aggressive autonomous retrotransposon LINE-1 ($b \geq 10^{-2}$, $\mu \geq 10^{-2}$ retrotransposon$^{-1}$ cell$^{-1}$ generation$^{-1}$), extinction is rapid, occurring in $\sim$ 100 – 10,000 generations. Conversely, parameter regimes corresponding to the group II intron Ll.LtrB ($10^{-3} \leq b \leq 10^{-2}$, $10^{-9} \leq \mu 10^{-6}$ retrotransposon$^{-1}$ cell$^{-1}$ generation$^{-1}$) can persist in low copy numbers ($\sim$ 1 per cell) for millions to tens of millions of generations. We additionally see that the small parameter regime where retrotransposons can proliferate to high copy numbers ($b \leq 10^{-2}$, $\mu \sim 10^{-3} - 10^{-4}$ retrotransposon$^{-1}$ cell$^{-1}$ generation$^{-1}$) persists for hundreds of thousands to millions of generations, and could be maintained longer with the inclusion of horizontal gene transfer.

## 6.8 Discussion

Our experimental collaborators found that NHEJ enhances the efficiency of LINE-1 integration and thus its lethality. They also found the same result for the group II intron L1.LtrB.

Additionally, our experimental collaborators found that both human L1H and bacterial L1.Ltrb expression results

in an exponential decrease in growth rate. I explained this observation by developing a simple model that assumes each transcript has a probability of integrating and disrupting essential genes affecting cell growth. The cell will survive as long as no essential genes are interrupted; this leads to the exponential growth defect. The measured growth defect of the retrotransposon ELH1 in the simple Moran model show the retrotransposon should go extinct within a couple hundred generations. These relatively short extinction times are consistent with the non-observation of retrotransposons in bacteria.

In the more detailed model I developed with Thomas Kuhlman, aggressive retrotransposons with parameters similar to the ones we measured for L1 would go extinct within 100 to 10000 generations. For parameter regimes corresponding to group II introns, the group II introns can persist in low copy numbers for millions of generations. Furthermore, retrotransposons can persist in high copy numbers if the growth defect is decoupled from the integration rate. In particular this requires suppression of the growth defect below $b \sim 10^{-2}$. Many of the features unique to eukaryotes, including alternative RNA splicings enabled by the spliceosome, spatial and temporal decoupling of transcription and translation by the nuclear membrane, or utilizing existing junk DNA already present in the genome as a large, non-vital target in which to sink integrations have been hypothesized to have arisen specifically to mitigate the physiological effects of retroelements in eukaryotes [13, 14, 12].

## 6.9 Supplement: Experimental Details

The following supplement provides experimental details conducted by Thomas Kuhlman and his lab.

### 6.9.1 Effects of Retroelement Expression on Growth

To assess the effects of L1 expression on bacteria, we first transformed pTKIP-TL1H/EL1H constructs into *E. coli* BL21(DE3), a strain that expresses T7 polymerase [137]. A decrease in growth rate in response to increasing L1 expression is immediately apparent in cultures titrated with IPTG (Fig. 6.7B-C). To test the generality of this effect, we next assessed the effects of L1 expression on *Bacillus subtilis*. In contrast to *E. coli*, *B. subtilis* is a Gram-positive bacterium able to repair DNA double strand breaks through nonhomologous end joining (NHEJ) in a manner similar to eukaryotes [138]. Hence, we hypothesized that *B. subtilis* would be more resistant to L1 and cleavage of DNA as a result of ORF2p endonuclease than *E. coli*, which lacks capacity for NHEJ repair. Instead, we find the opposite: wildtype *B. subtilis* cannot survive transformation with pHCMC05-EL1H (Fig. 6.7D). Conversely, we obtain high yield transformation of EL1H into *B. subtilis* strains with NHEJ repair enzymes Ku (*ykoV*), LigD (*ykoU*), and both Ku and LigD knocked out [139]. A Miller assay of expression level from the positive control plasmid pHCMC05-lacZYAX expressing *E. coli's* metabolic lac enzymes from the hyper-spank promoter shows that expression is weak

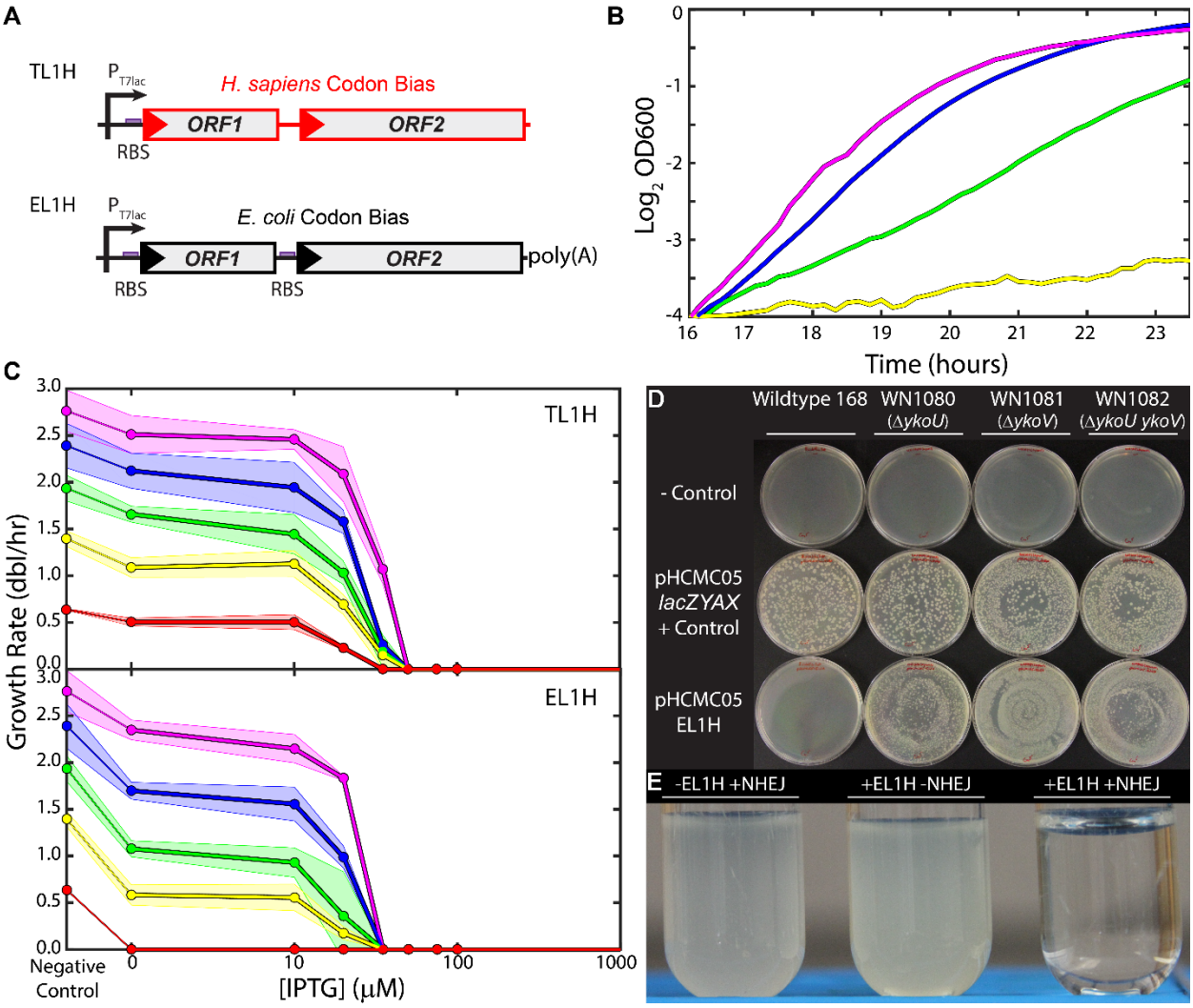


Figure 6.7: **Bacterial L1 elements and effects on growth**. (A) LINE-1 constructs used in this study. Top: TL1H has human codon bias (indicated by red), and was modified for expression in *E. coli* using a bacterial T7*lac* promoter and a consensus Shine Dalgarno ribosomal binding site driving *ORF1* (RBS, purple box). Bottom: EL1H is driven by $P_{T7lac}$ and has consensus RBS for *ORF1* and *ORF2*. EL1H has a 100 bp 3' poly(A) tract and has *E. coli* codon bias (indicated by black). (B) L1 is detrimental to *E. coli* growth. Example growth curves for BL21(DE3) pTKIP-TL1H growing in M63 glucose medium including 0 (magenta), 10 μM (blue), 20 μM (green), and 35 μM (yellow) IPTG. (C) Growth response as a function of [IPTG] for BL21(DE3) pTKIP-TL1H (top) and pTKIP-EL1H (bottom) in various media; magenta - RDM glucose, blue - RDM glycerol, green – cAA glucose, yellow – M63 glucose, red – M63 glycerol. Growth rates were determined using the slope of the best fit regression of the initial linear portion of $Log_2$(OD600) versus time, as in (B). Points are the average of three independent replicates, and shaded regions indicate the standard deviation. (D) Wildtype *B. subtilis* cannot survive transformation with EL1H (first column), while NHEJ knockouts relieve sensitivity (second column: Δ*ykoU*; third column Δ*ykoV*; fourth column Δ*ykoU* Δ*ykoV*). First row: negative control (TE buffer only); second row: positive control (pHCMC05-lacZYAX).; third row: pHCMC05-EL1H. We performed transformations in four independent replicates with identical results. (E) Example *E. coli* BL21(DE3) cultures in RDM glucose grown for 20 hours. Left - pTKIP, pUC57-NHEJ; middle – pTKIP-EL1H, pUC57; right – pTKIP-EL1H, pUC57-NHEJ. All cultures contain no IPTG and 100 ng/ml aTc.

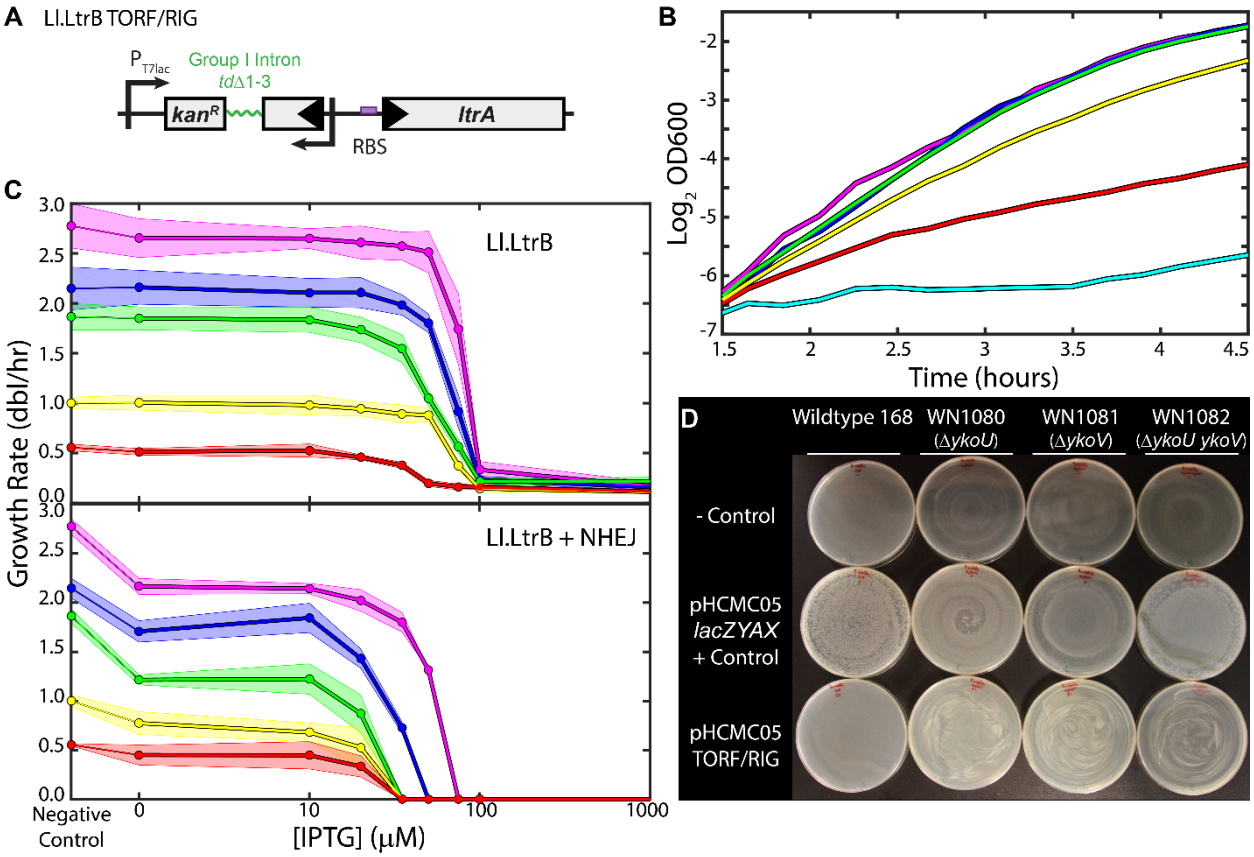


Figure 6.8: **Effects of Ll.LtrB on bacterial growth.** (A) The Ll.LtrB construct TORF/RIG. TORF/RIG drives the expression of the Ll.LtrB group II intron, with the ltrA coding sequence towards the 3' end of the intron driven by a strong RBS. TORF/RIG includes a kanamycin resistance gene encoded in the opposite orientation whose coding sequence is disrupted by the group I intron tdΔ1-3 for determination of retrotransposition frequencies. (B) Expression of TORF/RIG is detrimental to *E. coli* growth. Example growth curves for BL21(DE3) pET-TORF/RIG growing in M63 glucose medium including 0 (magenta), 10 μM (blue), 20 μM (green), 35 μM (yellow), 50 μM (red), and 100 μM (cyan) IPTG. (C) Growth response as a function of [IPTG] for BL21(DE3) pET-TORF/RIG pZA31-tetR (top) and pET-TORF/RIG pZA31-NHEJ (bottom) in various media; magenta - RDM glucose, blue - RDM glycerol, green – cAA glucose, yellow – M63 glucose, red – M63 glycerol. Growth rates were determined using the slope of the best fit linear regression line of $Log_2$(OD600) versus time, as in (B). Points are the average of three independent replicates, and shaded regions indicate the standard deviation. (D) Wildtype *B. subtilis* cannot survive transformation with pHCMC05-TORF/RIG (first column), while NHEJ knockouts somewhat relieve sensitivity (second column: Δ*ykoU*; third column Δ*ykoV*; fourth column Δ*ykoU* Δ*ykoV*). First row: negative control (TE buffer only); second row: positive control (pHCMC05-lacZYAX).; third row: pHCMC05-TORF/RIG. We performed transformations in four independent replicates with identical results.

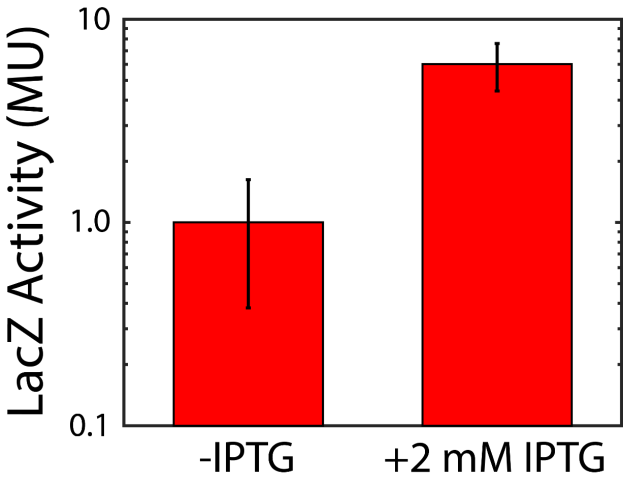


Figure 6.9: **Expression from the hyper-spank promoter of pHCMC05 in *Bacillus subtilis*.** LacZ activity of uninduced (left) and induced (right) *B. subtilis* 168 transformed with pHCMC05-lacZYAX was measured with a Miller assay.[140] Bars are the mean of six independent replicates and error bars are the standard deviation

but leaky in *B. subtilis* (Fig. 6.9). We conclude that wildtype *B. subtilis* is extremely sensitive to very low levels of L1H expression, and that this growth defect is enhanced by NHEJ repair.

We next cloned and expressed the *B. subtilis* NHEJ enzymes (BsKu and BsLigD) in *E. coli* under the control of the aTc inducible PLtet01 promoter [96]. We first verified that BsKu and BsLigD were functional in *E. coli* by ensuring their ability to rescue strains where we induced the homing endonuclease I-SceI to create double stranded chromosomal breaks at chromosomally integrated I-SceI recognition sites[141, 125, 126] [Fig. 6.10]. We then verified the enhancement of lethality of LINE-1 by NHEJ by cotransformation of BL21(DE3) with plasmids expressing LINE-1 and NHEJ enzymes. We find that even low leakage expression of EL1H without addition of IPTG is lethal to *E. coli* with concomitant induction of expression of NHEJ enzymes with 100 ng/ml aTc (Figure 6.7E).

To quantify the effect of L1 expression on *E. coli* growth, we measured the growth rate as a function of expression level by titration with IPTG and periodic measurement of optical density in a variety of growth media (Fig. 6.7B-C). Even with no induction, leaky expression of L1 significantly reduces the growth rate relative to the parent strain carrying an empty plasmid, and complete growth arrest occurs at IPTG concentrations of 35 – 50 μM (Fig. 6.7C).

We measured the transcriptional response function of the T7*lac* promoter by quantitative reverse transcription PCR (qRT-PCR, Fig. 6.11A-D) of L1 mRNA extracted from bacteria grown at those IPTG concentrations where cultures survive. This yielded the response function shown in Fig. 6.11E. The resulting dose-response as a function of L1 RNAs per cell is shown in Fig. 6.1A, with data from TL1H as blue points, EL1H as red points, and EL1H + NHEJ as black points. The normalized growth rate decreases exponentially with increasing numbers of L1 RNAs, and growth conditions do not affect this response. Solid lines in Fig. 6.1A correspond to fits to the exponential function , where L is the average number of L1 RNAs per cell and the parameter b quantifies the growth defect and sensitivity to L1

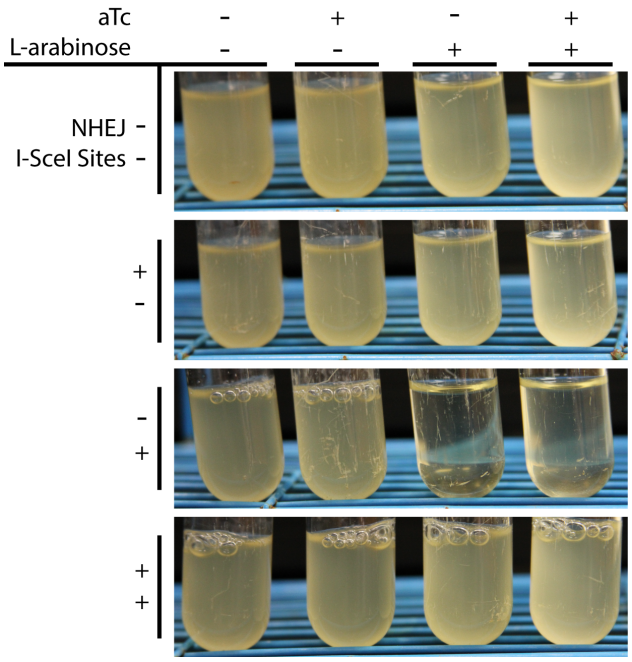


Figure 6.10: ***B. subtilis* NHEJ enzymes function in *E. coli*.** Turbidity of cultures grown for $\sim$ 36 hours at $30^oC$ after inoculation with identical amounts of cells. Bacterial strains are MG1655 Δlac carrying the plasmid pTKRED, which expresses the homing endonuclease I-SceI when induced with L-arabinose. Additional plasmids and modifications are, from top to bottom - first row: pUC57-kan; second row: pUC57-kan-NHEJ; third row: pUC57-kan with I-SceI sites integrated at the atpI chromosomal locus [141, 125, 126]; fourth row: pUC57-kan-NHEJ with I-SceI sites integrated at the atpI chromosomal locus. Columns correspond to different inducer conditions – first column: 0 aTc, 0 L-arabinose; second column: 100 ng/ml aTc, 0 L-arabinose; third row: 0 aTc, 0.4% w/v L-arabinose; fourth row: 100 ng/ml aTc, 0.4% L-arabinose. Lack of turbidity in row 3, columns 3 and 4 demonstrate that I-SceI double strand breaks are lethal to *E. coli* [125]. Recovery of turbidity in row 4, columns 3 and 4 demonstrate that even low, leakage expression of *B. subtilis* NHEJ enzymes rescue *E. coli* growth.

expression. We find that, on average, each L1 transcript yields a decrease in *E. coli's* growth rate of $\sim 0.83 \pm 0.06\%$ (TL1H) or $1.9 \pm 0.6\%$ (EL1H) in the absence of NHEJ, and $\geq 45 \pm 1.6\%$ with NHEJ.

We find that Ll.LtrB also yields a growth defect as a function of expression level, as qualitatively reported by Coros, et al. 2005 [121]. Using Ll.LtrB expressed from pET-TORF/RIG in *E. coli*, we measured the growth rate as a function of expression level by titration with IPTG and periodic measurement of optical density in a variety of growth media (Fig. 6.8B-C), yielding the dose response shown in Fig. 6.1B as red points. As with L1, simultaneous expression of *B. subtilis* NHEJ enzymes significantly enhances the lethality of Ll.LtrB to *E. coli* (Fig. 6.1B, black points). As might be expected due to the ability of LtrA maturase to excise Ll.LtrB from interrupted genes, the growth defect resulting from Ll.LtrB is weaker than that from L1, with each Ll.LtrB transcript reducing growth rate by $0.11 \pm 0.02\%$ in the absence of NHEJ and $0.82 \pm 0.11\%$ with NHEJ. The Ll.LtrB growth defect is also evident in plating assays to determine retrotransposition efficiency (described below). Induction of Ll.LtrB expression with 100 µM IPTG reduces the number of viable colony forming units (cfus) per milliliter per OD by $\sim$ 10x. Simultaneous induction of Ll.LtrB with 100 µM IPTG and induction of NHEJ enzymes with 100 ng/ml anhydrotetracycline reduces viable cfus/OD/ml by $\sim$ 100x, while induction of expression of NHEJ enzymes alone has no detectable effect.

Finally, we attempted to transform Ll.LtrB into *B. subtilis* as the plasmid pHCMC05-TORF/RIG, with Ll.LtrB under control of the lacI-regulated hyper-spank promoter. As with LINE-1, we find that wildtype *B. subtilis* 168 cannot survive transformation with Ll.LtrB, while knockouts for the NHEJ genes *ykoU*, *ykoV*, and both *ykoU* and *ykoV* are transformed with high yield (Fig. 6.8D). However, NHEJ knockouts do not alleviate the grow defect as significantly as with LINE-1, with Ll.LtrB transformants growing slowly to form very small colonies after $\sim$ 24 hours of growth at $37^oC$.

### 6.9.2 L1 Successfully Integrates into *E. coli's* Chromosome

We next addressed the question of how L1 is functioning in bacteria and the molecular mechanisms causing growth defects. Since both ORF2p and LtrA contain an endonuclease domain, expression of these proteins alone may damage genomic DNA and halt growth without being accompanied by successful retrotransposition. However, the hypothesis that DNA damage by endonucleases is primarily responsible for growth defects appears inconsistent with observation of NHEJ enhancing the growth defect. We now report multiple lines of evidence indicating that L1 successfully integrates into *E. coli's* chromosome, and that NHEJ enhances the efficiency of retrotransposition of both L1 and Ll.LtrB.

First, we grew cultures carrying EL1H with 30 µM IPTG for $\sim$ 48 hours. Surviving bacteria were collected and transformed with the plasmid pTKRED, which expresses the homing endonuclease I-SceI [141, 125, 126] resulting in *in vivo* digestion and curing of pTKIP-EL1H. After screening of the resulting cultures for appropriate antibiotic

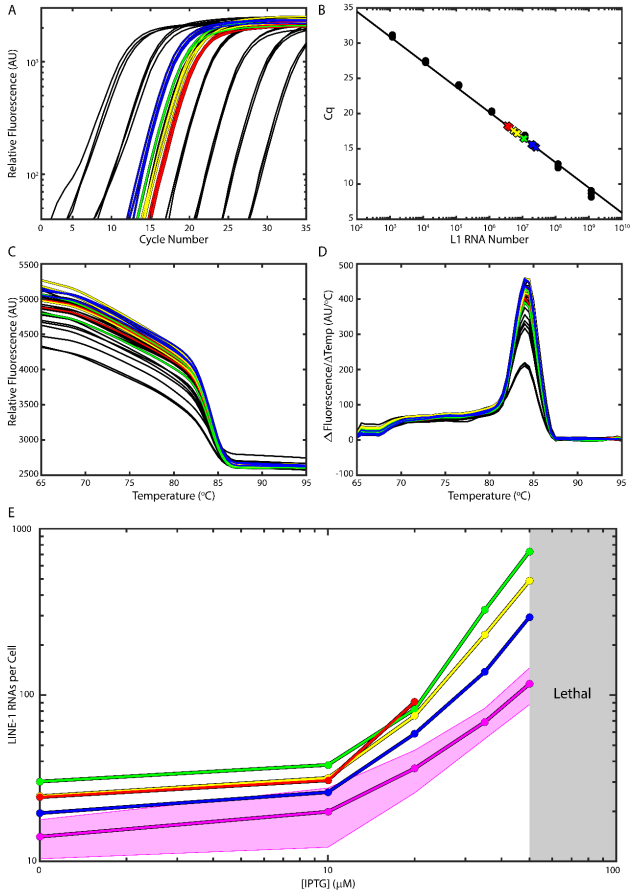


Figure 6.11: **Quantitative RT-PCR to determine T7*lac* promoter response function.** (A) Amplification curves of reverse transcribed serial 10x dilutions of *in vitro* transcribed TL1H RNA as an absolute standard (black), along with reverse transcribed RNA extracted from BL21(DE3) pTKIP-TL1H grown in M63 glucose medium with 0 (red), 10 μM (yellow), 20 μM (green), and 50 μM (blue) IPTG. (B) Absolute quantification of TL1H RNA numbers. Black circles are critical cycle numbers (Cq) of the *in vitro* standards from (A), colored crosses are Cqs of BL21(DE3) pTKIP-TL1H RNA with the threshold at $\sim$ 200 AU. PCR efficiency was 90.5%. (C) Melting curves and their unimodal derivatives (D) resulting from qRT-PCR, demonstrating clean amplification of TL1H cDNA. Melting temp of the amplicon was 84.5 oC. (E) RNA was extracted from BL21(DE3) pTKIP-TL1H grown in RDM glucose (magenta), RDM glycerol (blue), M63 glucose (yellow), cAA glucose (green), or M63 glycerol (red) with 0, 10, 20, 35, or 50 μM IPTG and quantified through qRT-PCR (Figure 6.11). Concentrations of IPTG higher than 50 μM were nonviable in all media except M63 glycerol, where concentrations higher than 20-35 μM were generally nonviable. The number of RNAs determined by qRT-PCR was divided by the number of cells added to the reaction, determined by measurement of OD600 and plating performed at the time of harvest. Shaded magenta region shows the standard error of the mean of four experimental replicates for samples prepared in RDM glucose. The standard errors of other samples are similar, but not shown for clarity. The number of LINE-1 RNAs per cell obtained for each growth and induction condition thus obtained were used as the x-axis in Figure 6.1.

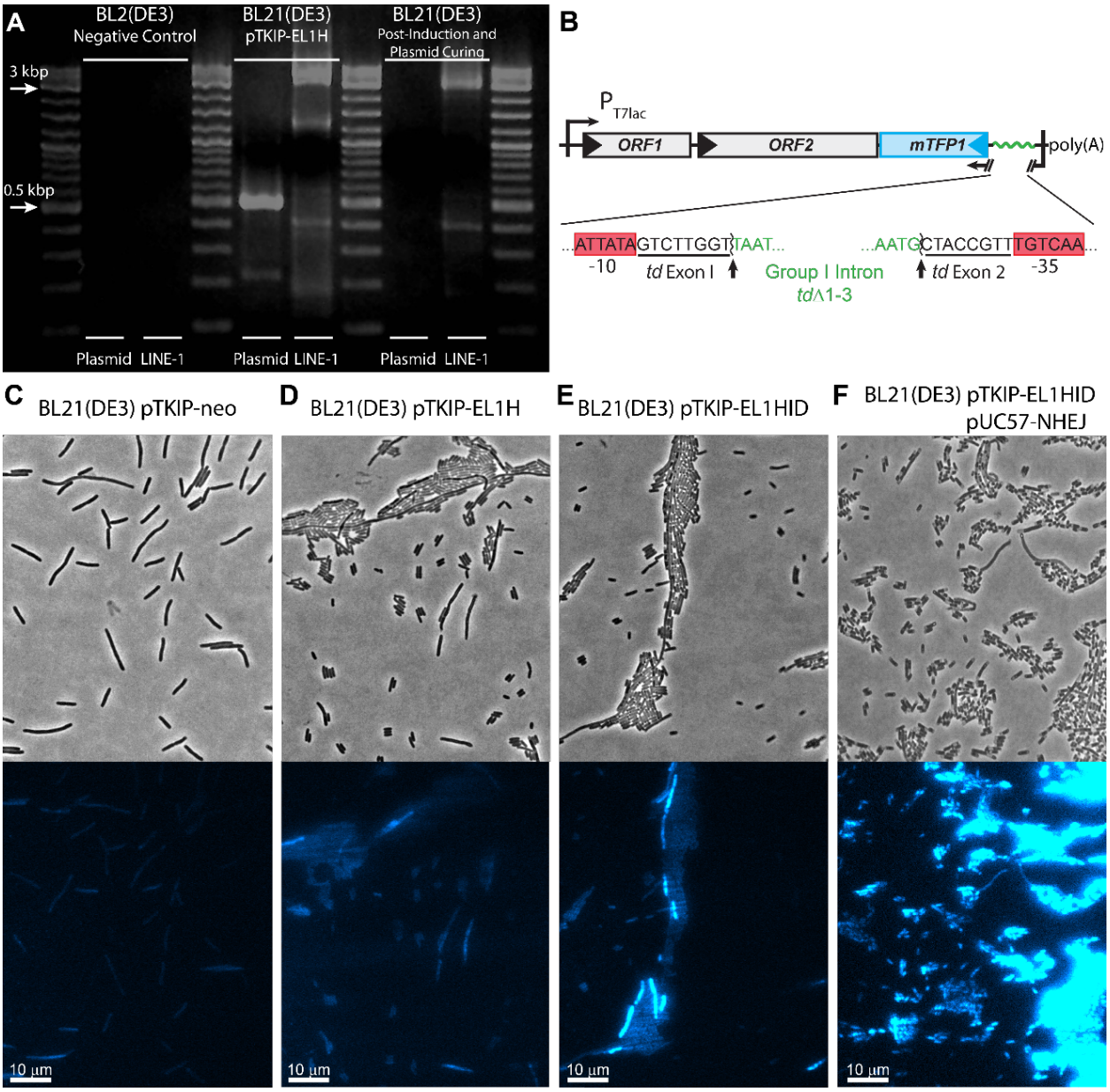


Figure 6.12: **L1 integrates into the *E. coli* genome.** (A) Non-clonal colony PCR to detect EL1H (LINE-1 lanes) and pTKIP (Plasmid lanes). Left: BL21(DE3) negative control. Middle: BL21(DE3) pTKIP-EL1H positive control. Right: Strain post EL1H exposure and plasmid curing. (B) EL1HID, a construct for detecting successful retrotransposition of EL1H in individual cells by fluorescence. The integration detection cassette (ID) consists of *mTFP1* with consensus $\sigma$70 promoter and RBS. -10 and -35 core promoter sequences are split by the group I intron tdΔ1-3 (sequences shown below). Upon successful retrotransposition the cell fluoresces blue. (C) Induced BL21(DE3) pTKIP-EL1HID are visibly fluorescent with UV illumination. (D-F) Phase contrast (top) and fluorescence microscopy (bottom) of induced (20 µM IPTG) (D) BL21(DE3) pTKIP-neo negative control, (E) BL21(DE3) pTKIP-EL1H, (F) BL21(DE3) pTKIP-EL1HID, and (G) BL21(DE3) pTKIP-EL1HID pUC57-NHEJ (0 IPTG, 5 ng/ml aTc).

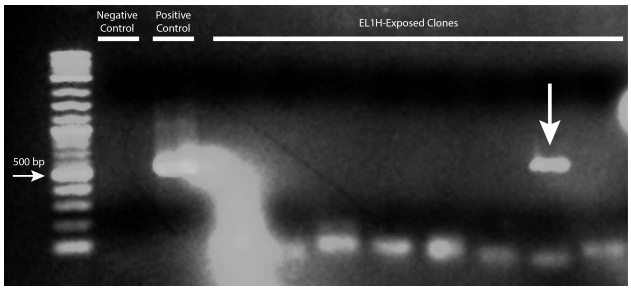


Figure 6.13: **Detection of full-length EL1H genomic integrants.** Representative 2% agarose gel electrophoresis of colony PCR of eight isolated colonies of BL21(DE3) that had been exposed to EL1H and cured of pTKIP-EL1H using primers that anneal to the 5' end of EL1H and produce a 500 bp amplicon. BL21(DE3) was used as a negative control, and BL21(DE3) pTKIP-EL1H as a positive control. The large fluorescent smear near the positive control band was a result of excess ethidium bromide staining. Since EL1H RNA is reverse transcribed and integrated starting from the 3' end, presence of the 5' end indicates complete integration. Out of 80 colonies tested, we found 3 colonies yielding this 500 bp product indicating complete integration of EL1H.

resistances, we performed colony PCR to verify plasmid loss and to attempt to amplify L1 from genomic DNA. An example obtained from exposed cultures is shown in Fig. 6.12A. Post-curing strains generate no product corresponding to presence of the pTKIP plasmid, yet we were able to amplify EL1H from non-clonal samples. We subsequently isolated single, clonal colonies of EL1H-exposed *E. coli* and attempted to amplify a 500 bp segment containing the 5' end of *ORF1* by colony PCR. We detected a positive signal in 3 out of 80 colonies screened (Fig.6.13).

As an additional phenotypic test for successful retrotransposition, we modified EL1H by inserting a cassette between the 3' end of *ORF2* and the poly(A) tract for detecting integration in individual live cells by fluorescence (Fig. 6.1) [122, 15, 142]. This cassette consists of a gene encoding mTFP1 [143], a bright teal fluorescent protein, driven by a consensus *E. coli* $\sigma$70 promoter and RBS. The -10 and -35 sequences of the mTFP1 promoter are split by the group I intron $td\Delta 1-3$ [144], preventing transcription of mTFP1 from the $\sigma$70 promoter in the original construct. Upon transcription of EL1H, the $td\Delta 1-3$ intron catalyzes its own excision, reconstituting the mTFP1 promoter in the EL1H RNA. Finally, if this part of EL1H RNA is successfully reverse transcribed and integrated into the genome, individual cells undergoing successful retrotransposition can be detected by fluorescence microscopy.

We transformed this construct, EL1HID, into BL21(DE3) and grew cultures with weak induction of L1 expression. We have previously observed that dead *E. coli* produce stronger autofluorescence than live cells, raising the possibility that any observed fluorescence is simply due to a higher proportion of dead cells. However, fluorescence microscopy shows that cultures carrying EL1HID contain a subpopulation of cells (-NHEJ: $\sim$ 1%; +NHEJ: $\sim$ 80%) whose total fluorescence is $>$10x brighter than cells in any control strains.

pET-TORF/RIG contains a retromobility indicator gene (RIG) that functions similarly to the mTFP1 cassette employed in EL1HID [121], in that the coding sequence of a kanamycin resistance gene carried by Ll.LtrB is interrupted by the td$\Delta$1-3 group I intron. Consequently, the frequency of successful ectopic retrotransposition can be determined

by plating cultures on selective medium containing kanamycin. We performed retrotransposition plating assays with BL21(DE3) pET-TORF/RIG cells carrying pUC57-cat-NHEJ, a chloramphenicol-resistant high copy number plasmid with a pUC origin of replication. To prevent confounding effects from possible incompatibility between pUC57's pUC origin and pET-TORF/RIG's pBR322 origin, we also performed retrotransposition assays using pZA31-NHEJ, a medium copy number plasmid with a p15A origin of replication [96]. We find the efficiency of Ll.LtrB retrotransposition in BL21(DE3) pET-TORF/RIG cells carrying empty pUC57-cat or pZA3-tetR plasmids as negative controls to be $3.0 \pm 0.9 \times 10^{-9}$, consistent with measurements by Coros et al. [121]. In contrast, the efficiency of retrotransposition of cells carrying pUC57-cat-NHEJ was $4.6 \pm 0.4 \times 10^{-6}$, while those carrying pZA31-NHEJ was $1.5 \pm 0.3 \times 10^{-7}$. Hence, we find that bacterial NHEJ increases the efficiency of Ll.LtrB retrotransposition by 2 – 3 orders of magnitude.

# Part III

# Stochastic Dynamics of Ants

# Chapter 7

# Stochastic Dynamics of Ants

Bistability is ubiquitous. It has been used to describe flip-flop circuits in electronics, regulation motifs in cellular signaling, and a particle in a double well potential in mechanics. Perhaps the most familiar mechanism for bistability is that of a particle executing a Brownian walk in a double well potential. In the absence of noise the particle will tend to settle to a fixed point of one of the potential wells. If the noise or temperature of the system is increased the particle can be kicked out of one well and settle into the other. For this type of bistability scientists have asked and answered questions such as: What is the stationary probability distribution of the particle? What is the mean switching time of the particle? And how stable are the states relative to one another[145]? In this section I will describe another type of bistability, where surprisingly the bistability is created by multiplicative noise.

This alternative type of bistability is created by multiplicative noise, such as that which arises from intrinsic demographic stochasticity. For example, a particle sits in a potential well with one fixed point, but the multiplicative noise is greatest at the fixed point and vanishes at the boundaries of the well. So one can imagine the dynamics of a particle in such a system will be such that as the particle relaxes towards the bottom of the potential well it is immediately kicked out and experiences less kicks as it gets closer to a boundary. These boundary states are metastable and the particle can switch between them. This kind of bistability was first observed in the Togashi-Kaneko reaction scheme [146] and later understood to be caused by multiplicative noise by Ohkubo et al. [147]. The characteristic equation of this type of bistability is given by $\dot{z} = -z + s\sqrt{1-z^2}\eta$, where $\eta$ is Gaussian noise, $z$ is the bi-stable quantity switching between -1 and 1, and $s$ is a quantity controlling the strength of the noise. Notice that when the noise vanishes, $s = 0$, the system is deterministic with a fixed point at 0. This hints at a phase transition unique to this type of bistability. When the noise is small, the bistability vanishes and the system relaxes to its fixed point. Figure 7.1 displays the stationary probability distribution for systems that have this type of multiplicative noise-induced bistability for different strengths of noise.

This kind of bistability can arise from systems with autocatalytic reaction networks [148]. Instead of introducing the noise by hand, the noise comes about as a consequence of the underlying individual level model and the discreteness of the chemical copy numbers. The form of the noise is derived by expanding the Master Equation with respect to system size using the Kramers-Moyal expansion. This type of bistability has been shown to occur in systems with

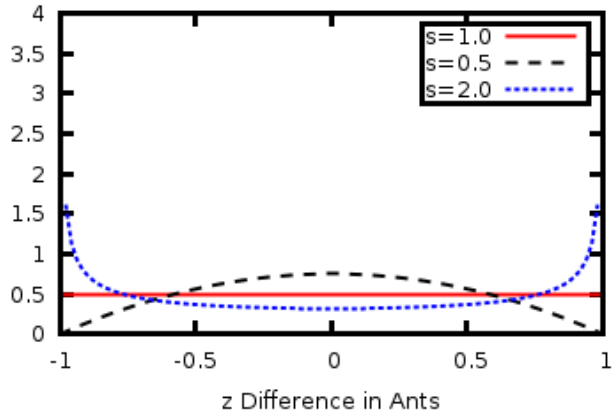


Figure 7.1: The stationary probability distribution for bistability characterized by multiplicative noise. The distribution is plotted for different strengths of niose, s=0.5, 1.0, and 2.0

recruitment, such as ants foraging from two food sources spaced equally away from a nest [1].

## 7.1 Direct Recruitment Model

In previous work by Tommaso Biancalani, a simple model of ant foraging from two identical food sources was proposed. In this proposed model there are two types of ants, ants foraging from food source 1, denoted by $x_1$, and ants that forage from food source 2, denoted by $x_2$. There is a total fixed number of ants, $N$, foraging at all times. Ants can directly recruit one another to forage from their food source. For example, an ant $x_1$ could recruit an ant $x_2$ to start to forage from food source one. In this case the ant $x_2$ would become an $x_1$. This recruitment of ants is autocatalytic. Finally, in Biancalani's model there is a small chance that an ant can spontaneously start to forage from the other food source. The set of reactions describing this model is as follows:

$$\begin{aligned} x_1 + x_2 &\xrightarrow{r} 2x_2, & x_2 + x_1 &\xrightarrow{r} 2x_1 \\ x_1 &\xrightarrow{\varepsilon} x_2, & x_2 &\xrightarrow{\varepsilon} x_1 \end{aligned} \tag{7.1}$$

Since the total number of ants is a conserved quantity, the behavior of this system can be completely described by one variable $z = x_1 - x_2$, the difference in number densities of ants foraging from source one and source two. Writing down the Master Equation for this set of reactions and using a Kramers-Moyal expansion in system size, Biancalani found that $z$ obeyed the following stochastic differential equation:

$$\dot{z} = -z + \sqrt{\frac{N_c}{N}}\sqrt{1 + 2\varepsilon - z^2}\,\eta\,\tau \tag{7.2}$$

Where $N_c = 1/\varepsilon$. This equation has a stationary probability solution of:

$$P(z) = \frac{C_0}{(1+2\varepsilon - z^2)^{1-N/N_c}} \tag{7.3}$$

As can be seen from this stationary solution when $N < N_c$ the system is mainly in $z = 1$ or $z = -1$. Observations of trajectories of $z$ in this steady state show the system is switching between the states $z = 1$ and $z = -1$; that is, the ants are all foraging from one food source then foraging from the other food source. This corresponds to the autocatalytic process being much more important than the random spontaneous decisions of individual ants. When there are enough ants, $N > N_c$, then the ants will start to forage equally from both food sources. In this system the strength of the noise in the stochastic pde is controlled by the population size: at small populations the strength of the fluctuations is large and at large population sizes the strength is small. In this model there exists a critical population size above which bistability vanishes.

Biancalani's model would predict that if the population size is small enough, then the ants would bistably forage first from one and then the other food source. However, this does not seem to be experimentally the case. Experiments by Beckers et al. examined ants foraging from two food sources placed equidistant from an ant colony. The ants foraged preferentially from one of the food sources with roughly 80% of the ants foraging from that food source [149]. Switching was not observed but perhaps this could be due to the duration of the experiment being too short.

Biancalani's original model exhibits bistability but leaves out many of the important details of foraging ants. Ants have three main types of recruitment: tandem recruitment, group recruitment, and trail recruitment (mass recruitment). In tandem recruitment, a scout guides one recruit to the source. In group recruitment, a scout guides a group of ants to the food source. In trail recruitment, a trail of pheromones is laid by the scout on the way back to the nest after finding a food source. Subsequent ants can reinforce the trail. Tandem recruitment and group recruitment usually occur along with trail recruitment. One of the most important details that was ignored in Biancalani's model was that some ants do recruitment through pheromones.

Inspired by an ant foraging model in Netlogo, I further modified that model to exhibit similar behavoirs as observed in Biancalani's model (see Fig.7.2). In this modified Netlogo model, I was able to observe bistable foraging of ants when the population size was small enough. I also saw that I could modify the critical population size above which ants will forage equally from both food sources simply by modifying the evaporation rate of the pheromones. This led me to create a simple extension of Biancalani's model to include pheromones which I describe in the next section.

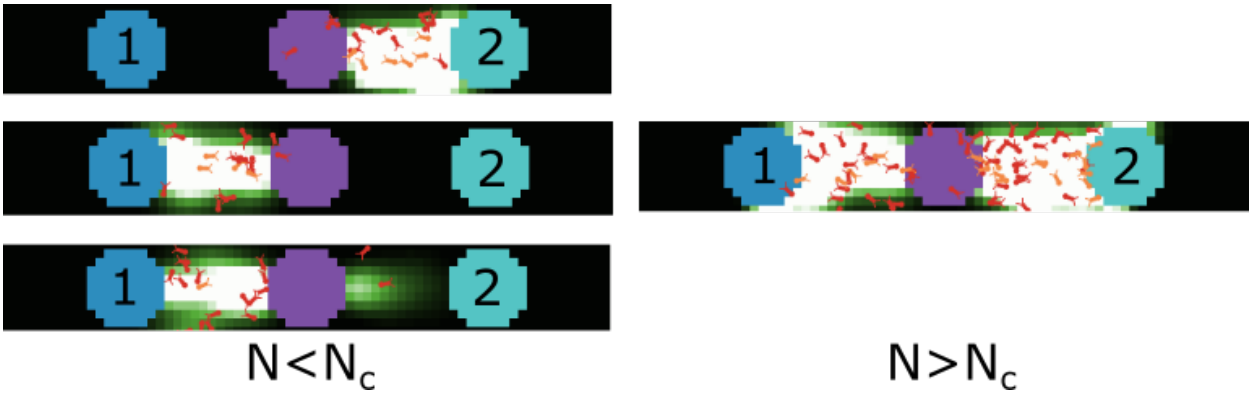


Figure 7.2: Netlogo simulations of ants foraging from two food sources. The food sources are indicated by the numbers 1 and 2, the purple circle in the center is the location of the ant colony. The green color gradient shows the amount and location of the pheromones. The figures on the left show ants bistably foraging when their population size is small, $N < N_c$. The figure on the right shows ants foraging equally from both food sources when their population is large, $N > N_c$.

## 7.2 Stochastic Model for Ant Foraging with Pheromones

By considering pheromones I am able to examine the role of memory (how long the pheromones last on a surface) and to examine if the pheromones have an effect on the critical population size and thus, the switching time. I derive experimental predictions for how the critical switching size depends on the evaporation rate of the pheromones and the rate at which the pheromones are created. This theory is rich enough that it can be systematically compared with experiment. It would also be possible to compare our predictions with earlier theoretical work[150, 151] which does not correctly represent stochasticity but does model phenomenologically the pheromone memory effect. These models do not seem to predict bistability for small colony sizes, but instead predict that large colonies are more effective at exploiting a single source.

The following individual level model of ant foraging consists of ants foraging from two food souces labeled 1 and 2. Ants foraging from food source one are $x_1$ and those foraging from food source are $x_2$. The pheromones produced by ants $x_1$ are called $a_1$ and those produced by $x_2$ are labelled $a_2$.

$$\begin{aligned} x_1 + a_2 &\xrightarrow{r} x_2 + a_2, & x_1 &\xrightarrow{b} x_1 + a_1, & a_1 &\xrightarrow{d} \emptyset \\ x_2 + a_1 &\xrightarrow{r} x_1 + a_1, & x_2 &\xrightarrow{b} x_2 + a_2, & a_2 &\xrightarrow{d} \emptyset \\ & & x_1 &\xleftrightarrow{\varepsilon} x_2 \end{aligned} \tag{7.4}$$

The parameter $\varepsilon$ represents the small rate at which ants spontaneously start to forage from the other source and $N$ is the total number of ants foraging. The rate $r$ is the rate at which pheromones recruit ants from the other source, $b$ is the rate at which pheromones are produced by ants and $d$ is the rate at which pheromones evaporate. This set of reactions

produces the following transition rates:

$$
\begin{aligned}
T_1(x_1+\frac{1}{N},x_2-\frac{1}{N},a_1,a_2|x_1,x_2,a_1,a_2) &= ra_1x_2+\varepsilon x_2 \qquad & T_3(x_1,x_2,a_1+\frac{1}{N},a_2|x_1,x_2,a_1,a_2) &= bx_1 \\
T_2(x_1-\frac{1}{N},x_2+\frac{1}{N},a_1,a_2|x_1,x_2,a_1,a_2) &= ra_2x_1+\varepsilon x_1 \qquad & T_4(x_1,x_2,a_1,a_2+\frac{1}{N}|x_1,x_2,a_1,a_2) &= bx_2 \\
T_5(x_1,x_2,a_1-\frac{1}{N},a_2|x_1,x_2,a_1,a_2) &= da_1 \qquad & T_6(x_1,x_2,a_1,a_2-\frac{1}{N}|x_1,x_2,a_1,a_2) &= da_2
\end{aligned} \tag{7.5}
$$

Using these transition rates the general Master Equation can be written using raising and lowering operators as

$$
\begin{aligned}
\partial_t P &= [(\varepsilon_{x_1}^-\varepsilon_{x_2}^+-1)T_1+(\varepsilon_{x_1}^+\varepsilon_{x_2}^--1)T_2+(\varepsilon_{a_1}^--1)T_3+(\varepsilon_{a_2}^--1)T_4+(\varepsilon_{a_1}^+-1)T_5+(\varepsilon_{a_2}^+-1)T_6]P \\
&\approx \frac{1}{N}[-\partial_{x_1}(T_1-T_2)-\partial_{x_2}(T_2-T_1)-\partial_{a_1}(T_5-T_3)-\partial_{a_2}(T_6-T_4) \\
&\qquad +\frac{1}{2N^2}(\partial_{x_1}-\partial_{x_2})^2(T_1+T_2)+\frac{1}{2N^2}\partial_{a_1}^2(T_2+T_5)+\frac{1}{2N^2}\partial_{a_2}^2(T_4+T_6)]P,
\end{aligned} \tag{7.6}
$$

where the raising and lowering operators are $\varepsilon_x^\pm f(x)=f(x\pm\frac{1}{N})\approx(1\pm\frac{1}{N}\partial_x+\frac{1}{2N^2}\partial_x^2+...)f(x)$. Using the Taylor expansion in system size and dropping terms $O(\frac{1}{N^3})$ and higher, yields a Fokker-Planck equation. We can rescale time to $t/N\to t$ and obtain the Fokker-Planck equation in the form

$$
\partial_t P(\mathbf{x},t)=[-\partial_{x_i}A_i+\frac{1}{2N}\partial_{x_i}\partial_{x_j}B_{ij}]P(\mathbf{x},t) \tag{7.7}
$$

with

$$
A_i=\begin{bmatrix} T_1-T_2 \\ T_2-T_1 \\ T_3-T_5 \\ T_4-T_6 \end{bmatrix}
$$

$$
B_{ij}=\begin{bmatrix} (T_1+T_2) & -(T_1+T_2) & 0 & 0 \\ -(T_1+T_2) & (T_1+T_2) & 0 & 0 \\ 0 & 0 & T_3+T_5 & 0 \\ 0 & 0 & 0 & T_4+T_6 \end{bmatrix}.
$$

The corresponding Langevin equations (with zero mean noise) are $\partial_t x_i=A_i+\frac{1}{\sqrt{N}}\xi_i$ where $\langle\xi_i(t)\xi_j(t')\rangle=B_{ij}\delta(t-$

$t'$). The noise can be decoupled by making the transformation $\xi_i = G_{ij}\eta_j$, where $\mathbf{G}\mathbf{G}^T = \mathbf{B}$. We choose

$$G_{ij} = \begin{bmatrix} -\sqrt{\frac{T_1+T_2}{2}} & \sqrt{\frac{T_1+T_2}{2}} & 0 & 0 \\ \sqrt{\frac{T_1+T_2}{2}} & -\sqrt{\frac{T_1+T_2}{2}} & 0 & 0 \\ 0 & 0 & \sqrt{T_3+T_5} & 0 \\ 0 & 0 & 0 & \sqrt{T_4+T_6} \end{bmatrix}.$$

so that $\langle \eta(t)\eta^T(t')\rangle = \mathbf{G}^{-1}\langle \xi(t)\xi^T(t')\rangle(\mathbf{G}^{-1})^T = \mathbf{G}^{-1}\mathbf{B}\delta(t-t')(\mathbf{G}^{-1})^T = \mathbf{G}^{-1}\mathbf{G}\mathbf{G}^T(\mathbf{G}^{-1})^T\delta(t-t') = I\delta(t-t')$ showing that the noise is now delta correlated.

After decoupling the noise and making the change of variables $w = x_1 + x_2$, $z = x_1 - x_2$, $c_1 = a_1 + a_2$, and $c_2 = a_1 - a_2$ we rescale time so that $d \times t \to t$. To simplify the equations we can use the Gaussian sum rule for white noise to obtain:

$$\begin{aligned} \partial_t w &= 0 & \partial_t c_1 &= \frac{b}{d}w - c_1 + \sqrt{\frac{\frac{bw}{d}+c_1}{N}}\eta_{c_1} \\ \partial_t z &= \frac{r}{d}(c_2 w - c_1 z) - \frac{2\varepsilon z}{d} + \sqrt{\frac{r(c_1 w - c_2 z) + 2\varepsilon w}{dN/2}}\eta_z & \partial_t c_2 &= \frac{b}{d}z - c_2 + \sqrt{\frac{\frac{bw}{d}+c_1}{N}}\eta_{c_2} \end{aligned} \tag{7.8}$$

These equations show that total number of ants $w$ is conserved. From the equation for $c_1$, the total amount of pheromones decouples from the rest of the system. The stationary probability distribution for $c_1$ is obtained by solving the corresponding Fokker-Planck equation with $\partial_t P(c_1,t) = 0$ and a zero probability current. We obtain

$$P(c_1) = C_0 e^{-2N(K+c_1)}(K+c_1)^{4KN-1} \tag{7.9}$$

where $C_0$ is the normalization constant and $K = \frac{b}{d}w$. Note that the peak of this probability distribution is given by $c_1 = K - \frac{1}{2N}$. This exact result can be intuitively understood by comparing it to the outcome of a linear noise approximation, where we obtain a Gaussian centered around $K$ with variance $\frac{K}{N}$.

To solve for the stationary distribution for $z$, we assume that the noise for $c_1$ and $c_2$ is small, so that their corresponding equations are deterministic. Additionally let us assume that the dynamics for $c_1$ and $c_2$ are sufficiently fast compared to $z$ that they can be approximated by their fixed points $c_1 = \frac{bw}{d}$ and $c_2 = \frac{b}{d}z$. This leads to

$$\partial_t z = -\frac{2\varepsilon z}{d} + \sqrt{\frac{\frac{b}{d}r(w^2 - z^2) + 2\varepsilon w}{dN/2}}\eta_z. \tag{7.10}$$

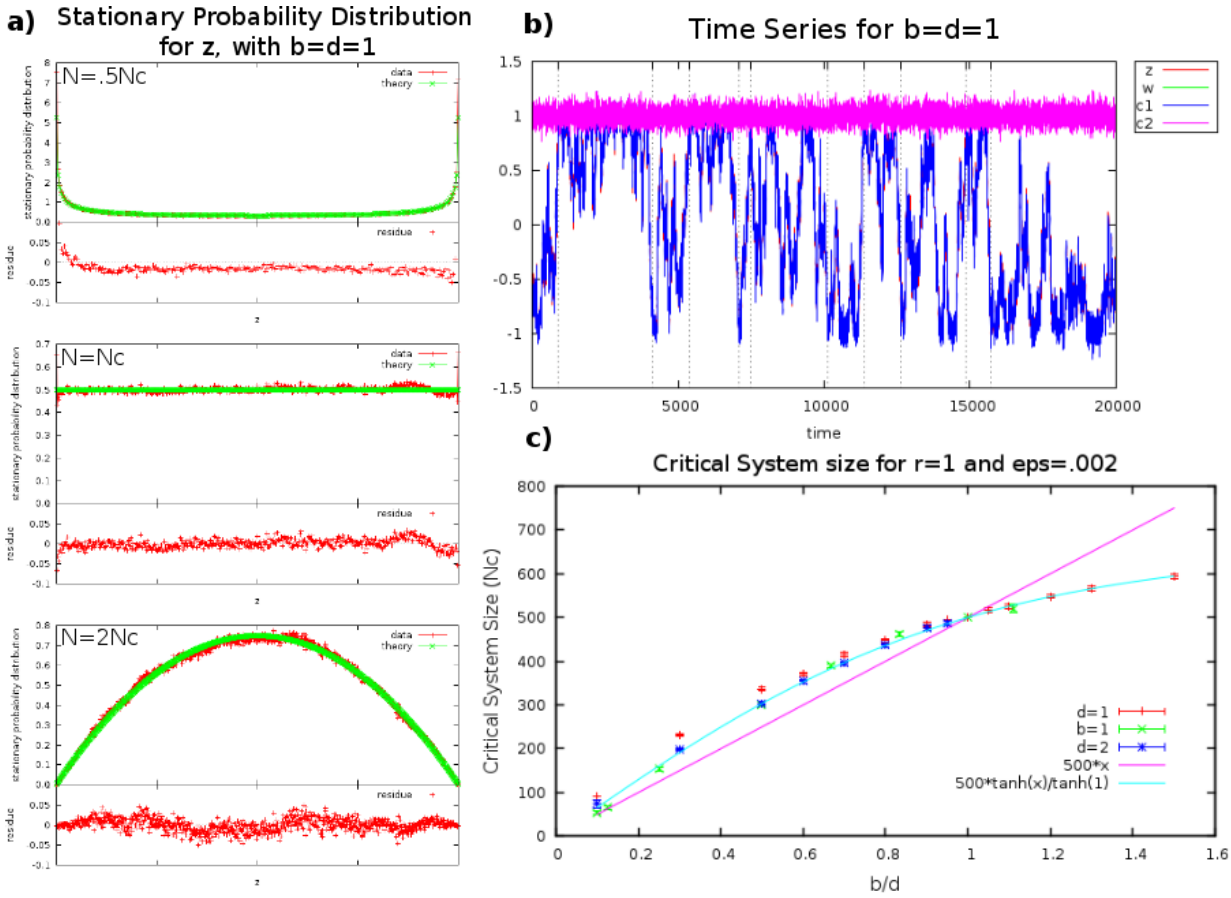


Figure 7.3: Comparison of Gillespie simulations to theory for $\varepsilon = .002$, $r = 1$, and $b = d = 1$. Also plotted is the measured critical system size and an example time series of a simulation.

This equation can be solved for its stationary distribution in the same way as before and has a solution:

$$P(z) = \frac{C_{z0}}{(w^2 - z^2 + 2\frac{\varepsilon w d}{rb})^{1-\frac{N}{N_c}}} \tag{7.11}$$

where $N_c = br/d\varepsilon$. From the distribution for $z$, we obtain the stationary probability distribution for $c_2$, by relaxing our criterion that $c_2$ is deterministic and using the linear noise approximation: $c_2 = \frac{b}{d}z + \sqrt{\frac{2\frac{bw}{d}}{N}}\eta_{c_2}$. This implies that the probability distribution for $c_2$ is the probability distribution for $z$ convolved with a Gaussian.

$$P(c_2) = \int_{-\infty}^{\infty} P_z(z)\frac{1}{\sqrt{2\pi K/N}} e^{\frac{-\left(c_2 - \frac{b}{d}z\right)^2}{2K/N}} dz = \int_{-1}^{1} P_z(z)\frac{1}{\sqrt{2\pi K/N}} e^{\frac{-\left(c_2 - \frac{b}{d}z\right)^2}{2K/N}} dz \tag{7.12}$$

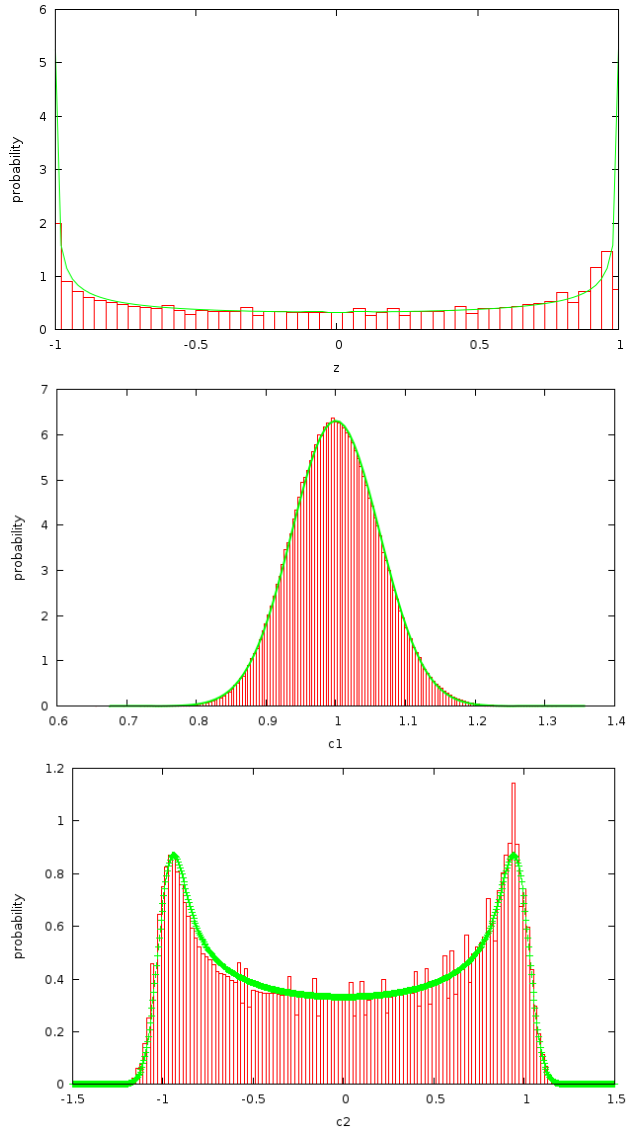


Figure 7.4: Comparison of Gillespie simulations to theory for $\varepsilon = .002$, $r = 1$, and $b = d = 1$ for $z$, $c_1$ and $c_2$.

## 7.3 Simulations and Discussion

The results of these approximate analytic calculations were tested by performing a Gillespie simulation. From the simulation the stationary probability distribution and switching times of the ant population were measured for various parameters. I found that when $b/d = 1$ the simulation and the analytical formula agree remarkably well (See Fig. 7.4). For other values of $b/d$ there are systematic deviations as shown in Figure 7.3c. The critical system size as measured from the simulations differs from that predicted by the approximate theory. But the functional form of the stationary probability distribution is accurate as long as the measured critical system size $N_c$ is used in equation 7.11. This is probably due to the simplifying assumptions made, for example, making the pheromone deterministic and assuming

that the pheromone instantly goes to its fixed point.

As can be seen from the simulation and the analytical theory, the critical system size depends on both the rate of pheromone deposition and evaporation. Specifically, we found that the critical system size depends on the ratio of the deposition rate to the evaporation rate, $b/d$. Analytically, we found the critical system size should behave as $br/d\varepsilon$. From simulation, however, it appears that the critical system size behaves more as $(r/\varepsilon)\cdot(\tanh(b/d)/\tanh(1))$. These predictions for the critical system size potentially give experimentalists more opportunities to test multiplicative noise theories. Biancalani's theory is problematic to study experimentally because the number of ants foraging is not controllable by experimentalists. In my new theory, the evaporation rate and deposition rate of the pheromone could be controlled to effectively change the critical system size and observe the switch from bistable foraging to foraging equally between two food sources. One could imagine controlling the evaporation rate of the pheromone by the type of material the ants are allowed to forage on in addition to blowing a fan or heating the surface. This suggests that this model could be ripe for experimental testing.

## 7.4 Extensions

There are various possible extensions to this model that are of biological and experimental significance. The first extension is considering this model with asymmetric rates. By considering asymmetric rates we can answer questions such as what happens when ants forage from food sources of different quality. For example, in experiments by Beckers et al., ants were allowed to forage from two food sources, one which had 1M sugar water and the other with 0.1M sugar water. There is evidence that ants will put down more pheromones on trails corresponding to higher quality food [152]. Additionally, asymmetric rates will allow us to model food sources placed at unequal distances from the ant colony.

The ant model can also be extended spatially. We can model the spatial distribution of pheromone and ants. This will allow us to study additional effects such as when the pheromone trail evaporates faster than the ants can place it. Experiments indicate the existence of a phase transition as a function of number of foragers and a transition between disordered foraging and ordered foraging [153].

# References


[1] Tommaso Biancalani, Louise Dyson, and Alan J McKane. Noise-induced bistable states and their mean switching time in foraging colonies. *Physical review letters*, 112(3):038101, 2014.

[2] A.M. Turing. The chemical basis of morphogenesis. *Philos. Trans. R. Soc. London, Ser. B*, 237(641):37–72, 1952.

[3] Thomas Butler and Nigel Goldenfeld. Robust ecological pattern formation induced by demographic noise. *Phys. Rev. E*, 80(3):030902, 2009.

[4] Tommaso Biancalani, Duccio Fanelli, and Francesca Di Patti. Stochastic turing patterns in the brusselator model. *Physical Review E*, 81(4):046215, 2010.

[5] Thomas Butler and Nigel Goldenfeld. Fluctuation-driven turing patterns. *Physical Review E*, 84(1):011112, 2011.

[6] D. Schneider and R. E. Lenski. Dynamics of insertion sequence elements during experimental evolution of bacteria. *Res Microbiol*, 155(5):319–27, 2004.

[7] Bao Ton-Hoang, Ccile Pasternak, Patricia Siguier, Catherine Guynet, Alison Burgess Hickman, Fred Dyda, Suzanne Sommer, and Michael Chandler. Single-stranded dna transposition is coupled to host replication. *Cell*, 142(3):398–408, 2010.

[8] Arnaud Le Rouzic, Thibaut Payen, and Aurlie Hua-Van. Reconstructing the evolutionary history of transposable elements. *Genome Biology and Evolution*, 5(1):77–86, 2013.

[9] S. Kohl and R. Bock. Transposition of a bacterial insertion sequence in chloroplasts. *Plant J*, 58(3):423–36, 2009.

[10] C. Parisod, C. Mhiri, K. Y. Lim, J. J. Clarkson, M. W. Chase, A. R. Leitch, and M. A. Grandbastien. Differential dynamics of transposable elements during long-term diploidization of nicotiana section repandae (solanaceae) allopolyploid genomes. *PLoS One*, 7(11):e50352, 2012.

[11] Olga Novikova and Marlene Belfort. Mobile group ii introns as ancestral eukaryotic elements. *Trends in Genetics*, 2017.

[12] William Martin and Eugene V. Koonin. Introns and the origin of nucleuscytosol compartmentalization. *Nature*, 440(7080):41–45, 2006.

[13] Alan M. Lambowitz and Marlene Belfort. Mobile bacterial group ii introns at the crux of eukaryotic evolution. *Microbiol. Spectr.*, 3(1):MDNA3–0050–2014, 2015.

[14] Alan M. Lambowitz and Steven Zimmerly. Group ii introns: Mobile ribozymes that invade dna. *Cold Spring Harbor Perspectives in Biology*, 3(8):a003616, 2011.

[15] N. H. Kim, G. Lee, N. A. Sherer, K. M. Martini, N. Goldenfeld, and T. E. Kuhlman. Real-time transposable element activity in individual live cells. *Proc Natl Acad Sci USA*, 113(26):7278–83, 2016.

[16] Simon A Levin. Hypothesis for origin of planktonic patchiness. *Nature*, 259:659, 1976.

[17] D. T. Gillespie. The average number of generations until fixation of a mutant gene in a finite population. *J Comput Phys*, 22:403–434, 1976.

[18] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[19] A. Gierer and H. Meinhardt. A theory of biological pattern formation. *Biol. Cybern.*, 12(1):30–39, 1972.

[20] JD Murray. A pre-pattern formation mechanism for animal coat markings. *J. Theor. Biol.*, 88(1):161–199, 1981.

[21] William G Wilson, Susan P Harrison, Alan Hastings, and Kevin McCann. Exploring stable pattern formation in models of tussock moth populations. *Journal of animal ecology*, 68(1):94–107, 1999.

[22] V. Castets, E. Dulos, J. Boissonade, and P. De Kepper. Experimental evidence of a sustained standing Turing-type nonequilibrium chemical pattern. *Phys. Rev. Lett.*, 64(24):2953–2956, 1990.

[23] J. Raspopovic, L. Marcon, L. Russo, and J. Sharpe. Digit patterning is controlled by a Bmp-Sox9-Wnt Turing network modulated by morphogen gradients. *Science*, 345(6196):566–570, 2014.

[24] G. Theraulaz, E. Bonabeau, S.C. Nicolis, R.V. Solé, V. Fourcassié, S. Blanco, R. Fournier, J.L. Joly, P. Fernández, A. Grimal, et al. Spatial patterns in ant colonies. *Proc. Natl. Acad. Sci. U.S.A.*, 99(15):9645, 2002.

[25] Han-Sung Jung, Philippa H Francis-West, Randall B Widelitz, Ting-Xin Jiang, Sheree Ting-Berreth, Cheryll Tickle, Lewis Wolpert, and Cheng-Ming Chuong. Local Inhibitory Action of BMPs and Their Relationships with Activators in Feather Formation: Implications for Periodic Patterning. *Developmental Biology*, 196(1):11–23, apr 1998.

[26] Akiko Nakamasu, Go Takahashi, Akio Kanbe, and Shigeru Kondo. Interactions between zebrafish pigment cells responsible for the generation of Turing patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 106(21):8429–34, may 2009.

[27] A.D. Economou, A. Ohazama, T. Pornaveetus, P.T. Sharpe, S. Kondo, M.A. Basson, A. Gritli-Linde, M.T. Cobourne, and J.B.A. Green. Periodic stripe formation by a Turing mechanism operating at growth zones in the mammalian palate. *Nature Genetics*, 44(3):348–351, 2012.

[28] Patrick Müller, Katherine W Rogers, Ben M Jordan, Joon S Lee, Drew Robson, Sharad Ramanathan, and Alexander F Schier. Differential diffusivity of nodal and lefty underlies a reaction-diffusion patterning system. *Science*, 336(6082):721–724, 2012.

[29] Tommaso Biancalani, Farshid Jafarpour, and Nigel Goldenfeld. Giant amplification of noise in fluctuation-induced pattern formation. *Physical Review Letters*, 118:018101, 2017.

[30] Natalie S. Scholes and Mark Isalan. A three-step framework for programming pattern formation. *Current Opinion in Chemical Biology*, 40:1–7, oct 2017.

[31] E. Andrianantoandro, S. Basu, D.K. Karig, and R. Weiss. Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol.*, 2(1), 2006.

[32] S. Basu, Y. Gerchman, C.H. Collins, F.H. Arnold, and R. Weiss. A synthetic multicellular system for programmed pattern formation. *Nature*, 434(7037):1130–1134, 2005.

[33] N.J. Guido, X. Wang, D. Adalsteinsson, D. McMillen, J. Hasty, C.R. Cantor, T.C. Elston, and JJ Collins. A bottom-up approach to gene regulation. *Nature*, 439(7078):856–860, 2006.

[34] A. Levskaya, A.A. Chevalier, J.J. Tabor, Z.B. Simpson, L.A. Lavery, M. Levy, E.A. Davidson, A. Scouras, A.D. Ellington, E.M. Marcotte, et al. Synthetic biology: engineering Escherichia coli to see light. *Nature*, 438(7067):441–442, 2005.

[35] Alexandra M Tayar, Eyal Karzbrun, Vincent Noireaux, and Roy H Bar-Ziv. Propagating gene expression fronts in a one-dimensional coupled system of artificial cells. *Nature Physics*, 11(12):1037–1041, 2015.

[36] J. Stricker, S. Cookson, M.R. Bennett, W.H. Mather, L.S. Tsimring, and J. Hasty. A fast, robust and tunable synthetic gene oscillator. *Nature*, 456(7221):516–519, 2008.

[37] Tal Danino, Octavio Mondragón-Palomino, Lev Tsimring, and Jeff Hasty. A synchronized quorum of genetic clocks. *Nature*, 463(7279):326–330, 2010.

[38] Ye Chen, Jae Kyoung Kim, Andrew J Hirning, Krešimir Josić, and Matthew R Bennett. Emergent genetic oscillations in a synthetic microbial consortium. *Science*, 349(6251):986–989, 2015.

[39] Mehdi Sadeghpour, Alan Veliz-Cuba, Gábor Orosz, Krešimir Josić, and Matthew R. Bennett. Bistability and oscillations in co-repressive synthetic microbial consortia. *Quantitative Biology*, 5(1):55–66, mar 2017.

[40] Jesus Fernandez-Rodriguez, Felix Moser, Miryoung Song, and Christopher A Voigt. Engineering RGB color vision into Escherichia coli. *Nat Chem Biol*, 13(7):706–708, jul 2017.

[41] Takayuki Sohka, Richard A Heins, Ryan M Phelan, Jennifer M Greisler, Craig A Townsend, and Marc Ostermeier. An externally tunable bacterial band-pass filter. *Proceedings of the National Academy of Sciences*, 106(25):10135–10140, 2009.

[42] Chenli Liu, Xiongfei Fu, Lizhong Liu, Xiaojing Ren, Carlos KL Chau, Sihong Li, Lu Xiang, Hualing Zeng, Guanhua Chen, Lei-Han Tang, et al. Sequential establishment of stripe patterns in an expanding cell population. *Science*, 334(6053):238–241, 2011.

[43] E.C. Pesci and B.H. Iglewski. The chain of command in Pseudomonas quorum sensing. *Trends Microbiol.*, 5(4):132–134, 1997.

[44] P.S. Stewart. Diffusion in Biofilms. *J. Bacteriol.*, 185(5):1485–1491, 2003.

[45] J.P. Pearson, C. Van Delden, and B.H. Iglewski. Active efflux and diffusion are involved in transport of Pseudomonas aeruginosa cell-to-cell signals. *Journal of bacteriology*, 181(4):1203–1210, 1999.

[46] E.C. Pesci, J.P. Pearson, P.C. Seed, and B.H. Iglewski. Regulation of las and rhl quorum sensing in Pseudomonas aeruginosa. *J. Bacteriol.*, 179(10):3127–3132, 1997.

[47] K. Brenner, D.K. Karig, R. Weiss, and F.H. Arnold. Engineered bidirectional communication mediates a consensus in a microbial biofilm consortium. *Proc. Natl. Acad. Sci. U.S.A.*, 104(44):17300, 2007.

[48] T.S. Gardner, C.R. Cantor, and J.J. Collins. Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403(6767):339–342, 2000.

[49] P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37:17–33, 1950.

[50] Shigeru Kondo and Takashi Miura. Reaction-diffusion model as a framework for understanding biological pattern formation. *Science (New York, N.Y.)*, 329(5999):1616–20, sep 2010.

[51] Aric Hagberg and Ehud Meron. Pattern formation in non-gradient reaction-diffusion systems: the effects of front bifurcations. *Nonlinearity*, 7(3):805, 1994.

[52] Luciano Marcon, Xavier Diego, James Sharpe, and Patrick Müller. High-throughput mathematical analysis identifies turing networks for patterning with equally diffusing signals. *eLife*, 5:e14022, 2016.

[53] E. A. Gaffney and N. A. M. Monk. Gene expression time delays and Turing pattern formation systems. *Bulletin of mathematical biology*, 68(1):99–130, 2006.

[54] S. Seirin Lee, E. A. Gaffney, and N. A. M. Monk. The Influence of Gene Expression Time Delays on GiererMeinhardt Pattern Formation Systems. *Bulletin of Mathematical Biology*, 72(8):2139–2160, November 2010.

[55] Mads Kærn, Timothy C. Elston, William J. Blake, and James J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, jun 2005.

[56] A. J. McKane and T. J. Newman. Predator-prey cycles from resonant amplification of demographic stochasticity. *Phys. Rev. Lett.*, 94:218102, Jun 2005.

[57] Martin Howard and Andrew D. Rutenberg. Pattern formation inside bacteria: Fluctuations due to the low copy number of proteins. *Phys. Rev. Lett.*, 90:128102, Mar 2003.

[58] Martijn Wehrens, Pieter Rein ten Wolde, and Andrew Mugler. Positive feedback can lead to dynamic nanometer-scale clustering on cell membranes. *The Journal of chemical physics*, 141(20):205102, 2014.

[59] Steven J. Altschuler, Sigurd B. Angenent, Yanqin Wang, and Lani F. Wu. On the spontaneous emergence of cell polarity. *Nature*, 454(7206):886–889, aug 2008.

[60] Y. Cao, D.T. Gillespie, and L.R. Petzold. Efficient step size selection for the tau-leaping simulation method. *J. Chem. Phys.*, 124:044109, 2006.

[61] D. Rossinelli, B. Bayati, and P. Koumoutsakos. Accelerated stochastic and hybrid methods for spatial simulations of reaction–diffusion systems. *Chem. Phys. Lett.*, 451(1-3):136–140, 2008.

[62] Salvador E Luria and Max Delbrück. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491, 1943.

[63] Patricia L Foster. Stress-induced mutagenesis in bacteria. *Critical reviews in biochemistry and molecular biology*, 42(5):373–397, 2007.

[64] Ranjit S Bindra, Meredith E Crosby, and Peter M Glazer. Regulation of dna repair in hypoxic cancer cells. *Cancer and Metastasis Reviews*, 26(2):249–260, 2007.

[65] Reuben S Harris, Gang Feng, Kimberly J Ross, Roger Sidhu, Carl Thulin, Simonne Longerich, Susan K Szigety, Malcolm E Winkler, and Susan M Rosenberg. Mismatch repair protein mutl becomes limiting during stationary-phase mutation. *Genes & development*, 11(18):2426–2437, 1997.

[66] C. Holmes, M. Ghafari, A Anzar, V Saravanan, and I Nemenman. Luria-delbruck, revisited: The classic experiment does not rule out lamarckian evolution. *arXiv*, 2017.

[67] A. P. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet*, 7(12):e1002384, 2011.

[68] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[69] B. McClintock. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A*, 36(6):344–355, 1950.

[70] V. P. Belancio, P. L. Deininger, and A. M. Roy-Engel. Line dancing in the human genome: transposable elements and disease. *Genome Med*, 1(10):97, 2009.

[71] B. Bodega and V. Orlando. Repetitive elements dynamics in cell identity programming, maintenance and disease. *Curr Opin Cell Biol*, 31C:67–73, 2014.

[72] P. A. Callinan and M. A. Batzer. Retrotransposable elements and human disease. *Genome Dyn*, 1:104–15, 2006.

[73] J. M. Chen, P. D. Stenson, D. N. Cooper, and C. Ferec. A systematic analysis of line-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum Genet*, 117(5):411–27, 2005.

[74] P. L. Deininger and M. A. Batzer. Alu repeats and human disease. *Mol Genet Metab*, 67(3):183–93, 1999.

[75] Jr. Kazazian, H. H., C. Wong, H. Youssoufian, A. F. Scott, D. G. Phillips, and S. E. Antonarakis. Haemophilia a resulting from de novo insertion of l1 sequences represents a novel mechanism for mutation in man. *Nature*, 332(6160):164–6, 1988.

[76] K. A. O'Donnell and K. H. Burns. Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mob DNA*, 1(1):21, 2010.

[77] N. G. Coufal, J. L. Garcia-Perez, G. E. Peng, G. W. Yeo, Y. Mu, M. T. Lovci, M. Morell, K. S. O'Shea, J. V. Moran, and F. H. Gage. L1 retrotransposition in human neural progenitor cells. *Nature*, 460(7259):1127–31, 2009.

[78] H. Kano, I. Godoy, C. Courtney, M. R. Vetter, G. L. Gerton, E. M. Ostertag, and Jr. Kazazian, H. H. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev*, 23(11):1303–12, 2009.

[79] L. Chao, C. Vargas, B. B. Spear, and E. C. Cox. Transposable elements as mutator genes in evolution. *Nature*, 303(5918):633–5, 1983.

[80] W. S. Reznikoff. *Transposable Elements*, pages 680–689. Academic Press, Oxford, 2009.

[81] D. A. Petrov, A. S. Fiston-Lavier, M. Lipatov, K. Lenkov, and J. Gonzalez. Population genomics of transposable elements in drosophila melanogaster. *Mol Biol Evol*, 28(5):1633–44, 2011.

[82] Sarah Schaack, Ellen J. Pritham, Abby Wolf, and Michael Lynch. Dna transposon dynamics in populations of daphnia pulex with and without sex. *Proceedings of the Royal Society B: Biological Sciences*, 2010.

[83] J. J. Shen, J. Dushoff, A. J. Bewick, F. J. Chain, and B. J. Evans. Genomic dynamics of transposable elements in the western clawed frog (silurana tropicalis). *Genome Biol Evol*, 5(5):998–1009, 2013.

[84] S. Venner, C. Feschotte, and C. Biemont. Dynamics of transposable elements: towards a community ecology of the genome. *Trends Genet*, 25(7):317–23, 2009.

[85] S. I. Wright, Q. H. Le, D. J. Schoen, and T. E. Bureau. Population dynamics of an ac-like transposable element in self- and cross-pollinating arabidopsis. *Genetics*, 158(3):1279–88, 2001.

[86] D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, and M. Blot. Genomic evolution during a 10,000-generation experiment with bacteria. *Proc Natl Acad Sci U S A*, 96(7):3807–12, 1999.

[87] C. E. Paquin and V. M. Williamson. Temperature effects on the rate of ty transposition. *Science*, 226(4670):53–5, 1984.

[88] J. L. Goodier. Retrotransposition in tumors and brains. *Mob DNA*, 5:11, 2014.

[89] A. Babic, A. B. Lindner, M. Vulic, E. J. Stewart, and M. Radman. Direct visualization of horizontal gene transfer. *Science*, 319(5869):1533–6, 2008.

[90] S. He, A. Corneloup, C. Guynet, L. Lavatine, A. Caumont-Sarcos, P. Siguier, B. Marty, F. Dyda, M. Chandler, and B. Ton Hoang. *The IS200/IS605 Family and Peel and Paste Single-strand Transposition Mechanism*. American Society of Microbiology, 2015.

[91] O. Barabas, D. R. Ronning, C. Guynet, A. B. Hickman, B. Ton-Hoang, M. Chandler, and F. Dyda. Mechanism of is200/is605 family dna transposases: activation and transposon-directed target site selection. *Cell*, 132(2):208–20, 2008.

[92] C. Guynet, A. B. Hickman, O. Barabas, F. Dyda, M. Chandler, and B. Ton-Hoang. In vitro reconstitution of a single-stranded transposition mechanism of is608. *Mol Cell*, 29(3):302–12, 2008.

[93] B. Ton-Hoang, C. Guynet, D. R. Ronning, B. Cointin-Marty, F. Dyda, and M. Chandler. Transposition of ishp608, member of an unusual family of bacterial insertion sequences. *EMBO J*, 24(18):3325–38, 2005.

[94] S. He, A. B. Hickman, F. Dyda, N. P. Johnson, M. Chandler, and B. Ton-Hoang. Reconstitution of a functional is608 single-strand transpososome: role of non-canonical base pairing. *Nucleic Acids Res*, 39(19):8503–12, 2011.

[95] Susu He, Catherine Guynet, Patricia Siguier, Alison B. Hickman, Fred Dyda, Mick Chandler, and Bao Ton-Hoang. Is200/is605 family single-strand transposition: mechanism of is608 strand transfer. *Nucleic Acids Research*, 2013.

[96] R. Lutz and H. Bujard. Independent and tight regulation of transcriptional units in escherichia coli via the lacr/o, the tetr/o and arac/i1-i2 regulatory elements. *Nucleic Acids Res*, 25(6):1203–10, 1997.

[97] M. P. Calos and J. H. Miller. The dna sequence change resulting from the iq1 mutation, which greatly increases promoter strength. *Mol Gen Genet*, 183(3):559–60, 1981.

[98] Michele L. Markwardt, Gert-Jan Kremers, Catherine A. Kraft, Krishanu Ray, Paula J. C. Cranfill, Korey A. Wilson, Richard N. Day, Rebekka M. Wachter, Michael W. Davidson, and Mark A. Rizzo. An improved cerulean fluorescent protein with enhanced brightness and reduced reversible photoswitching. *PLoS ONE*, 6(3):e17896, 2011.

[99] T. Nagai, K. Ibata, E. S. Park, M. Kubota, K. Mikoshiba, and A. Miyawaki. A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nat Biotechnol*, 20(1):87–90, 2002.

[100] N. C. Shaner, R. E. Campbell, P. A. Steinbach, B. N. Giepmans, A. E. Palmer, and R. Y. Tsien. Improved monomeric red, orange and yellow fluorescent proteins derived from discosoma sp. red fluorescent protein. *Nat Biotechnol*, 22(12):1567–72, 2004.

[101] Justin B. Kinney, Anand Murugan, Curtis G. Callan, and Edward C. Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proceedings of the National Academy of Sciences*, 107(20):9158–9163, 2010.

[102] Chitra R Nayak and Andrew D Rutenberg. Quantification of fluorophore copy number from intrinsic fluctuations during fluorescence photobleaching. *Biophysical Journal*, 101(9):2284–2293, 2011.

[103] A. Goni-Merno, M. Amos, and F. de la Cruz. Multicellular computing using conjugation for wiring. *PLoS One*, 8, 2013.

[104] A. Goni-Merno and M. Amos. Discrete modelling of bacterial conjugation. *arXiv*, 2012.

[105] S. E. Luria and M. Delbrck. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, 28(6):491–511, 1943.

[106] David A. Kessler and Herbert Levine. Large population solution of the stochastic luriadelbrck evolution model. *Proceedings of the National Academy of Sciences*, 110(29):11682–11687, 2013.

[107] Taekjip Ha. Single-molecule methods leap ahead. *Nat Meth*, 11(10):1015–1018, 2014.

[108] David J. Galas and Michael Chandler. Structure and stability of tn9-mediated cointegrates: Evidence for two pathways of transposition. *Journal of Molecular Biology*, 154(2):245–272, 1982.

[109] Matthew Scott, Carl W. Gunderson, Eduard M. Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of cell growth and gene expression: Origins and consequences. *Science*, 330(6007):1099–1102, 2010.

[110] Samir Acharya, Patricia L. Foster, Peter Brooks, and Richard Fishel. The coordinated functions of the e. coli muts and mutl proteins in mismatch repair. *Molecular Cell*, 12(1):233–246, 2003.

[111] C. R. Beck, P. Collier, C. Macfarlane, M. Malig, J. M. Kidd, E. E. Eichler, R. M. Badge, and J. V. Moran. Line-1 retrotransposition activity in human genomes. *Cell*, 141(7):1159–70, 2010.

[112] Sandra R. Richardson, Aurlien J. Doucet, Huira C. Kopera, John B. Moldovan, Jos Luis Garcia-Perez, and John V. Moran. The influence of line-1 and sine retrotransposons on mammalian genomes. *Microbiol. Spectr.*, 3(2), 2015.

[113] Sandra L. Martin. The orf1 protein encoded by line-1: structure and function during l1 retrotransposition. *J. of Biomed. and Biotechnol.*, 2006:45621, 2006.

[114] A. J. Doucet, J. E. Wilusz, T. Miyoshi, Y. Liu, and J. V. Moran. A 3 poly(a) tract is required for line-1 retrotransposition. *Mol. Cell*, 60(5):728–741, 2015.

[115] John V. Moran, Susan E. Holmes, Thierry P. Naas, Ralph J. DeBerardinis, Jef D. Boeke, and Haig H. Kazazian Jr. High frequency retrotransposition in cultured mammalian cells. *Cell*, 87(5):917–927, 1996.

[116] Victoria P. Belancio, Dale J. Hedges, and Prescott Deininger. Line-1 rna splicing and influences on mammalian gene expression. *Nucleic Acids Research*, 34(5):1512–1521, 2006.

[117] C. Meischl, M. Boer, A. Ahlin, and D. Roos. A new exon created by intronic insertion of a rearranged line-1 element as the cause of chronic granulomatous disease. *Eur J Hum Genet*, 8(9):697–703, 2000.

[118] Deborah Noack, Paul G. Heyworth, Peter E. Newburger, and Andrew R. Cross. An unusual intronic mutation in the cybb gene giving rise to chronic granulomatous disease. *Biochim Biophys Acta Mol Basis Dis*, 1537(2):125–131, 2001.

[119] A. M. Denli, I. Narvaiza, B. E. Kerman, M. Pena, C. Benner, M. C. N. Marchetto, J. K. Diedrich, A. Aslanian, J. Ma, J. J. Moresco, L. Moore, T. Hunter, A. Saghatelian, and F. H. Gage. Primate-specific orf0 contributes to retrotransposon-mediated diversity. *Cell*, 163(3):583–593.

[120] Jeffrey S. Han. Non-long terminal repeat (non-ltr) retrotransposons: mechanisms, recent developments, and unanswered questions. *Mobile DNA*, 1(1):15, 2010.

[121] Colin J. Coros, Markus Landthaler, Carol Lyn Piazza, Arthur Beauregard, Donna Esposito, Jiri Perutka, Alan M. Lambowitz, and Marlene Belfort. Retrotransposition strategies of the lactococcus lactis ll.ltrb group ii intron are dictated by host identity and cellular environment. *Molecular Microbiology*, 56(2):509–524, 2005.

[122] B. Cousineau, S. Lawrence, D. Smith, and M. Belfort. Retrotransposition of a bacterial group ii intron. *Nature*, 404(6781):1018–21, 2000.

[123] Kenji Ichiyanagi, Arthur Beauregard, Stacey Lawrence, Dorie Smith, Benoit Cousineau, and Marlene Belfort. Retrotransposition of the ll.ltrb group ii intron proceeds predominantly via reverse splicing into dna targets. *Molecular Microbiology*, 46(5):1259–1272, 2002.

[124] F. W. Studier, A. H. Rosenberg, J. J. Dunn, and J. W. Dubendorff. Use of t7 rna polymerase to direct expression of cloned genes. *Methods in enzymology*, 185:60–89, 1990.

[125] T. E. Kuhlman and E. C. Cox. Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Res*, 38(6):e92, 2010.

[126] H. Tas, C. T. Nguyen, R. Patel, N. H. Kim, and T. E. Kuhlman. An integrated system for precise genome modification in escherichia coli. *PLoS One*, 10(9):e0136963, 2015.

[127] H. D. Nguyen, Q. A. Nguyen, R. C. Ferreira, L. C. Ferreira, L. T. Tran, and W. Schumann. Construction of plasmid-based expression vectors for bacillus subtilis exhibiting full structural stability. *Plasmid*, 54(3):241–8, 2005.

[128] Arthur Beauregard, Venkata R. Chalamcharla, Carol Lyn Piazza, Marlene Belfort, and Colin J. Coros. Bipolar localization of the group ii intron ll.ltrb is maintained in escherichia coli deficient in nucleoid condensation, chromosome partitioning and dna replication. *Molecular Microbiology*, 62(3):709–722, 2006.

[129] P Armitage. The statistical theory of bacterial populations subject to mutation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–40, 1952.

[130] P. A. P. Moran. Random processes in genetics. *Math Proc. Cambridge*, 54:60–71, 1958.

[131] H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7:284–304, 1940.

[132] J. E. Moyal. Stochastic processes and statistical physics. *J. R. Stat. Soc. Series B (Methodological)*, 11:150–210, 1949.

[133] A. J. McKane, T. Biancalani, and T. Rogers. Stochastic pattern formation and spontaneous polarisation: the linear noise approximation and beyond. *Bull. Math. Biol.*, 76:895–921, 2014.

[134] M. Kimura and T. Ohta. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61:763–771, 1969.

[135] A. K. Hottes, P. L. Freddolino, A. Khare, Z. N. Donnell, J. C. Liu, and S. Tavazoie. Bacterial adaptation through loss of function. *PLoS Genet*, 9(7):e1003617, 2013.

[136] Ping Wang, Lydia Robert, James Pelletier, Wei Lien Dang, Francois Taddei, Andrew Wright, and Suckjoon Jun. Robust growth of ¡em¿escherichia coli¡/em¿. *Current Biology*, 20(12):1099–1103, 2010.

[137] F. William Studier and Barbara A. Moffatt. Use of bacteriophage t7 rna polymerase to direct selective high-level expression of cloned genes. *Journal of Molecular Biology*, 189(1):113–130, 1986.

[138] R. Bowater and A. J. Doherty. Making ends meet: repairing breaks in bacterial dna by non-homologous end-joining. *PLoS Genet*, 2(2):e8, 2006.

[139] Ralf Moeller, Erko Stackebrandt, Gnther Reitz, Thomas Berger, Petra Rettberg, Aidan J. Doherty, Gerda Horneck, and Wayne L. Nicholson. Role of dna repair by nonhomologous-end joining in bacillus subtilis spore resistance to extreme dryness, mono- and polychromatic uv, and ionizing radiation. *J. Bacteriol.*, 189(8):3306–3311, 2007.

[140] J. H. Miller. *Experiments in Molecular Genetics*. Cold Spring Harbor Laboratories, 1972.

[141] T. E. Kuhlman and E. C. Cox. A place for everything: chromosomal integration of large constructs. *Bioeng Bugs*, 1(4):296–9, 2010.

[142] Shuji Kubo, Maria del Carmen Seleme, Harris S. Soifer, Jos Luis Garcia Perez, John V. Moran, Haig H. Kazazian, and Noriyuki Kasahara. L1 retrotransposition in nondividing and primary human somatic cells. *Proceedings of the National Academy of Sciences*, 103(21):8036–8041, 2006.

[143] Richard N. Day, Cynthia F. Booker, and Ammasi Periasamy. Characterization of an improved donor fluorescent protein for frster resonance energy transfer microscopy. *Journal of biomedical optics*, 13(3):031203–031203, 2008.

[144] M. Belfort, P.S. Chandry, and J. Pedersen-Lane. Genetic delineation of functional components of the group i intron in the phage t4 td gene. *Cold Spring Harbor Symposia on Quantitative Biology*, 52:181–192, 1987.

[145] Crispin Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer, 2010.

[146] Yuichi Togashi and Kunihiko Kaneko. Transitions induced by the discreteness of molecules in a small autocatalytic system. *Physical review letters*, 86(11):2459, 2001.

[147] Jun Ohkubo, Nadav Shnerb, and David A. Kessler. Transition phenomena induced by internal noise and quasi-absorbing state. *Journal of the Physical Society of Japan*, 77(4), 2008.

[148] Tommaso Biancalani, Tim Rogers, and Alan J McKane. Noise-induced metastability in biochemical networks. *Physical Review E*, 86(1):010106, 2012.

[149] R Beckers, Jean-Louis Deneubourg, Simon Goss, and JM Pasteels. Collective decision making through food recruitment. *Insectes sociaux*, 37(3):258–267, 1990.

[150] Claire Detrain and Jean-Louis Deneubourg. Self-organized structures in a superorganism: do ants behave like molecules? *Physics of Life Reviews*, 3(3):162–187, 2006.

[151] Stamatios C Nicolis, Claire Detrain, Didier Demolin, and Jean-Louis Deneubourg. Optimality of collective choices: a stochastic approach. *Bulletin of mathematical biology*, 65(5):795–808, 2003.

[152] Ralph Beckers, Jean-Louis Deneubourg, and Simon Goss. Modulation of trail laying in the antlasius niger (hymenoptera: Formicidae) and its role in the collective selection of a food source. *Journal of Insect Behavior*, 6(6):751–759, 1993.

[153] Madeleine Beekman, David JT Sumpter, and Francis LW Ratnieks. Phase transition between disordered and ordered foraging in pharaoh's ants. *Proceedings of the National Academy of Sciences*, 98(17):9703–9706, 2001.