Uncovering Hidden Dynamics in Living Systems Using Bayesian Statistics and Single-Molecule Microscopy

by

Josh Karslake

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biophysics)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Julie S. Biteen, Chair
Associate Professor Ajit Jogelkar
Associate Professor Lyle Simmons
Associate Professor Sarah Veatch
Assistant Professor Kevin Wood

Josh Karslake

joshkars@umich.edu

ORCID iD: 0000-0002-3818-2888

I’d like to dedicate this Thesis to my family, whose support and encouragement helped me achieve my goals.

ACKNOWLEDGEMENTS

My time in graduate school has been both one of the most difficult and the most fulfilling experiences of my life. The work presented here would not have been possible or nearly as good without the help and support of numerous people. First off I would like to thank my advisor, Professor Julie Biteen, for believing in me and providing me the space to expand and change my project while also giving me the support and guidance to follow through when things weren't always going as planned. I would also like to thank my committee members for their guidance and support throughout my graduate school years, especially Professor Sarah Veatch for all she has done to help me get where I am today.

I would also like to thank all my collaborators on the various projects, their hard work made it so that I could succeed and their dedication inspired me to work hard at these projects as well. I am also indebted to my lab members, both past and present. You all have taught me so much and not just about science. I would especially like to mention Steve Lee, Hannah Tuson and Dave Rowland for their help and guidance and friendship.

Finally, I would like to thank my friends and family. Your support and encouragement has always meant a lot to me and I would not be here without it. Special thanks to my brother Jason and my wife Laura, my best friends and partners for years and without whom I would not have been able to accomplish my goals.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

BCR B-cell receptor

CD chromodomain

CPD Cumulative Probability Distribution

CSD chromoshadow domain

CTX cholera toxin

DP Dirichlet Process

DPMM Dirichlet process mixture model

(f)PALM (fluorescence) Photoactivated Localization Microscopy

EMCCD electron multiplying charge-coupled device

FP fluorescent protein

GFP Green Fluorescent Protein

MCMC Markov Chain Monte Carlo

MSD Mean-Squared Displacement

PSF Point Spread Function

SMAUG Single-Molecule Analysis by Unsupervised Gibbs

SPT single-particle tracking

STORM Stochastic Optical Reconstruction Microscopy

ABSTRACT

Fluorescence microscopy is a powerful technique for understanding the organization, structure and dynamics of cells. Single-molecule imaging techniques extend our ability to probe cellular systems down to a range of only a few tens of nanometers. Observing the motion of single molecules inside living cells and tracking their behavior can give insight into the native biochemical and biophysical environment of the molecule. If certain conditions, such as the cell being in equilibrium, are met, we can relate the motion observed to the functional role of the molecule. However, biological systems are complex and single-molecule data can be noisy, so care must be taken when analyzing single-particle tracking data sets such that supervisory biases and other external constraints are not placed on the analysis. There is a wealth of information hidden inside tracking data sets that careful analysis can uncover, leading to more concrete conclusions and informing further investigations.

In this Thesis, I present my work on expanding the scope and quality of single-particle tracking analysis and, using this new method, present my investigations of the dynamics involved in several complex biological questions in both prokaryotes and eukaryotes. Traditional curve-fitting analysis methods for single-particle tracking data require supervisory input and can suffer from parameter identifiability issues. Chapter II proposes a new analysis method for single-particle tracking data based on a nonparametric Bayesian statistical framework that we call Single-Molecule Analysis by Unsupervised Gibbs (SMAUG). The accuracy and precision of this method, as well as its ability to uncover the true dynamics, is investigated using realistic simulations and *in vitro* experimen-

tal systems. This new method increases the information available from tracking experiments while not sacrificing accuracy or precision, thus allowing for more rigorous conclusions. In addition, this method is also applied to *in vivo* data from two relevant biological systems and the analysis identifies potential biological roles for the uncovered diffusive states. Chapter 2 demonstrates a method for improving the scope of single-molecule analysis by introducing a new analysis framework that increases the information available.

Differential gene expression patterns are the basis of cellular biology. The markers that modulate which genes are active and which are silenced are called epigenetic markers. The dynamics involved in establishing, maintaining and removing these markers are not well understood. In Chapter III, I use single-particle tracking and the SMAUG algorithm to investigate the dynamics behind epigenetic silencing in a fission yeast model system and uncover the hidden complexity of the system. I present the dynamics uncovered for the key protein in the pathway, Swi6, in otherwise wild-type cells. This measurement resolves four distinct biochemical states. Then, using targeted mutation studies, I investigate and assign a biological role to each of the four identified states, and I uncover the impact that DNA compaction has upon the system. Overall, my application of single-particle tracking and SMAUG analysis to this system provides an example of expanding the scope of single-particle imaging techniques to complex systems and using the information obtained to gain biological insight.

Bacterial virulence is a complex pathway that requires precise timing and organization of the proteins involved to effect a response. In Chapter IV, I present my investigations deeper into the dynamics of *Vibrio cholerae* bacterial virulence that was started in Chapter II. Using single-particle tracking and SMAUG analysis I found three distinct biological states for the keystone protein TcpP in otherwise wild-type cells. Based on mutations of the protein sequence and other regions of the bacterial DNA, I present the biological roles for the states uncovered and discuss future investigations into the system using more mutation studies. The work in this chapter expands the scope of single-molecule imaging experiments by uncovering new information that can lead further developments against bacterial virulence. The work presented in this Thesis will have broad impact on the fields of biophysics and cell biology by both expanding the scope and quality of the information gathered of single-particle tracking experiments as well as by answering specific questions about the dynamics of several biological pathways.

CHAPTER I

Introduction

In this Thesis, I expand both the quantity and the quality of information that can be gained from single-particle tracking (SPT) experiments by designing a new analysis method free from supervisory biases and weak parameters. I then apply this method to several outstanding questions in microbiology that require a deeper mathematical foundation to answer. However, to fully grasp the details of this Thesis, I will use this chapter to explain the necessary background information that will be needed in all other sections of this work. As the work presented here is heavily reliant on super-resolution fluorescent imaging techniques, I will begin by discussing what fluorescence imaging is, and its drawbacks, before moving onto how these limitations are overcome in super-resolution imaging and how the information gathered from these experiments are typically analyzed before finally providing an outline for the subsequent chapters of the Thesis.

1.1 Fluorescence Microscopy

Cells are extremely small; bacterial cells are on the order of 1 to 3 microns in length and eukaryotic cells are generally about ten times larger. Therefore, in order to study these organisms, the ability to see and image them in high contrast and high resolution is needed. Light microscopy has thus been an invaluable tool for cell biology. Unlike other techniques with enough magnification to see inside cells, such as electron microscopy where the cells must be fixed and extensive sample preparation used, light microscopy can be performed on cells while they are still alive and in a minimally invasive manner. Live-cell light microscopy has been extremely useful for hundreds of years in efforts such as histology or bacterial cell sorting, both of which are based on the size and shape of cells [1].

However, light microscopy is not without limitations. The 400-750 nm wavelength range of visible light, as well as the low absorbance cross section of most biological samples, hinder light microscopy. While light microscopy is useful to determine size and shape based on cellular boundaries, it is less useful in determining internal structure, especially if the internal structure is small or low contrast. These limitations are also especially true for bacteria which are both very thin and rather transparent. In fact, despite bacteria having been imaged with light microscopy for hundreds of years, it was not until the early 1990s, with the publication of papers from Pogliano *et al.* [2] and Maddock and Shapiro [3], that an internal structure for bacteria became widely accepted (Fig. 1.1).

The technique of fluorescence microscopy has been used to overcome these limitations and has spurred investigations into the organization and dynamics of specific biomolecules. This technique utilizes the phenomenon of fluorescence, where a fluorophore absorbs a photon of a specific range of wavelengths (excitation spectrum) and then emits a photon at a longer wavelength (emission spectrum) [4]. More specifically, a molecule may absorb a photon of incident light and become excited and transition to a higher energy state. That molecule will then lose some energy through vibrational relaxation, called internal conversion, to the lowest energy state of S_1, a thermally equilibrated excited state. In order to return to the ground state it may eject a photon of its own, but due to the loss of energy through the vibrational modes that emitted photon will be of lower energy and longer wavelength than the incoming one, a phenomena known as the Stokes' Shift (Fig. 1.2A). By using dichroic filters that are reflective to the incident light and are transparent to the emitted light, we can collect only the light that has been emitted from

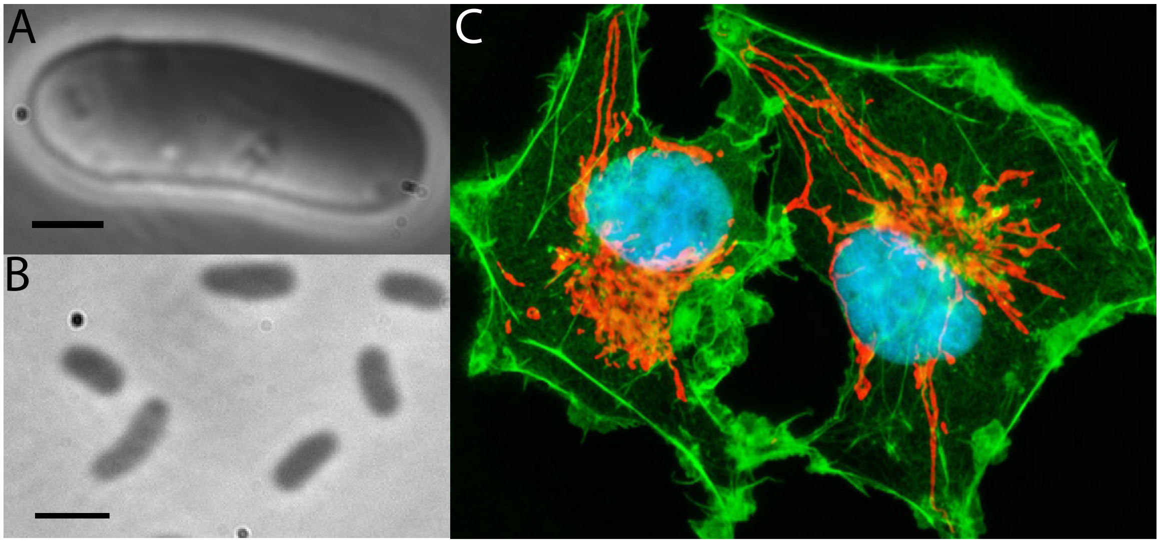


Figure 1.1: Optical Microscopy. Optical microscopy allows direct imaging of living cells but can hide inner structure. **A**: Phase-contrast image of the fission yeast *Schizosaccharomyces pombe*. **B**: Phase-contrast image of the bacteria *Vibrio cholerae*. **C**: Fluorescence microscopy provides better contrast and the ability to label specific cellular components, improving the ability to study cellular behavior. COS-1 cells stained for mitochondria (red), actin (green) and the nucleus (blue). Image taken from the Nikon MicroscopyU website at https://www.microscopyu.com/gallery-images/transformed-simian-virus-40-african-green-monkey-kidney-fibroblast-cells-cos-1-line-1. Scale bars: 2 µm

the sample, improving the contrast of the image. Originally many inorganic dyes were used to label specific parts of a cell such as the DNA, the cytoskeleton or the membrane, but with the discovery of Green Fluorescent Protein (GFP) [5] and concurrent advances in molecular biology [6], a vast majority of proteins of interest could be attached to GFP and become visible under proper illumination while the cell remains alive and active. Encoding a fluorescent protein (FP) label into the genome has the added benefit of maintaining the native expressive levels and control of the protein. These advances have led fluorescence microscopy to become an indispensable tool of cellular biology and an important aspect of many studies on proteins and other biomolecules [7]. Since that discovery, a vast array of FPs with differing properties have been characterized, allowing for multi-color labeling experiments on interactions between different cellular structures/proteins providing a more detailed picture into the complexity of cellular organization and processes [8].

1.2 Super-Resolution Fluorescence Imaging

While fluorescence microscopy has been incredibly useful for helping to understand the subcellular organization of cells, it too has limitations, chiefly that it cannot provide information on any structure that is smaller than the diffraction limit of the emitted light. The diffraction limit is the result of the fact that light has a wavelength. Thus an infinitesimally small emitter will appear larger when observed (Fig. 1.2B). This result was originally discovered in the field of astronomy in the 1800s when Airy described stars appearing not as points but as bright disks inside a telescope [9]. Many biological processes, such as DNA replication, protein transcription or extra-cellular signalling, occur at length scales well below this diffraction limit and thus remained hidden from investigation by microscopy for decades. In addition bulk fluorescent measurements can mask the fast, rare or transient interactions/events of biological processes behind ensemble averaging [10].

Today we describe this disk as the Point Spread Function (PSF) of the emitter, the result of a mathematical convolution of the light from a point source with the lenses and mirrors of a microscope. Two PSFs are considered resolveable if their separation, *d*, is greater than that given by the equation:

$$d = \frac{\lambda}{2NA} \tag{1.1}$$

where λ is the is the wavelength of the emission light and *NA* is the numerical aperture of the objective. Most FPs and dyes work in the visible range of light (400-750 nm) and even with modern oil-immersed objectives that are approaching the theoretical limit of NA = 1.5 the limits of a fluorescence image using visible light to a resolution in the range of 150-250 nm. For example, using the most commonly used FP from my lab, PAmCherry [11], which has an emission peak at 610 nm, and using our main objective, which has an NA of 1.4, our resolution limit is roughly 220 nm in a bulk experiment.

Beating the diffraction limit was the result of several complementary events that came together at roughly the same time. First, improvements in instrumentation allowed researchers to reliably detect the emission from a single fluorophore [12, 13], experiments which laid the groundwork that single-molecule detection was feasible. Further progression in instrumentation allows for single-molecule detection to be performed at room temperature today using widely available experimental setups.

The signal PSF collected from a single molecule isolated in time and space is extremely well approximated by a Gaussian function (Fig. 1.2 C). By fitting the intensity profile with this function the center position of the emitter can be reliably determined to a precision much smaller than the diameter of the PSF. Many studies can reliably reach resolutions of ~20-30 nm, roughly an order of magnitude improvement over bulk methods (Fig. 1.2 D) [14]. However, this fitting method requires isolated emitters. Emitter spacing must be such that any two PSFs do not overlap in time and space (a "sparse" density), a technical detail which requires attention to sample preparation or other methods to achieve. Many systems with fluorescently labeled samples, such as proteins tagged with genetically encoded FPs, could contain anywhere from dozens to thousands of emitters (or more!) depending on the number of native protein copies in the cell and the organism [15]. This resulting high density of emitters has extensive PSF overlap, limiting the ability of this fitting method to be used in biological samples, though some studies were able to label sparsely enough to take advantage of it [16, 17].

The second improvement was the development of methods that allow for isolating single emitters in time and space with the discovery of photoconvertible emitters. Photoconvertible fluorophores, both dyes and FPs, begin in either a dark (non-fluorescing) state or in a state whose fluoroesence emission spectrum is not in a range of interest and can be removed through the use of filters and thus appears dark on the camera. Upon illumination at the activation wavelength, usually ~400 nm, a subset of fluorophores will stochastically convert into a state in which visible photons can be absorbed (Fig. 1.3A). With proper tuning of the activation laser with intensity and duration, the subset of converted emitters can be sparse enough to not overlap and the molecules' emission patterns

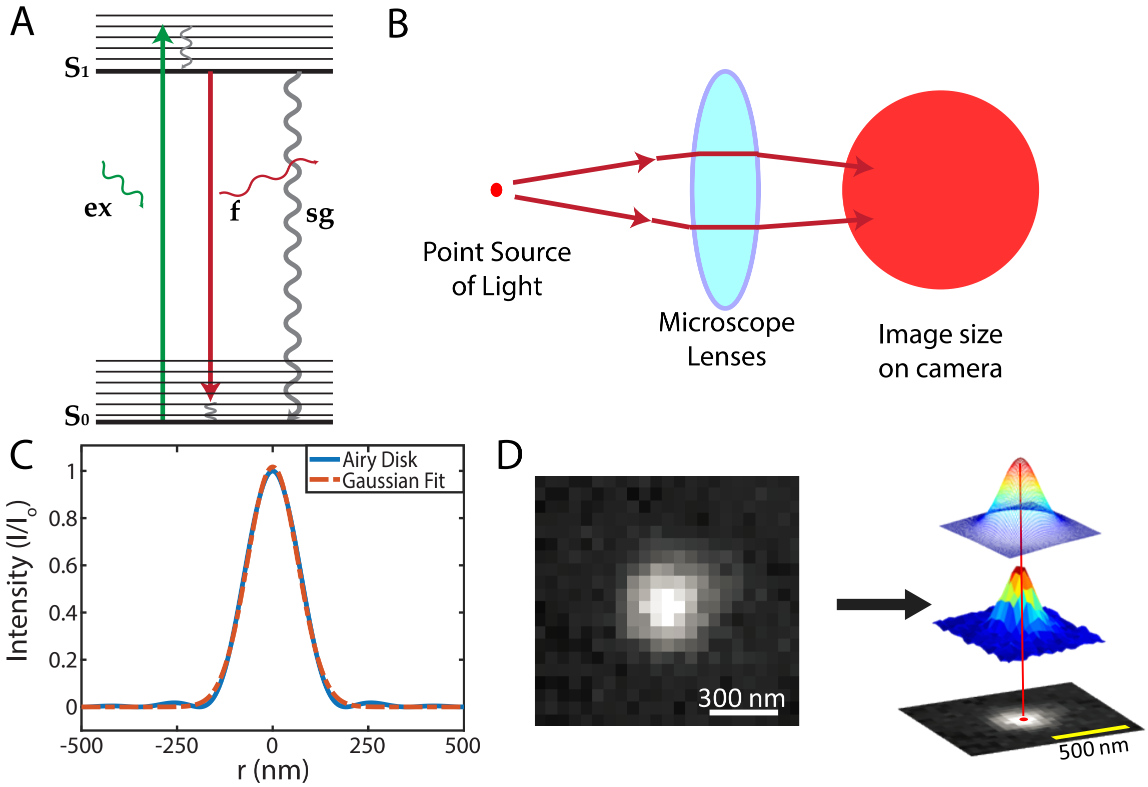


Figure 1.2: Fluorescence Microscopy and PSF Fitting. A: Simplified Jablonbski Diagram. S_0 is the ground state and S_1 is a singlet excited state. An incident exciting photon elevates the system to an excited state (ex). After some vibrational loss the excited electron returns to the ground state and the molecule can either emit a photon (f) or can decay nonradiatively (sg). An emitted photon will have less energy then the incoming photon had and so will be red-shifted. **B**: Due to light having a wavelength and interactions inside the microscope any point source will appear as a disk of light (PSF) on the camera, limiting the resolution of structures smaller than the PSF size. **C**: Using knowledge about the shape of the PSF allow for more precise localization of the emitter. The Airy function is well approximated by a Gaussian, here shown in 1 dimension. **D**: *left* The raw image of an isolated fluorophore as a disk of light. *right* The intensity profile is fit to a 2D Gaussian to more precisely locate the center. Figure A adapted from Ben Isaacoff, Figure C courtesy of Steve Lee and Figure D adapted from Tuson *et al.* [14]

can then be fit to a 2D Gaussian and their centers localized much more precisely. Then repeated rounds of activation and localization occur until enough molecules have been localized (Fig. 1.3B). This method using FPs has been named (fluorescence) Photoactivated Localization Microscopy ((f)PALM) [18, 19] while a similar method using dyes is called Stochastic Optical Reconstruction Microscopy (STORM) [20]. These methods effectively reduce the labeling density of emitting fluorophores enough that their PSFs can be fit and the centers localized much more precisely. Fig. 1.3 has a schematic of the experimental setup used in these types of experiments.

1.3 Single-Particle Tracking

Electron multiplying charge-coupled device (EMCCD) cameras are capable of achieving single-molecule detection while using aquistion times in the tens of milliseconds. As a result, even FPs or dyes that are dim (i.e. fluorophores with poor quantum efficiency) or that have short lifetimes are capable of being captured for several frames in a row before photobleaching. Connecting those localizations together into trajectories that describe the motion of the fluorophore is called single-particle tracking (SPT). SPT is a powerful tool of super-resolution microscopy as the dynamics of a molecule are directly related to the biochemical and biophysical state of that molecule. The most common descriptor of the dynamics of a molecule, or a set of similar molecules, is the diffusion coefficient as changes in the environment of a molecule directly change the values encoded in the diffusion coefficient. Molecules that are bound or confined behave differently than those that are free or actively trafficked [7]. SPT has the necessary spatial and temporal resolution to observe these events as it provides nanometer-scale resolution at millisecond-scale time. These lengths and times are the relevant scales on which much of cellular biology occur. In the past decade, SPT measurements have uncovered the positioning and interactions inherent to many different biological systems, from membrane proteins and lipids to transcription factors and DNA replication machinery, among others [22, 23].

There are many methods for linking a collection of localizations together through time to get trajectories [10]. Our lab uses an energy minimization approach in which all the localizations are put into a matrix and connections are made such that the energy of the matrix is minimized according to the Hungarian algorithm [24] using an exponential merit function that penalizes connections for spatial and temporal distance and intensity mismatch [25].

One method to determine diffusion coefficients from SPT data has been that, once these trajectories were constructed, the squared distances between all localizations in the trajectory is calculated as a function of increasing time lag between the localizations ("steps") [26,27]. The mean of these squared distances for each of the time lags is plotted to create a Mean-Squared Displacement (MSD) plot. Next, a linear fit to the data is performed by a non-linear least-squares method (Fig. 1.4 C) [28, 29], either fitting the MSDs of individual trajectories (i.e. fitting each grey line in Fig. 1.4C and taking the average of those fits) or by fitting the mean of all MSDs (red line in Fig. 1.4C). The slope of the fit line is directly proportional to the diffusion coefficient of the system and is an easy and intuitive measure of a system. However, the MSD method assumes a homogeneous system (or, at least, that each individual trajectory is representative of a single state) and returns a single parameter value for the slope of the fit line. This single value for diffusion is unlikely to be sufficient for many biological systems as molecules have many different functions and might, for example, switch between a bound and an unbound state which will have different dynamics poorly described by a single diffusion value. Some more recent efforts have been made to extend MSD fitting to include multiple diffusive states [30] but these methods remain reliant on user supervision, which can lead to biases and other issues.

To accommodate more complex and heterogeneous systems, a method of fitting the Cumulative Probability Distribution (CPD) of squared-step sizes was developed [31, 32]. Similar to the MSD method, collections of squared step sizes is calculated as a function of increasing time lag (Fig. 1.4D). Unlike the MSD method, the CPD method is a two step curve-fitting method that first fits the collection of squared step sizes directly using an expandable function such as the one below:

$$CPD(x^2,\tau) = 1 - \alpha_1 e^{\frac{-x^2}{MSD_{1,\tau}}} - \ldots - (1 - \sum_{i=1}^{K} \alpha_i) e^{\frac{-x^2}{MSD_{K,\tau}}} \tag{1.2}$$

where x^2 is the squared distance between two points, τ is the time lag in question, α_i is the fraction of the total steps allocated to the the ith term and $MSD_{i,\tau}$ is the MSD for the ith term at time lag τ. This MSD term is equal to $MSD = 2dD\tau + \epsilon^2$, where d is the dimension, D is the diffusion coefficient and ϵ^2 is the error. In the next step, the fit values for the MSD for each ith term is plotted as a function of τ and fit similarly to the MSD method described above (Fig. 1.4E). For example, if we have reason to suspect that there exist two distinct diffusive states within a system we would fit the collection of squared step sizes to a CPD function described above with two states, $K = 1, 2$. Next, we would plot the MSD_1 values for all τ values and fit that curve with a linear function to get the slope and thus the diffusion coefficient of that term. The α parameters inform us how much of the total each diffusive state exists in. We would then repeat for the MSD_2 fit values. In this manner we could in theory fit our data to any number of diffusive states.

While this method allows for multiple diffusive states, CPD fitting does suffer from several drawbacks including weak parameters and supervisory biases. Recent work from our lab published by Rowland *et al.* [27] improved the CPD method by fitting all the time lags simultaneously, thus reducing the parameter space, and removing the two step fitting method outlined above. Further, he implemented a penalty function for adding more diffusive states to the function, in an attempt to remove the desire to "overfit" the data and thereby remove some model selection and user bias. While an improvement, the method fails to completely remove the issues associated with the weak parameters in the fit function and to completely remove issues inherent to curve-fitting, such as the importance of starting values on the final parameter values. Other groups have also tried to address the weaknesses in current SPT analysis methods using various approaches, such as by implementing an expectation maximization algorithm or statistical approaches with more stringent penalty functions, with various degrees of success [30, 33, 34]. SPT experiments provide a wealth of data but without a rigorous mathematical analysis method, conclusions drawn from SPT experiments are, at best, subject to debate and at worst may be actually misleading. For example, if a system has two diffusive states but a given analysis method introduces or includes a third state conclusions about the biological role of that spurious state could lead to incorrect models and wild goose-chases for drugs that affect the state that is not present. SPT experiments have the unparalleled ability to observe biology in real time and at length scales that matter to the molecules themselves. Observations of binding and unbinding kinetics, detailed organizational structure and more are possible with this powerful technique and it is imperative that this wealth of data be underpinned with a foundation of mathematical rigor.

1.4 Thesis Outline

As stated above, my aim in this Thesis was to expand both the quantity and the quality of information that can be gained from SPT experiments by designing a new analysis method free from supervisory biases and weak parameters. I then applied this method to several outstanding questions in microbiology that required a deeper mathematical foundation in order to pursue.

In Chapter 2, I present a new method for the analysis of SPT experiments based on a non-parametric Bayesian statistical approach, which we call SMAUG. I first present the theory behind the approach and my application of this method to the specific dataset of SPT trajectories. I then set out to rigorously test this method using a variety of *in silico* and *in vitro* controlled experiments to validate its accuracy and precision. Finally, I demonstrate SMAUG on two real experimental systems: one in a prokaryotic system and one in eukaryotes.

In Chapter 3, I present my work in understanding the dynamics that govern the process of epigenetic silencing in a yeast model system using SPT and SMAUG. I begin by first providing some biological background about the system and explain why this system needs a method like SMAUG to understand the dynamics involved. I then present the findings for the dynamics uncovered in a minimally perturbed system in which the keystone protein is fluorescently tagged and tracked. Next, I present the finding from an extensive list of mutation studies that perturb the system in illustrative ways. Finally, I close the chapter with some conclusions about the reach and impact of this study.

In Chapter 4, I present my work in uncovering the dynamics of bacterial pathogenesis in *Vibrio cholerae* as a model system using SPT and SMAUG. again, I begin by providing the relevant background needed to understand the system and appreciate why it is a system that requires a method such as SMAUG to analyze. I then present the uncovered dynamics on the minimally perturbed system, followed by the results of targeted mutation experiments that help identify the behavior of the system. I then close the chapter discussing the conclusions and impact of this study on the field and for other similar bacterial systems.

In Chapter 5, I discuss the conclusions of my work in regards to its scope and impact as well as some future directions for these projects. I discuss how SMAUG will provide a foundation for further SPT experiments as it provides a rigorous analysis method from which investigators can draw concrete conclusions from their experiments and that can be widely adopted in the field. In addition, by using the examples from Chapters 3 and 4 as guides, this work demonstrates the strength and versatility of using SPT data to guide further hypotheses and studies.

The work presented in this Thesis has a broad impact for the scientific community by combining aspects of computer science, microbiology, biophysics and mathematics to answer fundamental questions in cellular biology. Additionally, it identifies and addresses several shortcomings in the realm of single-molecule super-resolution fluorescence microscopy by introducing a new analysis framework for the data gathered that can be easily and broadly adopted to expand the quality of information gained from these experiments.

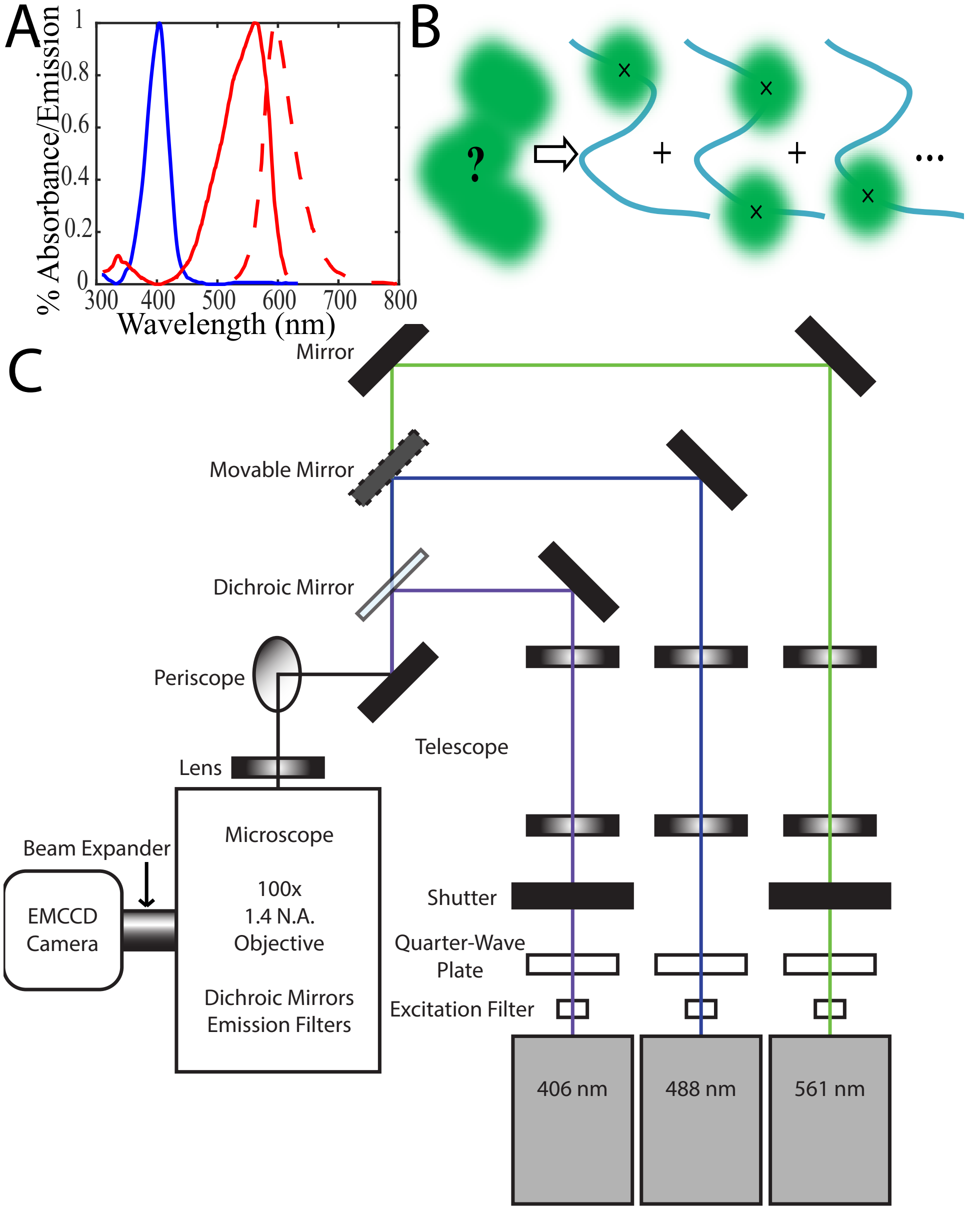


Figure 1.3: Experimental Setup. **A**: Spectra of the FP PAmCherry. Blue trace is the absorbance of the protein in the dark state. Upon activation the protein switches into a fluorescent state with an excitation (red solid) and emission (red dashed) spectrum in the visible range of light. **B**: Overlapping PSFs can hide smaller structures. A sample is bleached with an appropriate excitation laser, then a quick pulse of 405-nm light illuminates the sample and a small subset of fluorophores are activated and imaged with the excitation laser until they bleach. This process is then repeated until all probes are activated and imaged. Non-overlapping PSFs can be fit in order to more precisely localize their centers, leading to increased resolution and undercovering of previously hidden structure. **C**: Schematic of our lab's experimental setup for achieving super-resolution imaging using phtotoconvertible fluorophores. Figure B from Biteen [21]. Figure C from Tuson *et al.* [14]

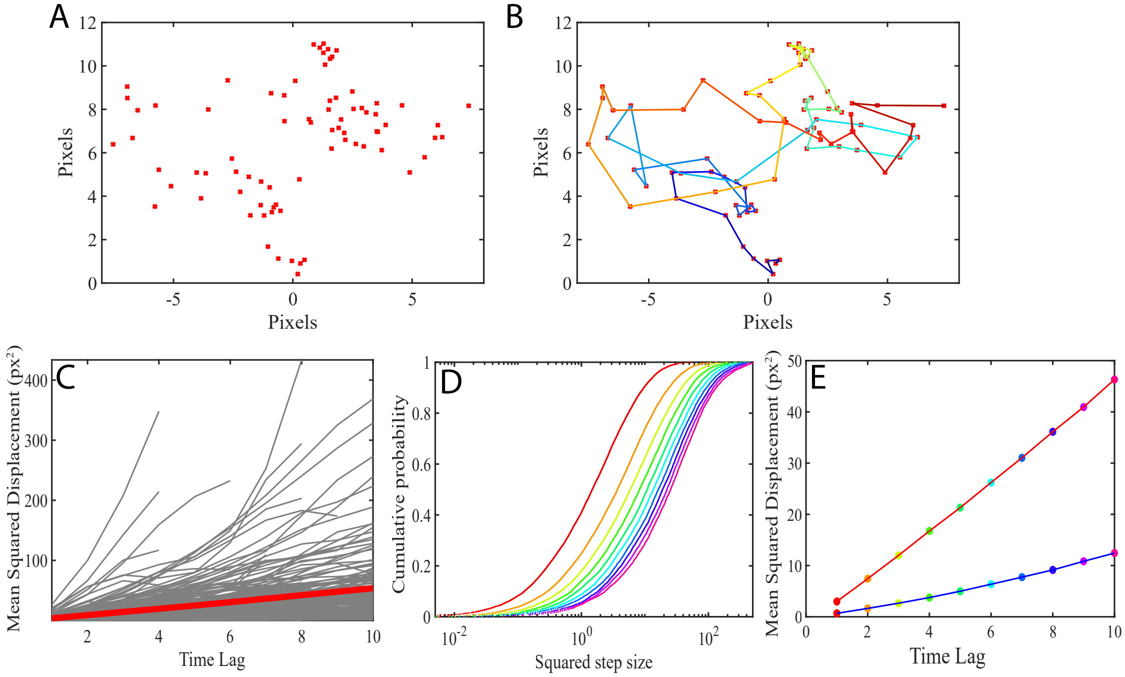


Figure 1.4: Single-Particle Tracking and Analysis. A: Localizations from one simulated molecule over time. **B**: Connecting the localizations from **A** into a trajectory reveals details about the molecule's motion, and thus its environment, at each time. Blue to red represents increasing time **C**: MSD curves for the simulated dataset. Individual trajectories' MSD curves are shown in gray. The mean of all MSD curves is shown in red and fitting this line gives an average diffusion value for the system but can hide heterogeneity. **D**: CPD curves for the same simulated dataset as in **C**. Red to purple represents increasing time lag. **E**: Each fit to the CPD curves returns a value for the MSD at that time lag. Plotting those values and then fitting that curve provides a diffusion value for that term. Circle colors correspond to the CPD curve in **D** with the same color. Data for **C, D** and **E** is the same dataset used in the 4-Term Test (Fig. 2.2) in Chapter II, Section 2.4 which in this case is fit to two states.

CHAPTER II

SMAUG: Analyzing Single-Molecule Tracks with Nonparametric Bayesian Statistics.

The work presented in the chapter has been submitted to the journal Biophysical Journal

Karslake, J.D., Donarski, E.D., Shelby, S.A., Demey, L.M., DiRita, V.J., Veatch, S.L., and Biteen, J.S. SMAUG: Analyzing single-molecule tracks with nonparametric Bayesian statistics. *Submitted March 2019* bioRxiv: 10.1101/578567 [1]

2.1 Introduction

As discussed in the previous chapter, super-resolution fluorescence microscopy is a powerful probe for subcellular biology and single-particle tracking (SPT) measurements have played a central role in measuring the regulation and dynamics of biomolecules inside living cells [22]. In this chapter, I address a key challenge in SPT analysis: our ability to interpret the data provided by high-quality experimental measurements is limited

[1] **Author contributions** - J.D.K. and J.S.B. designed the research. J.D.K. implemented the SMAUG algorithm, wrote the code, performed the simulations, and analyzed the data. J.D.K. and E.D.D. performed the *in vitro* and bacterial imaging experiments. S.A.S. performed the B cell imaging experiments. L.M.D. constructed the bacterial strains and performed the biochemical assays. All authors discussed the results. The manuscript was written and edited by all authors. All authors read and approved the manuscript.

by the analysis framework. To address these limitations in the state of the art, I developed a supervisory-free method for measuring heterogeneous single-molecule dynamics by applying nonparametric Bayesian estimation to SPT experiments, an approach we have termed a Single-Molecule Analysis by Unsupervised Gibbs (SMAUG). Bayesian statistical approaches provide a flexible and robust framework for estimating system parameter values from experimental data. In contrast to more familiar curve-fitting approaches (like those described in Section 1.3), which fit data to a pre-determined function by iteratively adjusting the parameters and checking the residuals, Bayesian approaches estimate the most probable parameters by investigating regions in parameter space where the value of the posterior probability function is very high in order to form a type of topological map of the parameter space. Bayesian approaches have been extensively reviewed, for instance in refs [35–37]. Recently, Bayesian analysis techniques have gained popularity in many fields of biophysics due to their robustness and flexibility. Several recent applications include: analyzing Forster resonance energy transfer (FRET) traces and stepwise photobleaching curves [38], increasing the ability to find and track molecules within single-molecule imaging movies [39], attaining more information from Mean-Squared Displacement (MSD) curves of tracked molecules [30], and mapping the local diffusion coefficients within a cell based on the single-molecule trajectories in each small constructed domain [40].

Parameter	Symbol	Description
Number of mobility states	K	Number of distinct mobility states present in the dataset
Diffusion Coefficient	D	$1 \times K$ vector of diffusion values ($\mu m^2/s$)
Localization Noise	ϵ^2	$1 \times K$ vector of localization noise (nm)
Weight Fraction	π	$1 \times K$ vector of weight fractions
Transition Matrix	T	$K \times K$ matrix where T_{ij} is the probability of transitioning from state i to state j
Theta	θ	A vector of vectors containing all the parameters

Table 2.1: SMAUG parameters. Set of parameters estimated by the SMAUG algorithm to describe a collection of single-molecule trajectories experiencing K mobility states

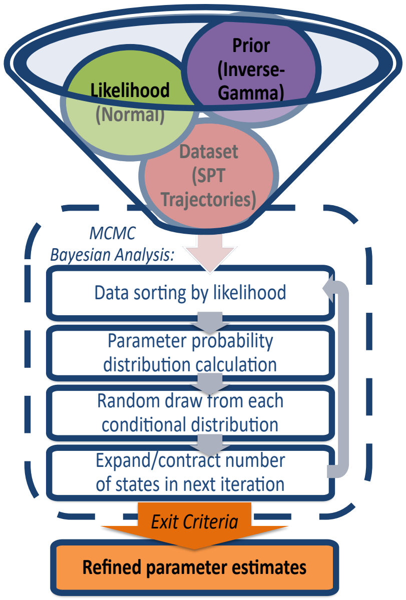

Prior (Inverse-Gamma)
Likelihood (Normal)
Dataset (SPT Trajectories)
MCMC Bayesian Analysis:
Data sorting by likelihood
Parameter probability distribution calculation
Random draw from each conditional distribution
Expand/contract number of states in next iteration
Exit Criteria
Refined parameter estimates

Figure 2.1: SMAUG algorithm. Graphical representation of the SMAUG algorithm, which combines the likelihood, prior, and dataset (top) into a Bayesian framework Markov Chain Monte Carlo (MCMC) algorithm that iterates through four steps to refine the parameter estimates (dashed line) until some exit criteria are satisfied.

In this chapter, I will introduce SMAUG, an algorithm that uses Gibbs sampling to implement a nonparametric Bayesian approach to estimate the most probable information about a heterogeneous collection of mobile molecules. In SPT experiments where multiple biochemical functions give rise to multiple observable mobility states, the SMAUG approach allows us to accurately and precisely determine the underlying parameters of the system. In such a system, the biophysical behavior is described by a set of mobility states, each with an average apparent diffusion coefficient and weight fraction, as well as by the likelihood of transitions between the various mobility states. We use SMAUG to extract these parameters from a collection of single-molecule trajectories free of any supervisory bias such as *a priori* model selection or parameter constraints. Importantly, we uncover novel information that could not be attained with many previous approaches: the probability that a molecule in one mobility state will transition to a different state. The full list of the parameters achieved by SMAUG and a schematic of the algorithm are presented in Table 2.1 and Fig. 2.1, respectively. In the next section I will present the theory of Bayesian inference and its application to SPT. I will then present validations of the SMAUG algorithm on a variety of simulated SPT datasets. Finally, I apply SMAUG to SPT experiments *in vitro*, in bacterial cells, and in eukaryotic systems. Overall, SMAUG provides a concrete mathematical framework that can interpret SPT datasets to provide novel biological insight.

2.2 Methods

Data Analysis

The analysis algorithms used are described in detail in the Theory section. All code and some test datasets are available on github at https://github.com/BiteenMatlab/SMAUG.

Simulated SPT Trajectories

Simulations of SPT experiments were constructed with custom-built MATLAB code (Matlab R2017b, The MathWorks). Each track was constructed with a random track length drawn from an exponential distribution with mean 10 localizations. Each step along the track could belong to one of several mobility states with corresponding diffusion coefficient, D_i. Mobility state labels were assigned for each localization by a random draw from the Transition Matrix. Steps along the trajectory were then constructed using a zero-mean Gaussian distribution with variance equal to $2D_i\Delta t$, where Δt is the frame imaging time. Camera noise and motion blur were applied as described in [41]. The "realistic" range imaging parameters were based on reference [26].

in vitro experiments

Fluoresbrite® microspheres with diameters of 100, 200, and 350 nm (Cat # 21636, Polysciences Inc.) in water were diluted 1:1 v/v with glycerol, and 5 µL of the 50% glycerol mixture was placed between two glass coverslips and imaged with a frame exposure time of 40 ms. Imaging was done in an Olympus IX71 inverted epifluorescence microscope with a 60x 1.20 NA water-immersion objective. Samples were excited by a 488 nm laser (Coherent Sapphire 488-50) with power density 140 W/cm^2. The fluorescence emission was filtered with appropriate filters and imaged on a 512 by 512 pixel Photometrics Evolve electron multiplying charge-coupled device (EMCCD) camera. Recorded single-molecule positions were detected and localized using home-built code as previously described [27], and connected into trajectories using the Hungarian algorithm [24].

***Vibrio cholerae* experiments**

V. cholerae cells containing a chromosomal fusion of the photoactivatable red fluorescent protein, PAmCherry, to TcpP, a membrane-localized transcriptional regulator (TcpP-PAmCherry) as the sole source of TcpP. TcpP-PAmCherry is expressed at the native *tcpP* locus (strain LD51) and cells were grown under conditions known to stimulate TcpP-mediated expression of virulence genes [42] (LB rich media at pH 6.5 and 30 °C). Once cells reached mid log-phase they were diluted into M9 minimal media, and then imaged at room temperature on agarose pads using a 406-nm laser (Coherent Cube 405-100; 102 W/cm^2) for photo-activation and a 561-nm laser (Coherent-Sapphire 561-50; 163 W/cm^2) for imaging. Continual images were collected with a 40-ms exposure time per frame in an Olympus IX71 inverted epifluorescence microscope with a 100x 1.40 NA oil-immersion objective. The fluorescence emission was filtered with appropriate filters and imaged on a 512 by 512 pixel Photometrics Evolve EMCCD camera. Recorded single-molecule positions were detected and localized as previously described using home-built code [27], and connected into trajectories using the Hungarian algorithm [24].

B-cell receptor (BCR) experiments

The BCR dynamics were measured in CH27 mouse lymphoma B cells (RRID:CVCL_ 7178) as described in [43]. Briefly, cells were transiently expressing full-length versions of Lyn kinase or LAT2 (linker for activation of T cells 2)/LAB (linker for activation of B cells) conjugated to mEos3.2. Endogenous, plasma membrane-localized BCR was labeled for 10 min at room temperature with 5 mg/mL goat anti-mouse IgM (Jackson ImmunoResearch; RRID: AB_2338477) f(Ab)1 fragments conjugated to both silicon rhodamine (SiR) dye (Spirochrome, Switzerland) and biotin. Cells were imaged in a live-cell buffer compatible with BCR signaling both before and after the addition of 1Âţg/ml streptavidin, which clusters and activates receptors. Imaging was performed on an Olympus IX81-XDC inverted microscope with a cellTIRF module, a 100x UAPO TIRF objective (NA = 1.49), and active Z-drift correction (ZDC). Excitation of the SiR dye was accomplished using a 647 nm solid-state laser (OBIS, 100 mW, Coherent, Santa Clara, CA). Photoactivation of mEos3.2 was accomplished with a 405 nm diode laser (CUBE 405-50FP, Coherent) with excitation using a 561 nm solid-state laser (Sapphire 561 LP, Coherent). All images were taken on an iXon-897 EMCCD camera (Andor, CT) at approximately 45 frames/s with an exposure time of 20 ms. Recorded single-molecule positions were detected, localized, and connected into trajectories as described in [44]. Data acquired for Lyn was reported previously [43] and reanalyzed for this work.

2.3 Theory

2.3.1 Bayesian Statistics

Using SMAUG, we interpret collections of single-particle trajectories: a set of single-molecule positions that are connected in time as the molecule moves along some path. The data set, $\boldsymbol{y}$, is therefore a distribution of step sizes that remain connected by their trajectories, and this data set as a whole is the consequence of the physical parameters that govern the motion of the individual single-molecules observed during an experiment (parameters summarized in Table 2.1). Here, we consider these physical parameters as a vector of parameters, $\boldsymbol{\theta} = \{\boldsymbol{D}, \boldsymbol{\epsilon}^2, \boldsymbol{\pi}, \boldsymbol{T}\}$, to ease notation.

Whereas traditional fitting algorithms assume a model function, *f*, and fit the function to the data by iteratively adjusting the parameters by some some described method until a cut-off is reached, Bayesian estimation instead looks to maximize the posterior probability distribution, $p(\theta|y)$, which is a measure of where in probability space the most likely set of parameters are that gave rise to the observed data. For some simple cases this calculation might be rather straightforward but for more complicated calculations we can take advantage of some mathematical tricks to arrive at a solution.

Treating both the data and the parameters as random variables instead of thinking of the data as fixed leads us to calculating the *joint probability*, $p(y, \theta)$, of the data and the parameters together. The place in parameter space where this function is highest should correspond to the most likely parameter values of the posterior as well. However what functional form that this joint probability calculation exists in might not be readily apparent but we can use the definition of the conditional probability for each "set" of random variables to arrive at an equation that can be manipulated further.

$$p(y, \theta) = p(y|\theta)p(\theta) = p(\theta|y)p(y) \tag{2.1}$$

In other words the joint probability of the combined random variables is equal to the probability of one of the random variables conditioned on the other set being treated as fixed. Simple manipulation of the two expressions yields Bayes' Rule, a general expression for calculating the posterior distribution:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{2.2}$$

Where $p(\theta|y)$ is the posterior distribution and describes how likely a set of parameters is given the data. The remaining factors of this calculation are: the likelihood, $p(y|\theta)$, a measure of how probable a the data is given a set of parameters; the prior probability of the parameters, $p(\theta)$, which encodes our knowledge and physical intuition about the system before any data is collected; and the marginal likelihood of the data, $p(y)$, also called the evidence. Because the evidence is hard to calculate and is independent of the parameters and is thus constant it is usually dropped and Bayes' Rule is more commonly rewritten as:

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{2.3}$$

Again, in the most general and simple of cases, the posterior distribution could be calculated by evaluating the data set at all possible parameter values and then looking for the point at which the calculation is maximized. However, for more complex cases where the posterior can be a mixture of several states and/or where several different posteriors distributions must be calculated, straightforward calculation of all parameter space is impractical if not impossible. In such cases, the main goal of a Bayesian algorithm remains the same: to calculate the posterior distribution in order to find regions of high probability that in turn describe the mostly probable estimates that explain the observed dataset. However, in these cases, the calculation requires methods that are more advanced.

Accordingly, SMAUG uses a Gibbs analysis based approach for the analysis of SPT data by embedding a Gibbs sampling scheme [45, 46] within a Markov Chain Monte Carlo (MCMC) framework [47]. SMAUG implements this Markov scheme iteratively in two broad steps (Fig. 2.1). In the first step, SMAUG calculates the posterior distribution of each of the parameters in $\boldsymbol{\theta}$ using Gibbs sampling. The Gibbs sampling method iteratively updates each parameter's posterior distribution individually while holding all other parameters constant to reduce an otherwise impossibly complex posterior calculation into manageable and calculable chunks. In the second step, new parameter values are sampled (i.e., new values for $\boldsymbol{\theta}$ are pulled from these calculated posterior distributions) and saved for the first step of the next iteration.

Here, in addition to being a complex mixture of distributions for one mobility state the posterior is also a mixture model of multiple mobility states that can be present in the dataset. Thus, a data-selection step precedes the sampling step described above. In this data-selection step, each data point in $\boldsymbol{y}$ is assigned to a particular mobility state of the mixture, and only the subset of data belonging to that state is used in the posterior calculations that describe it. Below, we describe our Gibbs sampling process first for a known number, K, of mixture states as it is more straightforward; afterward, we describe the process for expanding to an infinite number of states, which is necessary for SMAUG to learn the correct number of states present in a dataset.

2.3.2 Constructing a Gibbs Sampler for a known K

With the basics of Bayes' method discussed, I can now discuss how SMAUG actually achieves the steps of the Markov Chain process and estimates parameters for a SPT dataset. For a model of given complexity, K, we compute the conditional posterior distribution: the posterior distribution for one single parameter while all other parameters remain constant. Before we can preform the calculations we first need to assign data to the various K states. To identify which data point comes from which of the K mixture states, we introduce a latent variable, l_i, which labels each of the data points with the number of the state from which it was likely drawn. L_j is the set of all data points with $l_i = j$. For each l_i, we calculate the likelihood function, $p(y_i|\theta_j)$, (explained below) for each data point (i.e. each step in the set) belonging to each of the K states individually and then we draw the assignment using the categorical distribution with weights equal to the likelihood calculated for each state:

$$p(l_i = j|...) \propto p(y_i|\theta_j) \tag{2.4}$$

$$l_i \sim Cat(p(l_i = 1|...), p(l_i = 2|...), ..., p(l_i = K|...)) \tag{2.5}$$

The very first assignment of the data can be random (or some other method), as information about the likelihood has not yet been calculated. In every iteration, having sorted the data into subsets L_j that are relevant to each state, SMAUG then proceeds to the Gibbs sampler.

Two of the parameters we wish to find for our dataset are the diffusion coefficient values, $\boldsymbol{D_j}$, and the localization noise, $\boldsymbol{\epsilon}^2$. In a model derived by Berglund [41, 48], our dataset of step sizes from SPT experiments can be related directly to these quantities of interest. By this model, the measured steps sizes, Δx, are zero-mean Gaussian variables whose covariances are related to $\boldsymbol{D_j}$ and $\boldsymbol{\epsilon}^2$ by:

$$< \Delta x_i^2 >= 2D_j\Delta t(1-2R) + 2\epsilon_j^2 \qquad\qquad < \Delta x_i \Delta x_{i\pm1} >= 2D_j\Delta tR - \epsilon_j^2 \tag{2.6}$$

where Δt is the exposure time of the frame and R is the motion blur coefficient, which is set to 1/6 in the algorithm as our acquisition and exposure times are equal [41]. While SMAUG could in principle be adapted to include other physical manifestations like confinement, we focus here on apparent free diffusion and therefore we model the trajectories as from the result of a zero-mean Gaussian process as stated above. Specifically, the likelihood function, $p(y|\theta)$, for our system is a Gaussian, denoted $N(\mu, \sigma^2)$, with unknown mean, μ, and unknown variance, σ^2. For most purposes, these step size distributions should be zero-mean ($\mu = 0$), but we retain the unknown mean parameter to be as general as possible. Since we have specified that there are K distinct states in this dataset, we expand the likelihood to a Gaussian Mixture Model [49] that includes K such Gaussian distributions, each scaled by the amount of data in that state, expressed as the fraction of the whole, π_i (this weight parameter is discussed more below):

$$p(y|\pi_1\theta_1, ..., \pi_K\theta_K) \propto \pi_1 N(\mu_1, \sigma_1^2) + ... + \pi_K N(\mu_K, \sigma_K^2) \tag{2.7}$$

For the prior distribution for $\boldsymbol{D_j}$ and $\boldsymbol{\epsilon}^2$, SMAUG takes the conjugate prior to our likelihood: the Inverse-Gamma function, $IG(a, b)$. A conjugate prior is a prior distribution that when multiplied to the likelihood returns a posterior that is of the same mathematical family as the likelihood, simplifying the computation. Using the conjugate Inverse-Gamma has and added benefit that it is constrained to be positive valued, as diffusion cannot be negative. Additionally, if we have some other knowledge about the system (say from a previous experiment) we could encoded that knowledge into the prior, such as adjusting the mean of the Inverse-Gamma to influence the resulting variance distribution to reflect this previous knowledge. However, by default SMAUG uses flat priors to let the data speak for itself as much as possible but this can altered easily by changing the prior parameters in the code. By constructing the likelihood and the prior this way, SMAUG arrives at the full conditional posterior distribution function for diffusive motion: the Normal-Inverse-Gamma function, denoted $NIG(\mu_i, \sigma_i^2, a, b)$, the estimates for the covariances of which for the input step size distributions can be related to our parameters of interest by Equation 2.6:

$$p(\theta|y, l_1, ..., l_N) \sim \prod_{i \in L_j} NIG(\mu_j, \sigma_j^2, a, b) \tag{2.8}$$

With the conditional for $\boldsymbol{D_j}$ and $\boldsymbol{\epsilon^2}$ described, I will now turn to describing two other parameters: the weight fraction for each of the K states, $\boldsymbol{\pi_j}$, and the rows of the transition matrix, T_j. Each data point can only have come from a single mobility term and so the likelihood function for the weight fraction is the categorical distribution (which is a special case of the more generalized multinomial distribution where only a single outcome is observed). The prior distribution is chosen as the conjugate Dirichlet distribution, which is the multivariate generalization of the Beta distribution as has the property of always summing to 1. The resulting posterior distribution is also a Dirichlet distribution, denoted $DIR(a)$:

$$p(\pi_1, ..., \pi_K|...) \sim DIR(L_1 + c, ..., L_K + c) \tag{2.9}$$

where the L_js are the number of data points assigned to each mobility state described above and the vector (c, ..., c) is a vector of pseudo-counts that are used to describe the prior weight, for SMAUG we use a vector where all the values of this vector are the same and set to 1/K, a common uninformative prior.

Similarly, we can construct a transition matrix that describes the probability that a molecule in state i at time t will transition to any of the possible K state, including remaining in state i, at time t+1. Using the state assignments coded in the set of l_is and the trajectory information from the track data we can construct a $K \times K$ matrix that counts when subsequent steps within a trajectory change their assignment. $N_{i,j}$ counts the number of transitions from state i to state j. Each of the K rows of the transition matrix, $\boldsymbol{T}$, are then sampled using the Dirichlet distribution again with the counts in $\boldsymbol{N}$ as inputs and a vector of pseudo-transition counts, (c, ..., c), used as before to describe the prior weight:

$$p(T_i|\ldots) \sim DIR(N_{i,1} + c, \ldots, N_{i,K} + c) \tag{2.10}$$

Taken together, this collection of conditional posterior distributions is an effective method for calculating the full posterior of the system. In the second step, these newly defined distributions are used as the basis from which new parameter values are sampled to get the parameter values that will be used in the calculations of the next iteration. Iterating between assigning data based on the previous parameter values, then recalculating the distributions based on the new data assignments and finally sampling parameter values from the calculated distributions leads to an effective and efficient sampler for finding the most probable parameter values for a system of K states.

However, we rarely know *a priori* how many distinct states to include in an analysis; in fact learning this number is usually one of the principle goals of an SPT experiment [26, 27, 50]. We could possibly set some large upper bound for K, but then much of our computational power would be directed towards calculating states with zero occupancy, leading to a computationally inefficient process. Instead, in the next section, we outline a Dirichlet process mixture model (DPMM) method that allows the number of states present to expand or contract organically in response to the data based on a nonparametric Bayes approach.

2.3.3 Constructing a nonparametric Bayes sampler

Like before where we treated both the data and the parameters as random variables in order to construct an estimator, nonparametric Bayesian techniques rely on treating the probability distributions themselves as random. Random probability measures extend a finite-component mixture like the one described above in section 2.3.2 into an infinite-component mixture model needed for completely hands-off estimator (free from supervisory bias) [51]. SMAUG uses the Dirichlet Process (DP), $DP(\alpha, P_0)$, one such random probability measure which is generally described as a "distribution of distributions". Specifically, $DP(\alpha, P_0)$ is a distribution with base probability distribution, P_0 (such as a normal Gaussian or a beta distribution), and concentration parameter, α (which controls the variance around P_0). The DP can be seen as the infinite dimensional generalization of the standard Dirichlet distribution and, as with the standard Dirichlet, the "weights" drawn must sum to 1, which helps induce a clustering onto the infinite collection of possible states present in the nonparametric realization of the sampler.

A helpful visualization for understanding what a draw from a Dirichlet process looks like is the stick-breaking construction [52], which represents the total probability available to the system as a stick of unit length. First, a random sample, $\theta \sim P_0$, is drawn from the base probability measure P_0 (θ_1 can be a single value or a vector), and random weight, $V_1 \sim Beta(1, \alpha)$, is pulled from the Beta distribution. We give a probability weight of $\pi_1 = V_1$ to point mass θ_1. We then break the unit stick at V_1 and there now remains an amount of stick, $(1-V_1)$, to be allocated to the many other draws. We then break an amount $V_2 \sim Beta(1, \alpha)$ off the remaining stick and assign probability $\pi_1 = V_2(1 - V_1)$ to another point mass of probability $\theta_2 \sim P_0$. As we continue, the stick gets shorter and shorter and the weight assigned to each new draw from P_0 decreases with a rate that depends on the concentration parameter, α.Thus, our random probability measure is an arbitrarily large collection of segments of which only several have the vast majority of the probability weight; the rest of the segments have negligible mass. This stick-breaking construction can be summarized as:

$$P \sim \sum_{h=1}^{\infty} \pi_h \delta_{\theta_1}, \qquad \pi_h = V_h \prod_{l<h} (1 - V_h), \qquad V_h \sim Beta(, \alpha), \qquad \theta_h \sim P_0 \tag{2.11}$$

where the θ_h parameter value vectors are generated independently from $P + 0$, $\delta_{\theta_h}$is the point mass where the parameters θ_h are concentrated, and π_h is the probability weight associated with that point mass.

This method results in an infinitely large parameter space of which only a few states actually occupy any meaningful probability mass. SMAUG uses the Slice Sampler method from Walker to reduce the infinite state model that results from the distributions above, in Equation 2.11, to a model with only finitely many states capable of being calculated at each iteration [53]. The Slice Sampler method introduces another latent variable, u, which is drawn from the uniform distribution as $u_i \sim U(0, \pi_{l_i})$, for each data point in the set. Thus, any draw for u splits the infinite set of possible states into two categories: a finite set of states for which $\pi_j > u$ and an infinite set for which $\pi_j < u$. By looking for the minimum entry over the set of all u and seeing how many of the finitely many states with probability weight greater than that value there are we know the maximal size of the model we need to include for any iteration, i.e. the minimum value for u over the set provides an upper bound on the number of states, K, we need at any given time. Specifically SMAUG attempts in every iteration to satisfy the inequality:

$$\sum_{1}^{K'} \pi_j > 1 - min(u_1, ..., u_N) \tag{2.12}$$

where K' is the number of states present in the model at any time. In this way, only K' states need to be calculated, but over the course of sampling many iterations, we integrate over an "infinite" (or at least arbitrarily large) number of possible states. The value of K' can expand or contract over the course of the analysis with new terms being added when needed and terms whose occupancy is very low (i.e., states of a few data points or less) removed. In this way the sampler "learns" from the data itself how many clusters exist in the data set, removing the need to specify the number at the start.

Taken all together, SMAUG provides an efficient nonparametric Bayesian analysis framework for analyzing SPT data that returns accurate and precise estimates of the number of mobility states within a dataset, their diffusion coefficients, weights, localization errors and transitions in a hands-free manner. During each iteration of the sampler, SMAUG follows a simple stepwise process as outlined above (Fig. 2.1):

1. **First iteration only**: choose an initial number of states. This number should be selected to be several times bigger than the expected number for the experiment. Assign each of the data points to a state by some method.

2. Assign a vector of parameter values to each state, for instance by random draws from a base distribution, P_0, or by calculating the simple statistics (mean and variance) from the previous step's assignment.

3. **Second iteration onward**: Assign each data point a latent variable, u, and use these values of u to determine K', the number of states present for this iteration.

   (a) If K' is greater than the current number of states then states need to be added; assign each of the new states a weight by breaking the stick and pulling a parameter vector from P_0.

   (b) If K' is less than the current number of states then a state needs to be removed; remove the state with smallest weight and add its weight to the next smallest weight.

   (c) If a state's occupancy drops out (by receiving zero weight fraction), remove the values for that state from the parameter vector

4. Implement a Gibbs Sampler with the fixed number of states, K', from step (3). Assign labels and update parameters by calculating the conditional posterior distributions

described in equations 2.5 - 2.10 above, then sample from these distributions to collect new parameter values for the next iteration.

5. **Exit criteria**: Repeat steps (3) and (4), saving values periodically, until some cutoff criterion has been achieved, either based on performing some number of total iterations or attaining some convergence metric, then construct parameter estimates from the back half of saved iterations.

The efficient SMAUG algorithm we have built provides a flexible method for determining all the relevant parameters for an arbitrary SPT trajectory dataset without supervisory bias. For instance, the amount of data generated in step (4) can be controlled by not saving the parameters of interest every iteration (by default, SMAUG saves every tenth iteration to minimize any possible autocorrelation between iterations).

We demonstrate below that SMAUG accurately and precisely estimates for SPT experiments the number of mobility states, the diffusion coefficients, the weight fractions, the noise values, and the frequencies of transitions between states. To demonstrate the value and feasibility of this nonparametric Bayesian algorithm, we validate our method first by using simulated diffusion trajectories with realistic parameter values (section 2.4) and an in vitro experimental system (section 2.5), and then we apply SMAUG to subcellular tracking in bacterial cells and in eukaryotic cells (sections 2.6).

2.4 *in silico* Validations

To begin, We validated the SMAUG algorithm with a simulated dataset (ref table) containing 13,636 steps (1090 trajectories) drawn from a diffusive mixture with four distinct mobility states, $i = \{1, 2, 3, 4\}$. The diffusion coefficients for the terms were seeded with the values $D = \{0.005, 0.03, 0.09, 0.20\}$ μm^2/s, and the localization error for each localization, ϵ_j^2, were pulled from a distribution with a mean of 10 nm and a variance of 5 nm. The weight fractions of each term were: $\{\pi_1, \pi_2, \pi_3, \pi_4\} = \{0.196, 0.301, 0.291, 0.212\}$. The

Parameter	**Seed Values**	**True Values**	**SMAUG Results**
Number of Mobility States	4	4	4
Diffusion Coefficient ($\mu m^2/s$)	0.005, 0.03, 0.09, 0.20	0.005, 0.03, 0.09, 0.20	0.0051,0.0305,0.836,0.201
Standard Deviation	NA	NA	0.0003,0.0016,0.0082,0.0093
Localization Noise (nm)	10 ± 5	10 ± 5	5.7, 8.8, 9.6, 13.7
Weight Fractions	0.25,0.25,0.25,0.25	0.196,0.301,0.291,0.212	0.192,0.322,0.251,0.235
Standard Deviation	NA	NA	0.0076,0.0223,0.0234,0.0235
Transition Matrix	$\begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 \\ 0.1 & 0.8 & 0.1 & 0 \\ 0 & 0.1 & 0.8 & 0.1 \\ 0 & 0.1 & 0.1 & 0.8 \end{pmatrix}$	$\begin{pmatrix} .806 & .104 & .090 & 0 \\ .107 & .801 & .092 & 0 \\ 0 & .104 & .801 & .095 \\ 0 & .097 & .097 & .806 \end{pmatrix}$	$\begin{pmatrix} .812 & .132 & .033 & .023 \\ .069 & .776 & .124 & .031 \\ .023 & .166 & .676 & .134 \\ .022 & .047 & .145 & .785 \end{pmatrix}$

Table 2.2: 4 Term simulation values. Seed values, true values, and SMAUG results for the four-term simulation described in Fig. 2.2 and Fig. 2.3 . Total number of steps included is 13,636.

transition matrix was:

$$\text{T=}\begin{pmatrix} .806 & .104 & .090 & 0 \\ .107 & .801 & .092 & 0 \\ 0 & .104 & .801 & .095 \\ 0 & .097 & .097 & .806 \end{pmatrix}$$

where T_{ij} is the probability of a step in mobility state i preceding a step in mobility state j as mentioned above. In these realistic simulations, the track lengths and transitions are random events: each track length is pulled from an exponential distribution with a mean of 10 localizations, and $t_{i \to j}$, the likelihood of transitioning from state i to state j, is governed by a flat distribution. The seed values for the simulations, which gave rise to the simulated values, are given in Table 2.2. As opposed to methods that fit data to a specific model with a selected number of mobility states, K, one strength of the SMAUG algorithm lies in its ability to identify the correct number of mobility states. The algorithm was initialized with a large number of components ($K \geq 10$), but quickly collapsed to the correct number ($K = 4$) (Figure 2.2 A). In general, the model complexity is increased or decreased until it converges at the correct number, though SMAUG continues to explore state space by adding states on occasion and then removing them (e.g., the bumps up to 5 in Fig. 2.2 A). To construct parameter estimates for this case, we only use posterior draws from saved iterations where K is the convergence value (here $K = 4$) in the back half of all saved iterations (iterations 501-1000, red box in Fig. 2.2 B).

Each parameter is observed throughout the course of the simulation (Fig. 2.2 B, Fig. 2.3) and the terms are sorted by D in the final output. We use only the second half of saved iterations (red box in Fig. 2B) to construct estimates. The posterior distributions are plotted for several parameters (Fig. 2.2C, Fig. 2.3). Because these data points represent draws from the converged posterior distributions for the parameters, we use these histograms to calculate statistics about our estimates or construct confidence intervals. The mean values in all cases are close to the true values (black arrows in Fig. 2.2C and Fig. 2.3). At each step in the analysis, the best estimates for all parameters are generated, and each pair $\{D_i, \pi_i\}$

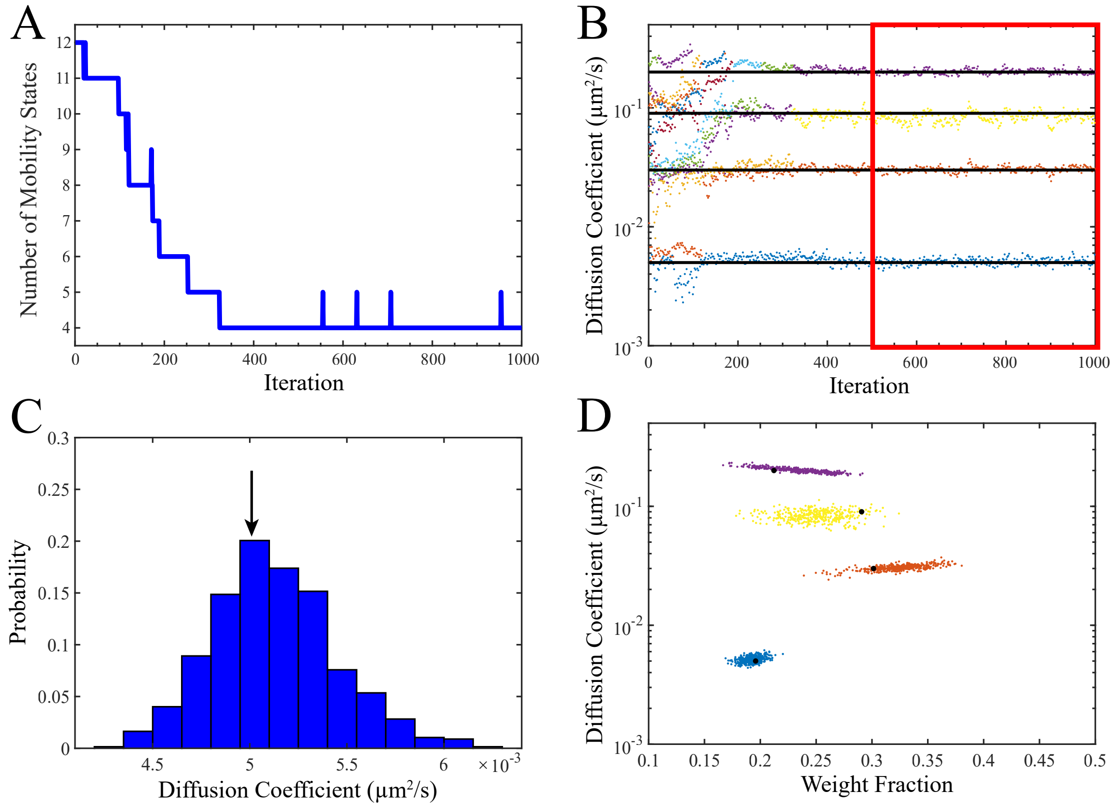


Figure 2.2: SMAUG Analysis of Simulated Test Data. A: SMAUG analysis of the simulated input data as a Gaussian Mixture Model. The algorithm initializes at a large number of states and quickly converges to the correct value of $K = 4$. However, after convergence additional states are added stochastically as the algorithm explores state space looking for other regions of high probability. **B**: Estimates of the diffusion coefficients, D, for each term (sorted in order of increasing D) as the algorithm progresses. Black lines are the true simulation values (Table 2.2). **C**: Histogram defining the probability of a given diffusion coefficient for the slowest term in the analysis (term 1 in Table 2.2). Histograms are constructed using the back half of saved iterations for the blue/slowest term (red box in **B**). Black arrow is the true value for the simulation. **D**: Diffusion Coefficients and weight fraction estimates for each saved iteration in the back half of the analysis run that also meets the $K = 4$ criterion. The analysis shows distinct clusters whose estimates do not overlap. Black dots are the true simulation values. Full histograms for all output values are in Fig 2.3

for every saved iteration is plotted as a point in Fig. 2.2 D. True values for the simulation (Table 2.2) are indicated by the large black data points in Fig. 2.2D.

To examine the ability of SMAUG to detect rare occurrences, we simulated a dataset (Table 2.3) containing 9,445 steps in which the majority of trajectories (95.5%) belonged to a fast diffusing state ($D_1 = 0.15\mu m^2/s$) while the rest belonged to a slower state ($D_2 =$

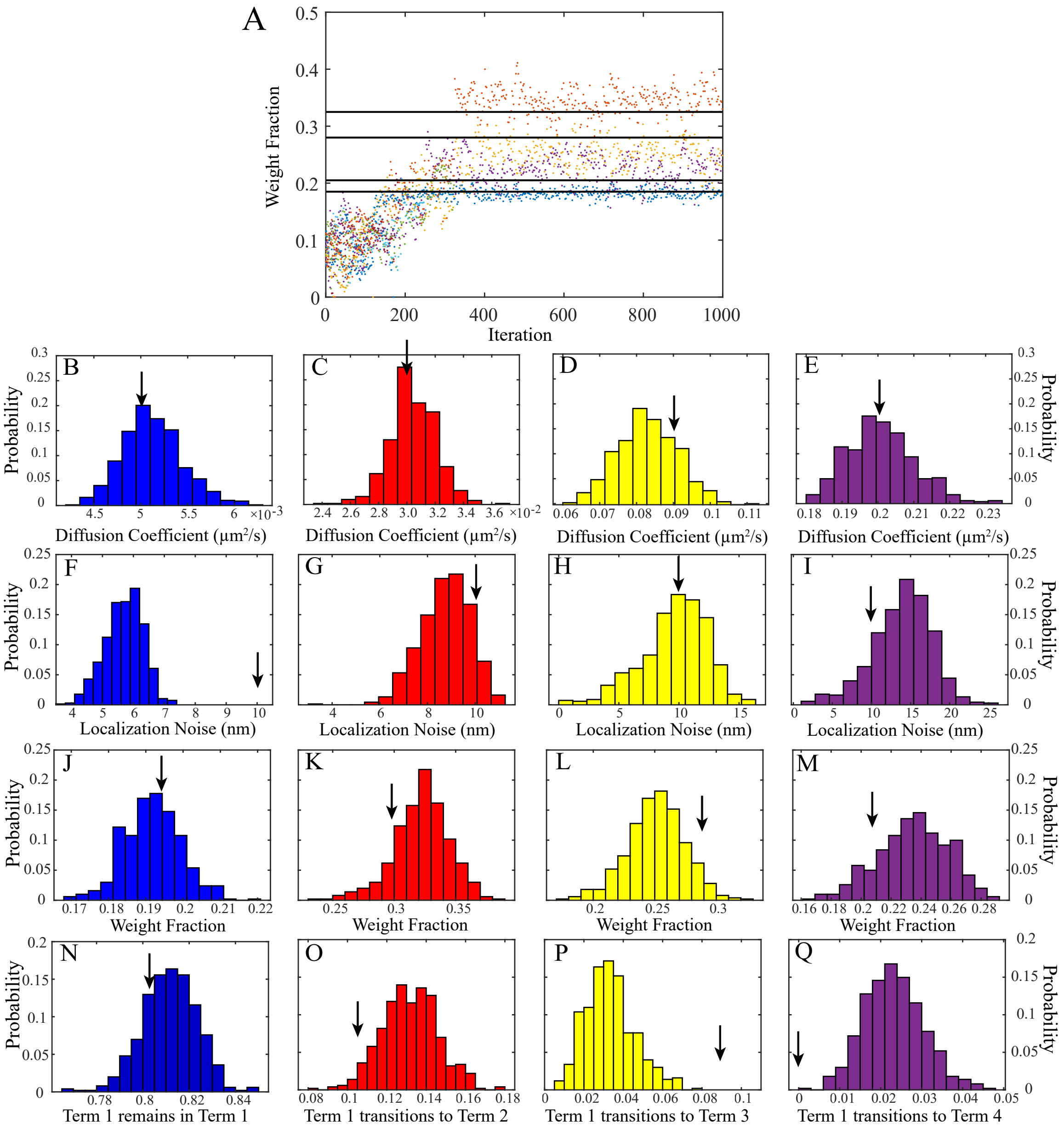


Figure 2.3: Full SMAUG analysis for simulated data. A: Estimates of the weight fraction for each term (sorted in order of increasing diffusion coefficient) as the algorithm progresses. Black lines are the simulation true values (2.2). **B-E**: Histograms of the diffusion coefficient estimates for each of the 4 terms over the back half of iterations. **F-I**: Histograms of the estimates of the localization noise for each of the 4 terms over the back half of iterations. **J-M**: Histograms of the estimates of the weight fractions for each of the 4 terms over the back half of iterations. **N-Q**: Histograms of the estimates for the transition matrix elements giving the probability that a step in Term 1 is followed by a step in Term 1 on the next step (**N**), that a step in Term 1 transitions to a step in Term 2 (**O**), that a step in Term 1 transitions to a step in Term 3 (**P**), or that a step in Term 1 transitions to a step in Term 4 (**Q**). Black arrows in B - Q are the true simulation values.

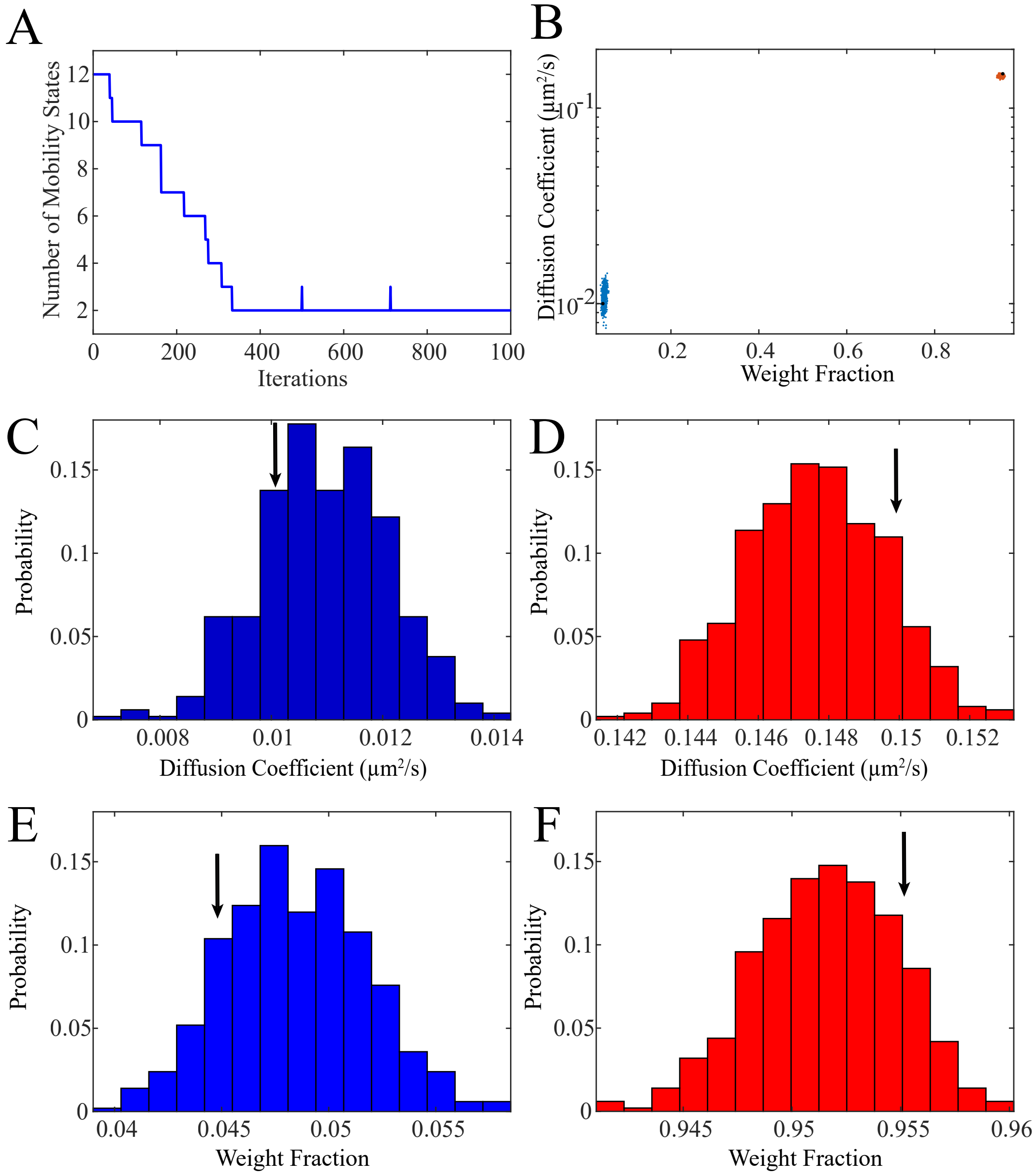


Figure 2.4: Rare states simulation. Full SMAUG analysis for the rare states simulation. **A**: Estimated mobility states over the course of the analysis run. SMAUG quickly converges to the correct value of $K = 2$, but continues to explore alternative hypotheses stochastically. **B**: Diffusion coefficient and weight fraction estimates for each saved iteration in the back half of the analysis run that also meets the $K = 2$ criterion. Black dots are the true simulation values. **C-D**: Histograms for the estimated diffusion coefficient values for the simulation. **E-F**: Weight fraction estimates. Black arrows are the true simulation values (Table 2.3).

Parameter	Seed Values	True Values	SMAUG Results
Number of Mobility States	2	2	2
Diffusion Coefficient ($\mu m^2/s$)	0.01, 0.15	0.01, 0.15	0.011, 0.146
Standard Deviation	NA	NA	0.0011, 0.0019
Localization Noise (nm)	10 ± 10	10 ± 10	12.1, 15.7
Weight Fractions	0.05, 0.95	0.045, 0.955	0.048, 0.952
Standard Deviation	NA	NA	0.0033, .0035
Transition Matrix	$\begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$	$\begin{pmatrix} 0.987 & 0.013 \\ 0.009 & 0.991 \end{pmatrix}$	$\begin{pmatrix} 0.969 & 0.031 \\ 0.001 & 0.999 \end{pmatrix}$

Table 2.3: Rare states simulation values. Seed values, true values, and SMAUG results for the rare-states simulation in FIG. Total number of steps included is 9,445.

$0.01 \mu m^2/s$). Furthermore, the transitions between states 1 and 2 were rare ($T_{12} = T_{21} = 0.01$). This distribution is relevant for experiments in which the binding events of biomolecules are rare, and analyzing this simulation explores the ability of SMAUG to confidently distinguish rare states from random events within a homogeneous distribution. SMAUG isolates the two distinct populations (Fig. 2.4) and accurately estimates their parameter values (Table 2.3). SMAUG can easily identify states whose occupancy is only a small fraction of the whole dataset.

2.5 Validations *in vitro*

We further tested the SMAUG method with an *in vitro* experimental system consisting of three different sizes of diffusing fluorescent beads in a 50/50 water/glycerol mixture. The Stokes-Einstein equation predicts a diffusion coefficient of $D = kT/6\pi\eta r$ for a particle of radius, r, undergoing Brownian motion in a fluid with viscosity, η; this equation predicts theoretical diffusion coefficients of $D = \{0.182, 0.319, 0.637\} \mu m^2/s$ for this system. SMAUG analysis of the bead trajectories (Table 2.4) correctly identified the number of distinct diffusors ($K = 3$) and estimated values of $D = \{0.168, 0.329, 0.675\} \mu m^2/s$ (Fig. 2.5). The distributions of the estimations of D at every saved iteration (Fig. 2.5) show that the theoretical D values are within the confidence intervals of the estimations. Furthermore, the transition matrix shows negligible transitions between states ($T_{(ij)(i \neq j)}$) < 0.03).

This observation is consistent with our attribution of each state to one bead size as beads cannot change sizes spontaneously and thus no transitions are allowed.

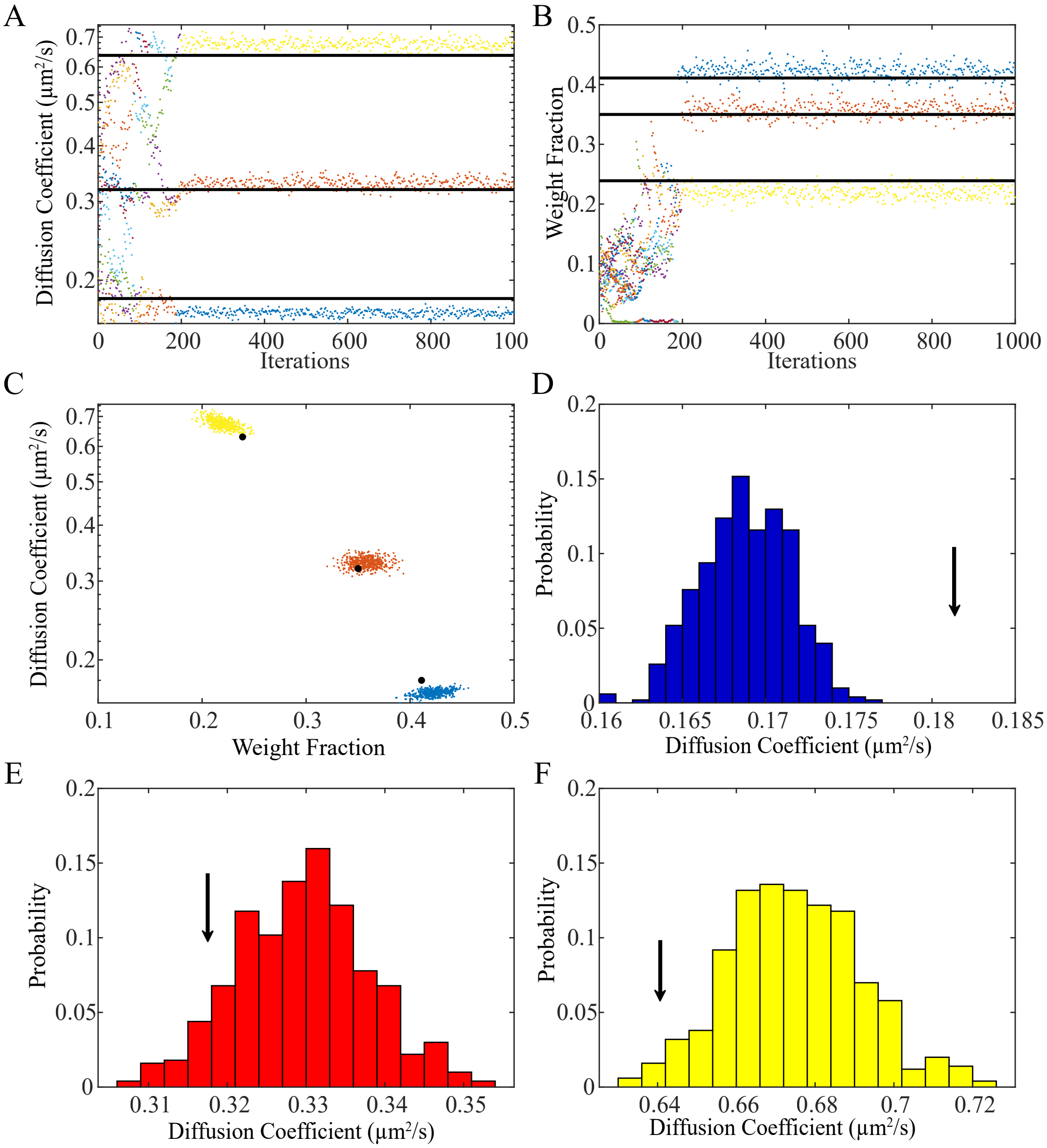


Figure 2.5: SMAUG analysis of the beads *in vitro* experiments. A: Diffusion coefficient estimates for the analysis run. Black lines are the theoretical values for diffusion of beads in 50% glycerol. **B**: Weight fraction estimates for the analysis. Black lines are the true value of the number of steps from each size of bead. **C**: Diffusion coefficient and weight fraction estimates for each saved iteration in the back half of the analysis run that also meets the $K = 3$ criterion. Black dots correspond to the black lines in **A** and **B**. **D-F**: histograms for the diffusion coefficient estimates. Black arrows represent the theoretical values (Table 2.4).

Parameter	Theoretical Values	SMAUG Results
Number of Mobility States	3	3
Diffusion Coefficient ($\mu m^2/s$)	0.182, 0.319, 0.637	0.168, 0.329, 0.675
Standard Deviation	NA	0.0027, 0.0083, 0.0166
Weight Fraction	0.411, 0.350, 0.239	0.423, 0.358, 0.219
Standard Deviation	NA	0.0110, 0.0119, 0.0106
Transition Matrix	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0.977 & 0.018 & .005 \\ 0.021 & 0.946 & .036 \\ .008 & 0.059 & 0.932 \end{pmatrix}$

Table 2.4: *in vitro* validation values. Theoretical values and SMAUG results for the diffusing beads experiments in Fig. 2.5. The theoretical diffusion coefficient is calculated from the Stokes-Einstein Equation. The theoretical weight fraction is based on taking the fraction of number of steps that came from each bead in the total combined data set. The theoretical transition matrix includes no transitions as the beads cannot spontaneously change sizes. Total number of steps included is 31,949.

Overall, SMAUG accurately determines the values for parameters of interest (Table 2.4): the number of distinct mobility states within a dataset; the diffusion coefficient, and the weight fraction, of each state; and the transition probabilities between the states at each iteration for these *in vitro* experiments.

2.6 Live-cell Investigations

2.6.1 Application to measuring protein cooperativity in living *Vibrio cholerae* bacteria cells

We extended SMAUG to live-cell single-molecule tracking to quantify the diffusion coefficients and distributions in biological systems. The pathogenic bacterium *V. cholerae* remains a global health concern, infecting millions each year leading to the diarrheal disease cholera [54]. The cholera toxin (CtxAB) and an adherence organelle called the toxin-coregulated pilus (TcpA-F) are key determinants of virulence that are under the regulatory control of ToxT, which itself is regulated by the membrane protein TcpP [42]. In collaboration with other membrane proteins (TcpH, ToxR, and ToxS), TcpP initiates the *V. cholerae* virulence cascade by binding to the promoter region of the *toxT* gene while remaining in the membrane (Fig. 2.6). Accurately measuring TcpP dynamics in live cells will enable investigations of this unusual membrane-localized mechanism of transcription activation. Previously, our lab used fusions to the photoactivatable fluorescent protein PAmCherry to show that TcpP-PAmCherry diffuses heterogeneously in living cells [26].

We tested the SMAUG algorithm on *V. cholerae* cells which encode a chromosomal copy of *tcpP-PAmCherry* that remains under the control of its native promoter (Methods 2.2). These TcpP-PAmCherry fusions were fully functional based on expression levels of downstream protein CtxB and the cells (strain LD51) exhibited wild-type growth rates (Fig. 2.7). Furthermore, we observed regular cell morphology under the microscope (Fig. 2.6B). We grew these cells under virulence-inducing conditions (Methods 2.2) and collected 11,403 steps from 2404 trajectories; representative trajectories are shown in Fig. 2.6B. Analysis of this dataset by SMAUG indicated a most probable interpretation of a $K = 3$-term model with diffusion coefficients of $D_i = \{0.006, 0.044, 0.368\}\mu m^2/s$ and weight fractions of $\pi_i = \{0.18, 0.53, 0.29\}$(Fig. 2.6 C&D, Table 2.5). The combined dataset for TcpP-PAmCherry trajectories results from four days of experiments in 111 cells. We then created 100 independent analysis runs using random sampling with replacement of the entire *V. cholerae* dataset of tracks and found that $K = 3$ was by far the most likely outcome (77 of the runs returned a 3-state model as the most likely (Table 2.5)).

Fig. 2.6D summarizes the key results for our measurements of TcpP-PAmCherry mobility: we observe three distinct mobility states, which we attribute to different binding states of the protein. Of the three states identified in our experiments, the intermediate state ($D_2 = 0.044\mu m^2/s$; red circle in Fig. 2.6D) is the most highly occupied state ($\pi_2 = 0.53$). TcpP exists in the membrane as either a monomer or a dimer [55], and we propose that the fastest diffusive state is free monomeric or dimeric TcpP-PAmCherry (yellow circle in Fig. 2.6C). We further hypothesize that TcpP association with other proteins in the membrane-most importantly its interaction partners, TcpH and ToxR-leads to its scanning the DNA for its binding target in the *toxT* promoter [26, 42]. We propose

Parameter	SMAUG Results	Parameter	Bootstrap Results*
Number of States	3	Number of runs with $K = 3$	77%
Diffusion Coefficient ($\mu m^2/s$)	0.0054, 0.046, 0.383	Mean Diffusion Coefficient ($\mu m^2/s$)	0.0053, 0.043, 0.355
Standard Deviation	0.0009, 0.0025, 0.0168	Standard Deviation	0.0011, 0.0043, 0.0181
Weight Fractions	0.193, 0.537, 0.270	Mean Weight Fractions	0.171, 0.540, 0.289
Standard Deviation	0.0095, 0.0094, 0.0079	Standard Deviation	0.0195, 0.0164, 0.0133
Transition Matrix	$\begin{pmatrix} .839 & .122 & .039 \\ .046 & .810 & .145 \\ .025 & .291 & .684 \end{pmatrix}$	Mean Transition Matrix	$\begin{pmatrix} .872 & .108 & .020 \\ .042 & .799 & .159 \\ .016 & .329 & .655 \end{pmatrix}$

Table 2.5: *V. cholerae* SMAUG analysis. Summary of measurements for TcpP-PAmCherry mobility in V. cholerae (Fig. 3). Total number of steps in the original dataset is 11,403. *Bootstrapping was performed over 100 rounds. 6% of analysis rounds yielded a 2-state model, 77% gave 3 states, 14% gave 4 states, and 3% gave 5 states. Mean and standard deviation values are constructed by using only the runs where $K = 3$. A new vector was created from the mean outputs of each of the 77 individual runs, and the mean and standard deviation of that vector are reported here.

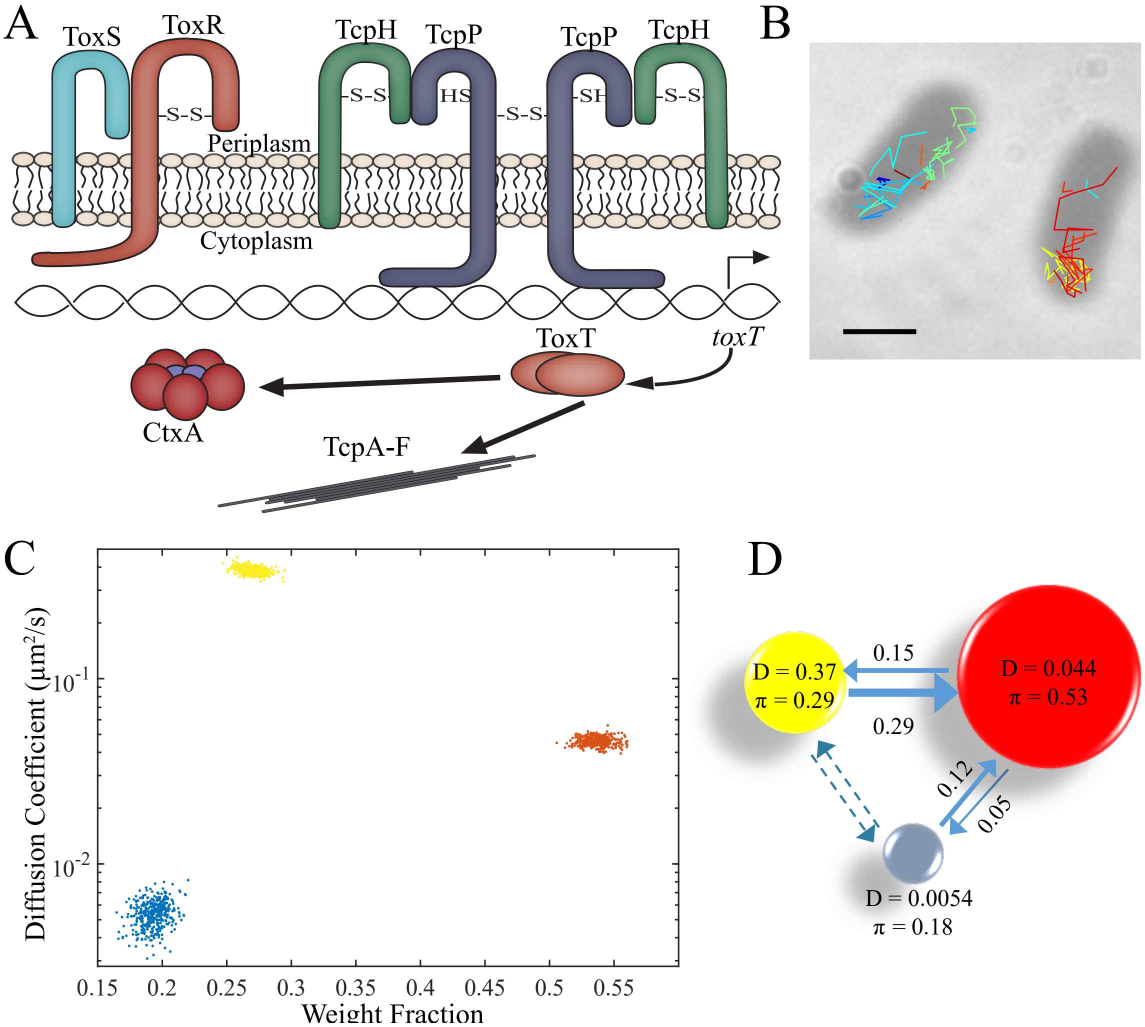


Figure 2.6: SMAUG analysis for bacterial imaging. A: Schematic of the *V. cholerae* virulence pathway. The membrane-bound protein TcpP binds the DNA directly along with other supporting proteins, leading to the hypothesis that the dynamics of TcpP reflect multiple mobility states. **B**: Representative image of individual TcpP-PAmCherry molecules trajectories inside live *V. cholerae* cells. Scale bar: 1 μm. **C**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis. SMAUG identifies three distinct clusters within the dataset. **D**: Cartoon depiction of the full SMAUG results for this dataset, including transition probabilities. Bubble colors correspond to the term colors in **C** and bubble sizes represent the weight fractions. Arrows between bubbles indicate the mean of the transition matrix elements for transitions between those terms. Dashed lines indicate transition probabilities that are negligible.

that the intermediate state is this protein complex DNA-searching state (red circle in Fig. 2.6D). Finally, to initiate the virulence cascade, this protein complex stops scanning and binds more tightly to the *toxT* promoter region of the DNA, and we propose the slowest

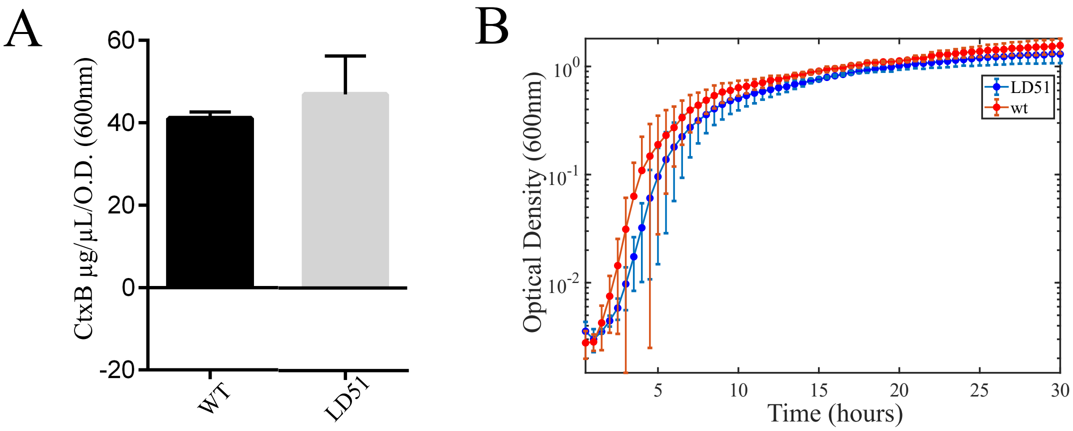


Figure 2.7: Biochemical validations of TcpP. A: CtxB levels in culture supernatants after 24 h in LB media (pH 6.5, 30 °C) show that LD51 expresses the same amount of CtxB protein as wild-type (WT) *V. cholerae* cells. **B**: Growth of WT and LD51 cells in LB (pH 6.5, 30 °C). OD600 nm values are an average of three biological replicates. Data courtesy of Lucas Demey

term (blue circle in Fig. 2.6D) is this promoter-bound state. Our model is further supported by the transition matrix (arrows in Fig. 2.6D and Table 2.5), which shows negligible transitions from the fastest to the slowest terms and instead outlines a path from the fastest state through the intermediate state to the slowest state, indicating that the TcpP monomers and dimers cannot directly bind the DNA, but rather that TcpP must form a complex with ToxR and/or TcpH before binding DNA and promoting *toxT* transcription. Testing these hypotheses to definitively assign the true nature of these identified mobility states will require further study (and is discussed in some detail in Chapter 4 of this Thesis). However, this analysis illustrates the utility of SMAUG for bacterial systems and provides a baseline to which future studies can be compared to more fully understand the mechanistic behavior of the *V. cholerae* virulence mechanism.

2.6.2 Application to antigen response in eukaryotic B cells

Finally, we applied our analysis method to investigate the dynamics of proteins involved in B-cell receptor (BCR) signaling. Situated in the plasma membrane of B cells, the

BCR recognizes and binds antigens, causing BCR clustering and initiating a downstream signaling pathway that results in BCR endocytosis and antigen processing. Following receptor clustering, the BCR is phosphorylated by the Src-family kinase Lyn [56], leading to recruitment and activation of the cytoplasmic kinase Syc which plays multiple roles in propagating the initial immune response. One target of Syc phosphorylation is the transmembrane adaptor protein LAB/LAT2, one of many proteins found within the BCR signalosome [54], a collection of proteins that localize, stabilize, and extend sites of BCR activation. Previously, it was found that membrane domains and lipid organization play a role in BCR activation by clustering BCR receptors upon antigen binding [43].

Using simultaneous two-color super resolution imaging, we analyzed the single-molecule trajectories of BCR and downstream protein Lyn or LAT2 at room temperature before and after stimulation by antigen addition [43] (Fig. 2.8 A-B). We split the trajectories into groupings of 1000 frames; each group contained on average 20,000 - 30,000 steps and occurred over 22 s, during which time frame we assume the dynamics do not change. In this way, we used SMAUG to analyze the evolution of the dynamics of the system over time. Before stimulation (Fig. 2.9C, left and Fig. 2.8D, first bar), the BCR dynamics are best described by three mobility states, with very little weight fraction in the slowest state (red). In other words, most BCR molecules are highly mobile. Immediately after stimulation (Fig. 2.8D, second bar), SMAUG finds four mobility states: the intermediate term is split into two mobility terms (brown and yellow). This finding may indicate a transition shortly after stimulation. Quickly, SMAUG returns only two mobility states, one of which is not observed pre-stimulation (blue) which we attribute to a new physiological state (Fig. 2.9). The most mobile terms have disappeared from the analysis as the system responds to antigen stimulation. This slower collection of mobility states persists for several minutes until the end of the measurement.

Simultaneously, we monitored the dynamics of Lyn or LAT2, and we matched the dynamics of the downstream protein with the response from the BCR itself. Analysis of LAT2 indicates four mobility states whose dynamics change greatly after BCR stimulation (Fig. 2.8E). Like BCR, the LAT2 dynamics slow over time post-stimulation: the slower LAT2 mobility state's population fraction increases and the faster LAT2 mobility stateâĂŹs population fraction decreases after stimulation. In contrast, for Lyn, a tyrosine kinase and the first protein in the downstream cascade, analysis with SMAUG consistently returns a three-term model with similar weight fractions and diffusion coefficients before and after BCR stimulation (Fig. 2.8F), with a slight change in the weight of the middle term occurring at 45 seconds and persisting through the end of the measurement. Consistent with this mobility analysis, we find that LAT2 colocalizes much more strongly with cross-linked BCR than does Lyn. A second analysis on different cells returns very similar results to those described above (Fig. 2.10). More studies are needed to assign biochemical and biophysical roles to the states uncovered by SMAUG, but this experiment proves the efficacy and utility of SMAUG analysis for both eukaryotic systems as well as for time series data.

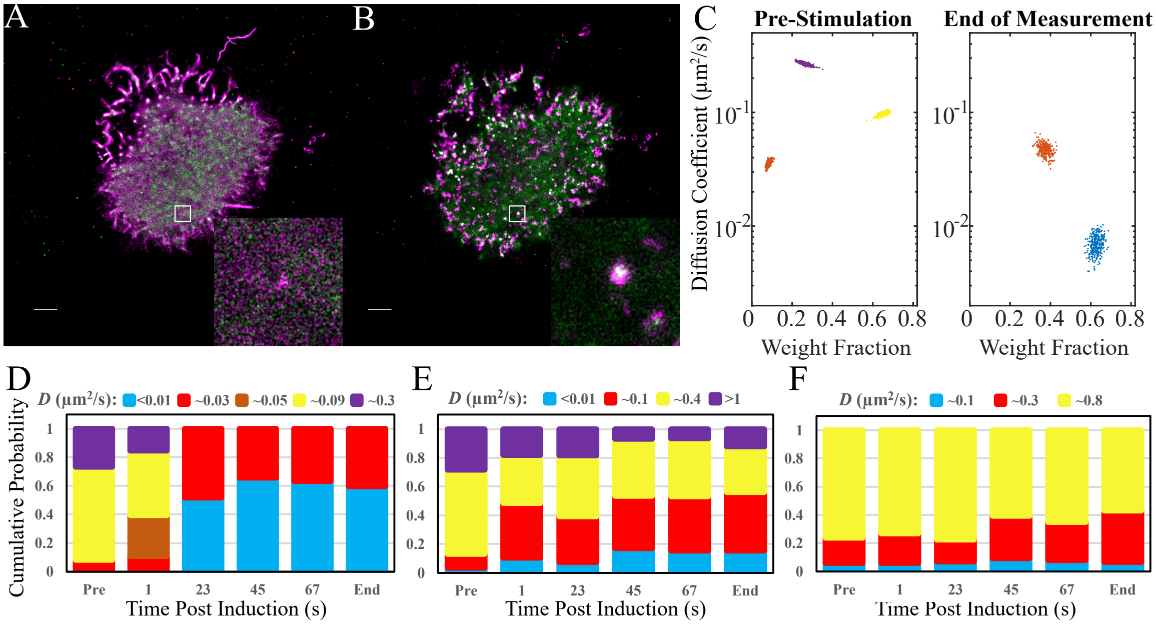


Figure 2.8: SMAUG analysis for single-molecule motion in a eukaryotic system. A: Super-resolution reconstruction image of BCR-SiR (magenta) and LAT2-mEos3.2 (green) in a representative B cell pre-stimulation. White: overlapping magenta and green signals. Inset is a higher resolution reconstruction of the 1.5 μm by 1.5 μm white boxed region. Scale bar: 2 μm. **B**: Super-resolution image of the cell in **A** 12.8 min post-stimulation. White: overlapping magenta and green signals. Inset shows same 1.5 μm by 1.5 μm white boxed region as in **A** at a higher resolution. Scale bar: 2 μm. **C**: Diffusion coefficient and weight fraction estimates for BCR molecules pre-stimulation and at the end of the measurement. Three distinct clusters are found pre-stimulation, but only two at the end of the measurement. **D**: Bar graphs showing the mean weight fraction of each identified state as a function of time for the BCR dataset. The bars labeled "Pre" and "End" correspond to the data in **C**. All other bars are labeled with the time post-stimulation. Identified mobility states are states whose estimates overlap in diffusion coefficient and weight fraction. A new, slower state (blue) emerges 23 s after antigen stimulation. **E**: The bar graphs for the weight fractions of LAT2 states over time show that the slowest mobility states (blue and red) increase in weight fraction relative to the faster terms (yellow and purple) suggesting the assembly of the BCR signalosome. **F**: The bar graphs for the weight fractions of Lyn states over time show that there is no change upon antigen stimulation and a slight overall decrease in mobility of the system starting at 45 seconds. The full cluster analysis is in Fig. 2.9.

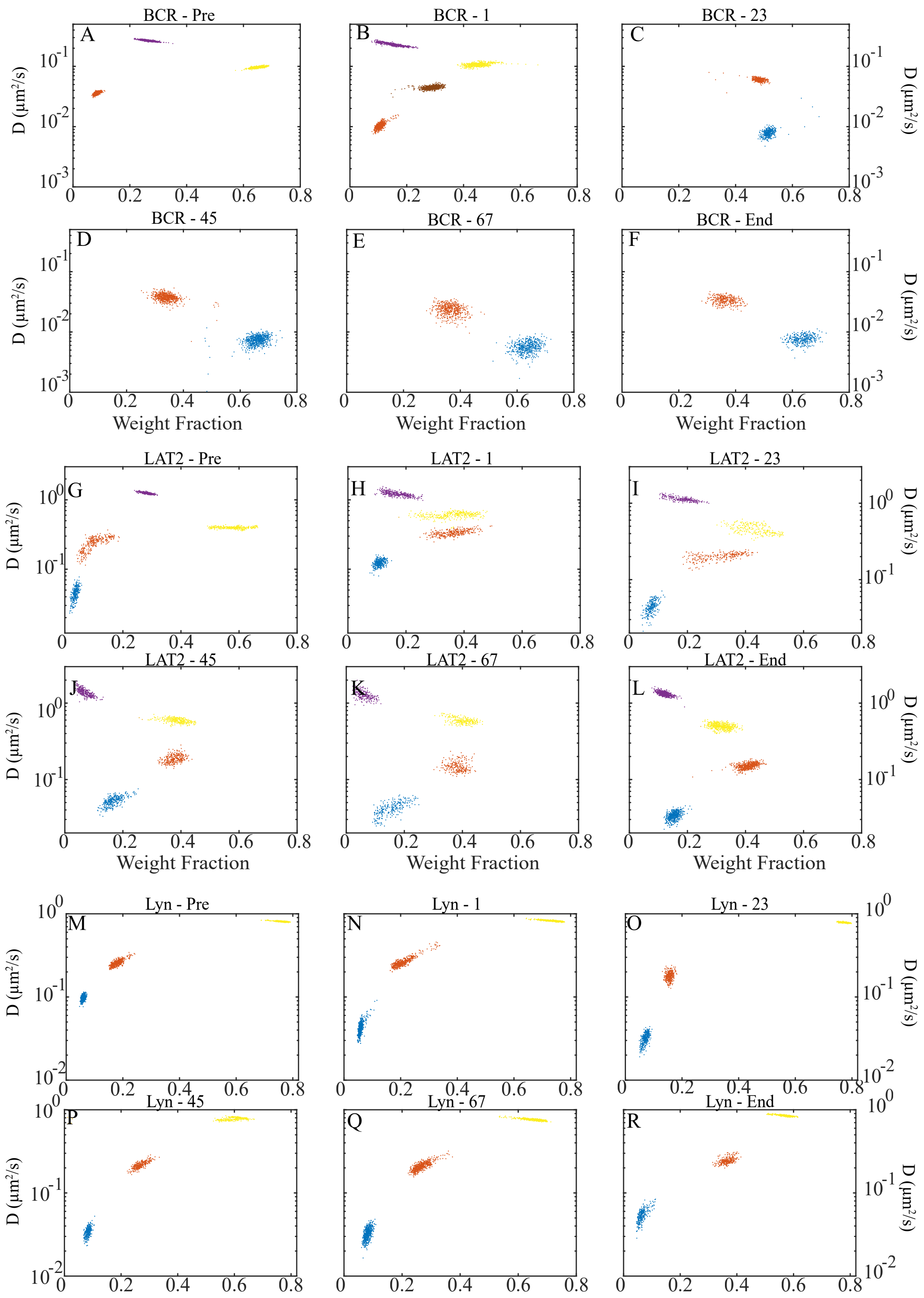

BCR - Pre
A
BCR - 1
B
BCR - 23
C
BCR - 45
D
BCR - 67
E
BCR - End
F
D (μm²/s)
10-1
10-2
10-3
0
0.2
0.4
0.6
0.8
Weight Fraction
LAT2 - Pre
G
LAT2 - 1
H
LAT2 - 23
I
LAT2 - 45
J
LAT2 - 67
K
LAT2 - End
L
100
10-1
Lyn - Pre
M
Lyn - 1
N
Lyn - 23
O
Lyn - 45
P
Lyn - 67
Q
Lyn - End
R
10-2

Figure 2.9

Figure 2.9 ***(previous page)*****: Full cluster analysis for the BCR, LAT2 and LYN molecules. A-F**: SMAUG analysis of the diffusion coefficients and weight fractions for BCR-SiR. **A** corresponds to the 'Pre' bar in Fig. 2.8D, **B-E** correspond to the middle bars, and '**F**' corresponds to 'End'. **A** and **F** are the same as in Fig. 2.8C. **G-L**: SMAUG analysis of the diffusion coefficient and weight fraction for the LAT2. **G** corresponds to the 'Pre' bar in Fig. 2.8E, **H-K** correspond to the middle bars and **L** corresponds to 'End'. **M-R**: SMAUG analysis of the diffusion coefficient and weight fraction for LYN. **M** corresponds to the 'Pre' bar in Figure 4F, **N-Q** correspond to the middle bars and **R** corresponds to 'End'.

2.7 Conclusions

Single-molecule experimental techniques have greatly enhanced the field of biophysics and our understanding of many biological problems. However, as SPT experiments are extended to include more complex systems, the need for a mathematically rigorous analysis method has increased. The SMAUG method we developed in this paper allows completely hands-free analysis of single-molecule tracking data by using a nonparametric Bayesian approach to fully characterize the posterior distributions of many of the relevant parameters and enables us to quantify the corresponding parameter uncertainties. This method allows more concrete and objective conclusions to be drawn from SPT experiments as it bypasses the issues of supervisory bias and model selection that can alter the data processing and the conclusions drawn.

However, certain limitations do exist for this method, both in the assumptions that the method makes and in the ability of SMAUG to resolve the dynamics of the system. The SMAUG algorithm, as mentioned before, assumes free diffusion. While we believe this is a reasonable assumption for the cases presented in this chapter, it could possibly bias results in systems where this is not the case and there is some active confinement or trafficking of the protein of interest. Effects from supra- or sub-diffusive behavior will, however, be mitigated somewhat because we use step sizes of 1 imaging frame where those effects are less pronounced as deviations from free diffusion are more pronounced

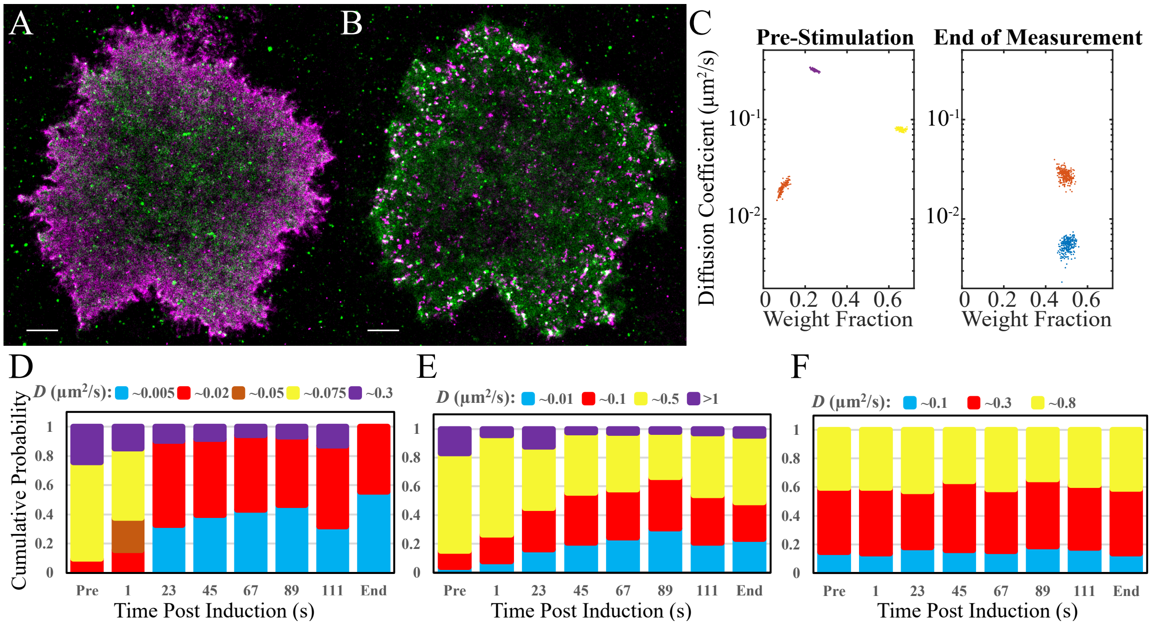


Figure 2.10: SMAUG analysis for single-molecule motion in a second B cell. A: Super-resolution reconstruction image of BCR-SiR (magenta) and LAT2-mEos3.2 (green) in a representative B cell pre-stimulation. Scale bar: 2 μm. **B**: Super-resolution image of the cell in **A** 12.8 min post-stimulation. Scale bar: 2 μm. **C**: Diffusion coefficient and weight fraction estimates for BCR molecules pre-stimulation and at the end of the measurement. **D**: Bar graphs showing the mean weight fraction of each identified state as a function of time for the BCR dataset. The bars labeled 'Pre' and 'End' correspond to the data in **C**. All other bars are labeled with the time post-stimulation. Identified mobility states are states whose estimates overlap in diffusion coefficient and weight fraction. **E**: The bar graphs for the weight fractions of LAT2. **F**: The bar graphs for the weight fractions of Lyn. Analysis shows similar results to the cells used in Fig. 2.8.

at longer time intervals. An analysis method by the Bathe lab can be used to test if the dataset resembles free diffusion [57]. Another pair of assumptions that go into the SMAUG algorithm is that the system under investigation is at equilibrium on the timescale of the experiment and that the dynamics of the biomolecules involved are in a steady state. If the system is undergoing rapid change that will alter the states, the SMAUG algorithm will still find the most probable parameters to describe the system, though that will likely be some amalgam or average of all the states present in the dataset. We overcame this limitation in the B-cell experiments (2.6.2) by analyzing small subsets of the total experiments, over timescales during which we assume the dynamics are roughly steady. Finally, SMAUG assumes that the diffusive states are indeed separate and distinct. If the dynamics underlying the dataset are drawn from a distribution whose component parameters overlap significantly, SMAUG's use of the slice sampler method will cluster in the data in a way that does not represent the underlying distribution. However, if given sufficient data to sample the entire continuous range of values well, it is probable that SMAUG would uncover a sufficiently large number of diffusive clusters that the resulting conclusions could still be useful in a biological sense. This brings up the point that SMAUG is dependent on the number of data points it has to work with. For the analysis in this chapter, and throughout this Thesis, we aim to surpass ten thousand steps for each dataset analyzed. This is because the number of resolvable states is proportional to the square root of the total data points and we have found that above ten thousand is sufficient to uncover the dynamics involved in our sorts of experiments where we expect less than 10 distinct diffusive states.

In this chapter, I began by outlining the theory behind constructing a nonparametric Bayesian analysis method as well as the specifics of constructing one for use in analyzing SPT data. I then used various realistic simulations and *in vitro* experiments to validate the accuracy and precision of the SMAUG method. Using the ability of this method to uncover hidden information, we then investigated the dynamics of two biological systems, one prokaryotic and one eukaryotic. SMAUG uncovered three distinct states for the dynamics of TcpP in wild-type *V. cholerae* cells. When SMAUG was applied to the dynamics of the BCR, it uncovered four total states for the BCR, one of which was not present pre-stimulation and two of which disappeared post-stimulation. Additionally, SMAUG found that the system rapidly shifted to from its pre-stimulation dynamics to its final state, the whole conversion taking less than 30 seconds total. The final state then lasted for several minutes. Similarly, it was found that the dynamics of LAT2 began with 3 distinct states and added a slower diffusing fourth state in concert with the shift of the BCR. Somewhat surprisingly, Lyn dynamics seem unaffected by alterations in BCR dynamics, indicating that any action of Lyn on BCR must occur within roughly ten secons or so and then resume its pre-stimulated dynamics.

As the work in this chapter demonstrates, the SMAUG analysis method provides researchers with a powerful tool for analyzing SPT experiments. Crucially, it removes supervisory biases while not sacrificing accuracy and precision in the estimation of the underlying dynamics under investigation. SMAUG allows researchers to draw concrete conclusions of the dynamics of biomolecules that can, in turn, provide mechanistic or biochemical insight into the environment or behavior of biomolecules in many systems relevant to cellular biology.

CHAPTER III

Investigating the Hidden Dynamics of Epigenetic Silencing in the Yeast Model System *S. pombe*

The work presented in the chapter is in preparation to be for publication

Karslake, J.D.*, Biswas, S.*, Biteen, J.S., and Ragunathan, K. Investigating the Hidden Dynamics of Epigenetic Silencing in the Yeast Model System *S. pombe. in prep* [1]

3.1 Introduction

Cells with the same genotype can display distinct phenotypes and gene expression patterns. This ability is the basis of all multicellular life, as cells take on specific roles but have no changes in the underlying sequence of their DNA. These patterns of gene expression that persist despite the removal of the initial signal and regardless of the external environmental are referred to as epigenetic states [58]. Epigenetic states have been shown to persist through many cell cycles and are inherited along with the genome itself [59]. Epigenetic states also play a role in single-cell organisms as well, even where

[1] **Author contributions** - J.D.K. and S.B. contributed equally to this work. All authors designed the research. J.D.K and S.B. preformed the live cell microscopy experiments and analyzed the results. J.D.K designed the image analysis algorithms. S.B. and K.R. constructed the yeast strains and preformed the biochemical assays.

cell specialization/differentiation into different cell types are not present. Daughter cells from a mitotic event might display different gene expression patterns from the mother cell despite having an identical genome. While this difference could arise from a random or unequal distribution of transcription factors during division, it can also arise from a change in the covalently bonded, post-transcriptional modifications that are attached to either the DNA itself or to various DNA-associated proteins. This collection of modifications is termed the "epigenome" of the organism, as it is another layer on top of the DNA sequence itself that helps determine gene expression. A phenotypic change in a cell that was the result of the random distribution process would not be expected to persist over the long term through many cell divisions and thus it would not be considered an epigenetic shift. However, directed changes in the post-translational modifications of DNA associated proteins or post-replicative modifications of the DNA itself are long-lived epigenetic changes.

In 1928, Heitz discovered that parts of the DNA of a moss species remain visible by staining even after mitosis has completed. When the cell was in interphase, most of the stained genetic material disappeared but a fraction stayed visible throughout the entire cell cycle [60]. He called the regions that remain visible "*heterochromatin*" because they were somehow different from the other, "*euchromatin*", regions though he did not know how or why. Later studies on other organisms showed that these heterochromatin regions were genetically silenced, and if a usually transcriptionally active gene were transposed near to these regions it could be silenced as well [61,62]. We now know that these regions are highly enriched for the types of post-replicative modifications to the DNA and/or post-translational modification of DNA-associated proteins mentioned above and discussed in more detail in the next section. Heterochromatin formation is highly correlated with the establishment of post-translational modifications that silence gene activity and the silencing of particular genes helps promote distinct phenotypic states that persist through generations [63, 64].

In this chapter, I discuss my investigation of one of the epigenetic gene-silencing pathways using super-resolution microscopy and single-particle tracking (SPT). To begin, I will discuss the relevant biological background and reasons for using single-molecule methods to probe this system. In the following sections, I will present the results of the dynamics of this system and how genetic mutations perturb the system in illuminating ways. Finally, I will conclude with some thoughts on the impact and reach of this study.

3.2 Biological Background

Regions of heterochromatin have been shown to have several general properties. In addition to the aforementioned gene silencing effect, heterochromatin regions are also generally sequestered to the nuclear periphery, replicate late in the cell cycle, and recombine less frequently than euchromatic regions [63]. The functional unit of heterochromatin is the nucleosome, which contains an octamer of histone proteins (dimers of H2A, H2B, H3, and H4) along with 147 base pairs of DNA which wrap around the histone octomer core [65]. The N-terminal region of some of these histones have accessible residues, especially lysine residues, that can be modified with various adapter molecules of which there exist at least 8 known classes [66, 67]. Acetylation of lysines generally corresponds to an increase in gene transcription whereas methlyation of histone lysines is generally correlated with increased silencing. Methylation of the lysine 9 residue (K9) of the H3 histone (H3K9me) is an especially important marker for the formation of heterochromatin as disruption of this residue results in the removal of silencing [68, 69].

The HP1 family of proteins recognizes the H3K9 methlyation mark specifically and is the key regulator of the formation of the heterochromatin-associated silencing complexes such as RITS and RDRC which are known to extend heterochromatin domains and silence gene expressions [70]. The model organism of fission yeast, *Schizosaccharomyces pombe*, lacks any other type of post-transcriptional epigenetic modifications and so is a good system in which to study the behavior of histone-mediated epigenetic silencing of genes without compounding factors, as *S. pombe* does not undergo direct DNA methylation. The homolog of HP1 in *S. pombe* is **Swi6**. Swi6 has no enzymatic activity of any kind but contains two distinct domains and a flexible hinge region between them directly related to its function as a recognition marker and scaffold. The chromodomain (CD) recognizes the H3K9me mark of the histone and can bind directly to it. The chromoshadow domain (CSD) is responsible for the dimerization of two Swi6 proteins into the functional unit and also has interaction areas that are needed to recruit other chromatin associated proteins [70]. Attached to the CD domain is the conserved ARK loop which, in dimeric unbound Swi6, interacts with the ARK loop on the other subunit of the dimer and auto-inhibits its binding to any other Swi6 molecules. Upon binding to the histone, however, these loops are displaced and become accessible for interactions with other Swi6 dimers (Fig. 3.1A). Interaction between neighboring Swi6 ARK loops into extended chains is thought to help extend regions of heterochromatin [71]. The hinge region is a flexible and highly positively charged region between the CD and the CSD. Less is known about the function of the hinge region between these domains, but recent work has suggested that this region is involved binding to DNA and/or nascent mRNA. Being highly positively charged, it is thought to trap and newly made RNA chains from silenced regions and escort these strands to degradation nearby complexes such as RITS and RDRC can cut and break these strands apart [72] (Fig. 3.1B).

The binding of Swi6 to the H3K9me mark is reversible and highly dynamic, it has been shown to occur on the timescale of milliseconds to seconds [73]. The bound Swi6 molecule then recruits other remodelers to the site in order to extend the region of heterochromatin. Clr4, a histone methlytransferase, is responisible for making the H3K9me mark and its recruitment to regions of heterchromatin extends the domain by methylating other nearby histones (Fig. 3.1C). Eventually, the presence of the remodelers recruits the protein Epe1 to the heterochromatin. Epe1, a demethylase, acts in opposition to Clr4 by removing the H3K9me mark from histones, thereby removing the Swi6 binding site

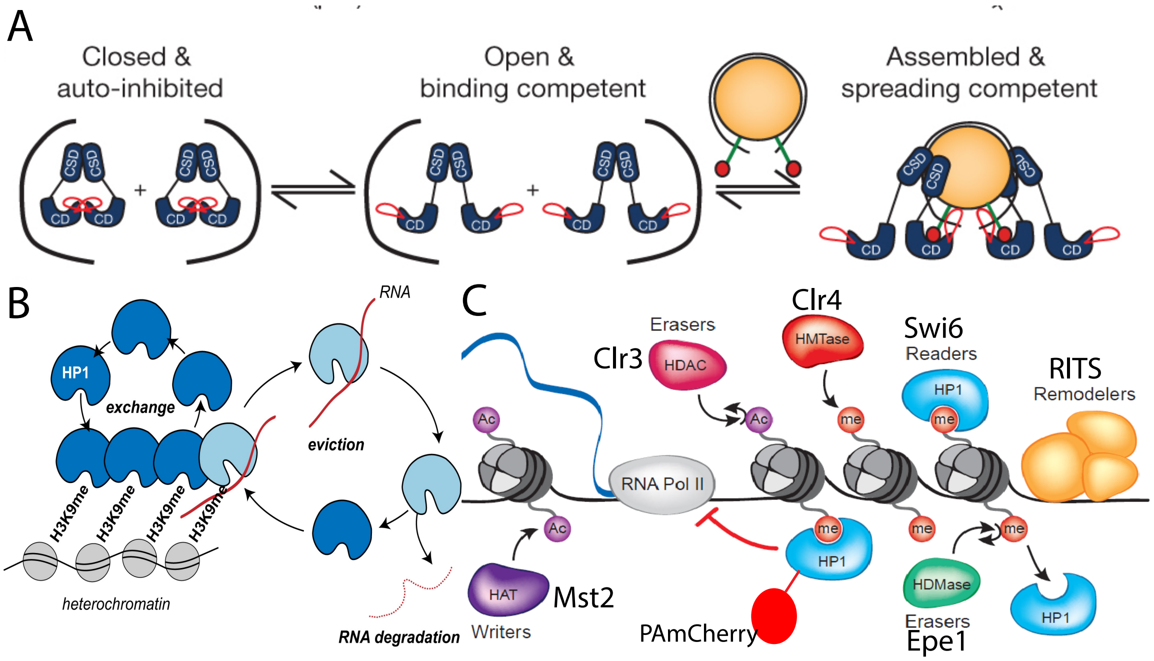


Figure 3.1: Swi6 structure and pathway. A: Schematic of the functional roles of Swi6. The CD binds to the H3K9me mark, the CSD is involved in the dimerization of two Swi6 molecules. These two domains are connected by a flexible linker region and the ARK loop (red) is thought to mediate higher order structures. Figure from [71]. **B**: The hinge region of Swi6 is thought to help keep silenced genes from being expressed by trapping newly made RNA chains and escorting them to degradation. Figure from [72]. **C**: Cartoon depiction of the histone modification machinery. Swi6 has been tagged with the fluorescent protein PAmCherry for tracking.

and contracting zones of heterochromatin to stop them from spreading. The interplay of these two proteins, Clr4 and Epe1, in extending and constricting regions of heterochromatin is further complicated by the effect of two other proteins, Mst2 and Clr3 (Fig. 3.1C). Mst2 and Clr3 are a part of the histone acetylation pathway, which promotes antisilencing and increased gene expression. Mst2 is the histone acetyltransferase and the Clr3 is the deacetylase. Thus Mst2 increases antisilencing by extending regions of euchromatin and Clr3 promotes silencing by cutting back regions of increased expression [70]. At any histone site, the balance of these opposing outcomes from the proteins involved will determine the functional state of the genes nearby (Fig. 3.1C).

In such a dynamic and complex system, several questions emerge. First, how are epigenetic states maintained over generations, which can range from hours to days, if the system is in such a constant flux? Second, what are the functional roles and steps inside the system that lead to these outcomes? To answer these questions, we need to identify distinct biochemical states within a complex system. Using super-resolution microscopy and SPT, along with an analysis method such as Single-Molecule Analysis by Unsupervised Gibbs (SMAUG) (Chapter II), I was able to look directly at the dynamics of this system by observing individual molecules inside living cells and uncover distinct biological states that are hidden or averaged over using other methods.

3.3 Methods

Strain Construction

Endogenous copy of Swi6 was deleted and replaced with natamycin antibiotic. PAmcherry was tagged at the N-term of Swi6 by Gibson cloning method. The strain containing PAmcherry-Swi6 was made by insertion of fluorophore tagged Swi6 at the *ura4+* locus using PCR-based gene targeting approach. In strains where WT copy of swi6 is intact, mNeongreen tagged Swi6 is inserted at *leu1+* locus. The deletions of the various RNAi and chromatin components were achieved by PCR-based gene-targeting approaches. To select colonies harboring the reporter gene PCR-based screening approach was applied. Mutation at different position of protein under investigation is done by site directed mutagenesis approach.

Microscopy experiments

S. pombe cells containing a chromosomal fusion of the photoactivatable red fluorescent protein, PAmCherry, to Swi6 as the sole source of Swi6, except for strains containing extra copies of Swi6 which inserted into the *trp1* locus. These cells were grown in Yeast Extract with Supplments (YES) media [74] at 25 °C for 2 days to mid-exponential phase. The cells were then diluted and allowed to grow for another 3 hours. The cells are then harvested by centifugation and the media exchanged for minimal media with n-propal galate added and plated on an agarose pad. Imaging occured at room temperature and only cells in G2 phase were selected. Phase-contrast images were obtained before each imaging experiment using a phase-condenser to illuminate the sample. Cells are imaged using a 406-nm laser (Coherent Cube 405-100; 102 W/cm^2) for photo-activation and a 561-nm laser (Coherent-Sapphire 561-50; 163 W/cm^2) for imaging. Continual images were collected with a 40-ms exposure time per frame in an Olympus IX71 inverted epi-fluorescence microscope with a 100x 1.40 NA oil-immersion objective. The fluorescence emission was filtered with appropriate filters and imaged on a 512 by 512 pixel Photometrics Evolve electron multiplying charge-coupled device (EMCCD) camera. Protein copy number experiments were carried out using a custom built MATLAB code. By hand, we identified single mCherry molecules at the end of movies, calculated the total intensity for several such molecules, and averaged them. We divided the intensity of the total fluorescence in the nucleus by this average value to roughly estimate the protein copy number. Recorded single-molecule positions were detected and localized as previously described using home-built code [27], and connected into trajectories using the Hungarian algorithm [24].

Data Analysis

Trajectory analysis was performed using home built MATLAB code, called SMAUG, that embeds a Hidden Markov Model embedded inside a Gibbs Sampler to estimate the parameters of interest in a data set(Chapter II, [75]). All trajectories for a given condition were bundled together and analyzed. Briefly, the SMAUG algorithm takes in a data set of trajectories and estimates the parameters of interest using a Bayesian statistical framework. Data is sorted probabilisticly into terms and then used to refine parameter estimates for that term and then the process is repeated. Over time, this iterative process converges onto the most likely values for the parameter estimates given the input data set. Parameters of interest from the data set are the number of mobility states within the set, the diffusion coefficient, weight fraction,and transition probabilities between the states.

3.4 Swi6 Motion in Wild-type Cells Display Complex Dynamics

To begin our investigations into this system, we first constructed a strain which contained a genetically-encoded, fluorescently labeled copy of Swi6 tagged at the N-terminus with the fluorescent protein (FP) mCherry at the endogenous *ura4* locus. These cells were shown to have normal cellular morphology under phase-contrast imaging (Fig. 3.2A). Imaging these cells under 561 nm excitation we detected a region of fluorescent signal from the nucleus containing a variable number of distinct spots (Fig. 3.2B). This fusion protein was shown to be functional, as seen by the silencing of a reporter gene inserted in the chromosome (Fig. 3.1C). *ura4* encodes a genes for Orotidine 5'-phosphate decarboxylase (ODCase), an enzyme that catalyzes one reaction in the synthesis of pyrimidine ribonucleotides. If 5-Fluoroorotic acid (FOA) is added to the media, the active ODCase will convert FOA into the toxic compound 5-fluorouracil, a suicide inhibitor of translation, causing cell death and allows for selection against yeast carrying the gene. Lanes 1 and 6 in Fig. 3.2C contain wildtype cells, in which this locus is silenced, and so cells grow normally when plated on media containing FOA. Lanes 2 and 5 contain cells which harbor the *clr4*Δ deletion, which removes the H3K9me mark and so the ODCase gene will be expressed, causing cells plated with FOA to die. Lanes 3 and 4 contain cells with the PAmCherry-Swi6 fusion as the only Swi6 in the cell and grow on media containing FOA, indicating that the gene remains silenced and the fusion protein is functional (Fig. 3.1C). Using our fluorescence setup, we captured images (such as in Fig. 3.2B) and, using a custom built MATLAB script, manually identified the nucleus region inside. The total intensity of the identified nuclear region was divided by the average integrated intensity of a single mCherry molecule to roughly estimate the number of Swi6 fusion proteins inside the cell. Total cells counted in this manner was 44 and the mean number of Swi6 molecules detected was 330, with a standard deviation of 130 molecules.

Next, to investigate the dynamics of the system, we constructed a new strain where Swi6 was tagged with the phtotoactivatable FP PAmCherry [11], still at the N-terminus. This protein is initially in a dark state and, by tuning the power of the excitation laser, we can stochastically activate a few molecules at a time into a fluorescent state. We can then use repeated rounds of this activation cycle to track single molecules inside the nucleus. We imaged these cells under the microscope collecting 19,273 steps from 2971 trajectories. Some representative trajectories are shown in Fig. 3.3A. The biochemical environment of Swi6 inside the cell should be directly related to its motion and so by identifying the different types of motion described by the trajectories of the molecules, we can learn about the behavior of the protein inside the cell. Of particular interest is: 1) information about the number of these distinct states of motion that exist, which I refer to as the "diffusive model" of the system; 2) what the value of the diffusion coefficient for each of these identified states in the diffusive model is along with the fraction of the total that each state represents and 3) what, if any, is the probability of transitioning between these states. Using our SMAUG Hidden Markov Model Gibbs sample (Chapter II, [75]) we use a nonparametric Bayesian probability framework to make estimates about the most probable values for these pieces of information in an iterative process. Briefly, SMAUG sorts each data point into a specific state by likelihood and then takes only the sorted data for each term to update the probability distribution that describes that term. In an

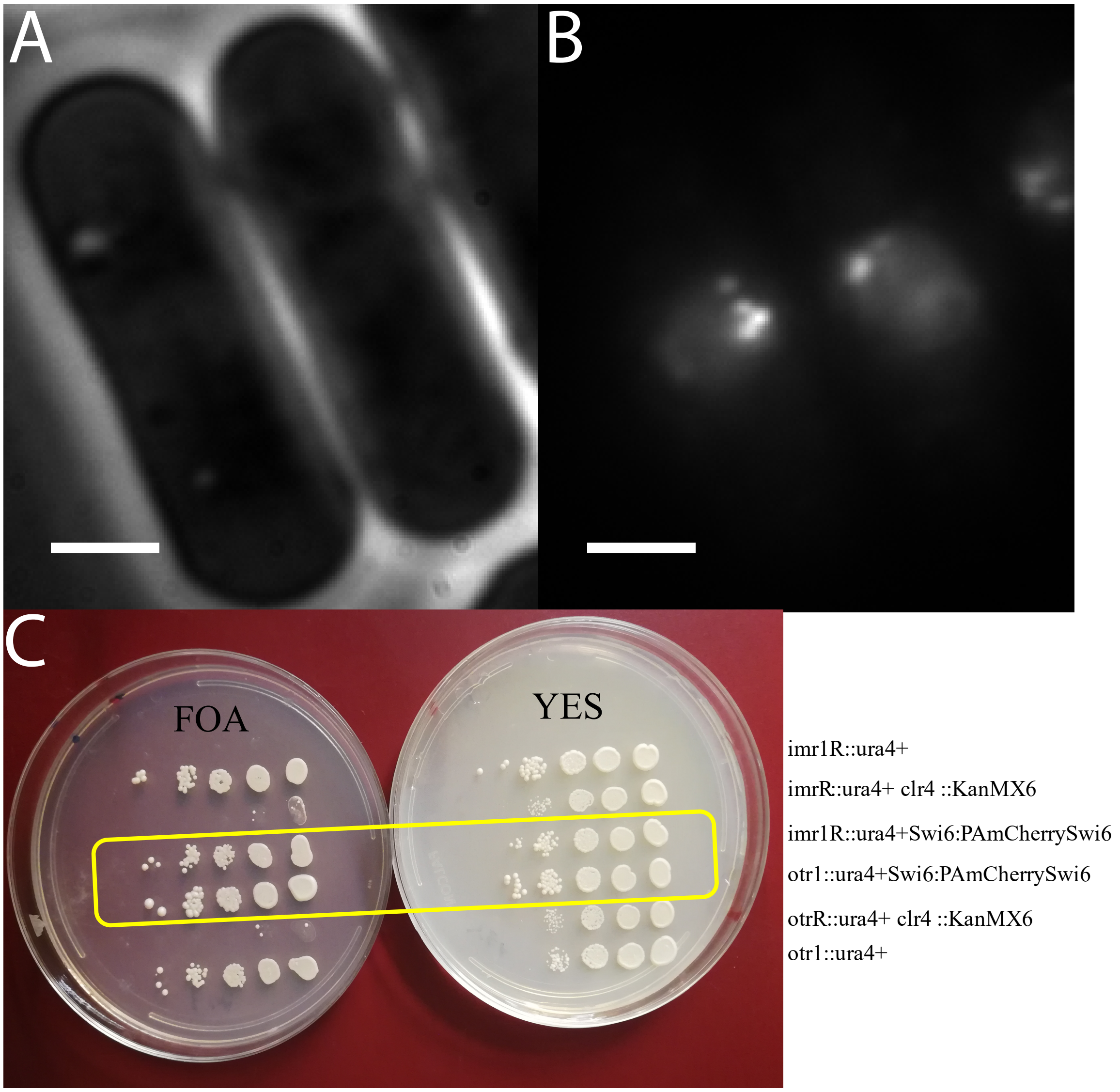

A
B
C
FOA
YES
imr1R::ura4+
imrR::ura4+ clr4 ::KanMX6
imr1R::ura4+Swi6:PAmCherrySwi6
otr1::ura4+Swi6:PAmCherrySwi6
otrR::ura4+ clr4 ::KanMX6
otr1::ura4+

Figure 3.2: Swi6 functionality. A: Phase-contrast image of representative *S. pombe* cells containing the fusion Swi6-mCherry and showing normal cell morphology. **B**: Same cells as in **A** imaged under 561 nm laser excitation showing fluorescence in the nuclei. **C**: Gene silencing assay that demonstrates the fusion Swi6 is functional. Cells plated on media containing 5-Fluoroorotic acid (FOA) will die if the usually silenced gene region *ura4* is no longer silenced. Lanes 1 and 6 contain wildtype cells that grow in both FOA containing media and rich media, indicating the gene is silenced. Lanes 2 and 5 contain strains that lack the ability to silence and thus these cells die when plated on media containing FOA. Lanes 3 and 4 are strains that contain a fusion PAmCherry-Swi6 and these cells live on media containing FOA, indicating the gene is silenced and the protein fusion is functional. Columns indicate serial 10-fold dilution of cells. Scale bars = 2 *μm*.

iterative process, these distributions are refined until the most likely parameter values are found. Crucially, SMAUG requires no user input or supervision and can identify the most likely parameters,including the diffusive model, without oversight.

Using the data set for the PAmCherry-Swi6, SMAUG identified the most probable diffusive model as having four distinct mobility states with diffusion coefficients of $D = \{0.007, 0.021, 0.081, 0.521\}$ $\mu m^2/s$ and with weight fractions of $\pi = \{0.23, 0.32, 0.20, 0.25\}$ (Fig. 3.3B and C) and with transition between states given by the matrix:

$$\text{T=}\begin{pmatrix} .85 & .13 & <.01 & <.01 \\ .09 & .78 & .08 & .04 \\ <.01 & .19 & .58 & .21 \\ <.01 & .07 & .16 & .75 \end{pmatrix}$$

where each value, T_{ij}, is the probability of transitioning from state i to state j. For example, the probability that a molecule in State 1 transitions into State 2 is 0.13 (Fig. 3.3).

We attribute these distinct states to various binding and biological roles of the protein. The most highly occupied state is the second slowest (red) state, which we hypothesis is a type of searching mechanism of the protein as it binds to histones looking for the H3K9me mark. We will refer to these states as State 1—4, with State 1 being the one with the slowest diffusion coefficient (blue in Fig 3.3B) and State 4 being the most mobile state (purple). For State 4, we propose that it is representing the freely diffusing dimer within the nucleus. State 1 we suggest is the bound state of the molecule with the Swi6 fusion bound to a H3K9-methylated histone and interacting with the various associated complexes. State 2 is hypothesized to be the Swi6 search process where the protein is transiently binding histones that lack the H3K9me mark. SMAUG analysis of these PAmCherry-Swi6 cells provides a baseline of dynamical processes for this system. We then perturbed the system using a variety of genetic mutations that are detailed below to better understand the role each of these states might be. A full list of the mutants utilized in this study is found in Tables 3.1, 3.2, and 3.3.

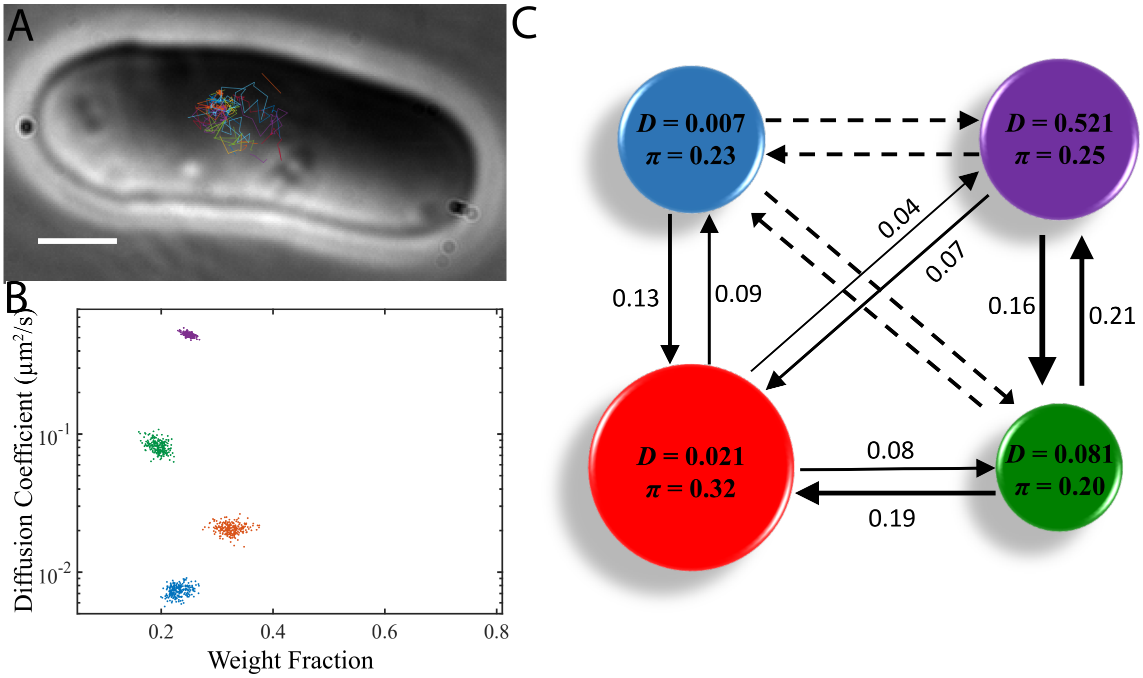


Figure 3.3: SMAUG analysis for the dynamics of PAmCherry-Swi6. A: Representative image of the trajectories of PAmCherry-Swi6 in one cell superimposed on top of the phase contrast image. All of the tracks are inside a volume determined to be the nucleus of the cell. Trajectory colors indicate separate tracks. Scale bar: $2\mu m$ **B**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis. SMAUG identifies four distinct clusters within the data set. **C**: Cartoon depiction of the full SMAUG results for this data set, including transition probabilities. Bubble colors correspond to the term colors in **B** and bubble sizes represent the weight fractions. Arrows between bubbles indicate the mean of the transition matrix elements for transitions between those terms. Dashed lines indicate transition probabilities that are negligible.

3.5 Swi6 Dynamics With Chromatin Remodeling Mutations Displays Altered Mobility

3.5.1 H3K9 Methylation Pathway Deletions

To discover how changing the heterochromatin state of the cell affects the dynamics of Swi6 and also uncover the biological roles of the states found in the wildtype analysis, we created several mutation strains (Table 3.1). We built strain *clr4*Δ, which lacks the H3K9me methyltransferase Clr4. In this strain, the deletion of the *clr4* gene removes the ability of the cell to methylate H3K9, an anti-silencing mutation that promotes more euchromatin and increased gene transcription. Consequently, with the Swi6 target removed we expected to observe a decrease or removal of State 1 (blue term in Fig. 3.3B,C) from the wildtype dynamics of Swi6. We imaged these *clr4*Δ cells to obtain a data set of 3,288 trajectories and 12,773 steps from 29 cells.

Accordingly, when these cells were analyzed using the SMAUG algorithm the most probable diffusion model was one with 3 states whose diffusion coefficients were $D = \{0.031,\ 0.104, 0.556\}\ \mu m^2/s$ and weight fractions $\pi = \{0.14, 0.25, 0.61\}$ (Fig. 3.4A). The lowest diffusive state from PAmCherry-Swi6 in otherwise wildtype cells has been removed and a state with diffusion value close to State 2 has been reduced in weight fraction (Fig. 3.4A, red). Additionally, the state with diffusion coefficient very similar to State 4 has been greatly increased in weight fraction and now represents the majority of the system (Fig. 3.4A, purple). This output from SMAUG supports the idea that State 1 is the state that represents Swi6 bound to the H3K9me target in regions of heterochromatin and with its removal a majority of the Swi6 is in a freely diffusive state throughout the nucleus. The transition matrix for this system still clearly shows a path for moving from highest to lowest diffusive state, indicating that while the H3K9me target is missing, the other biophysical states are present and the transitions between these states remain present (Fig. 3.4B).

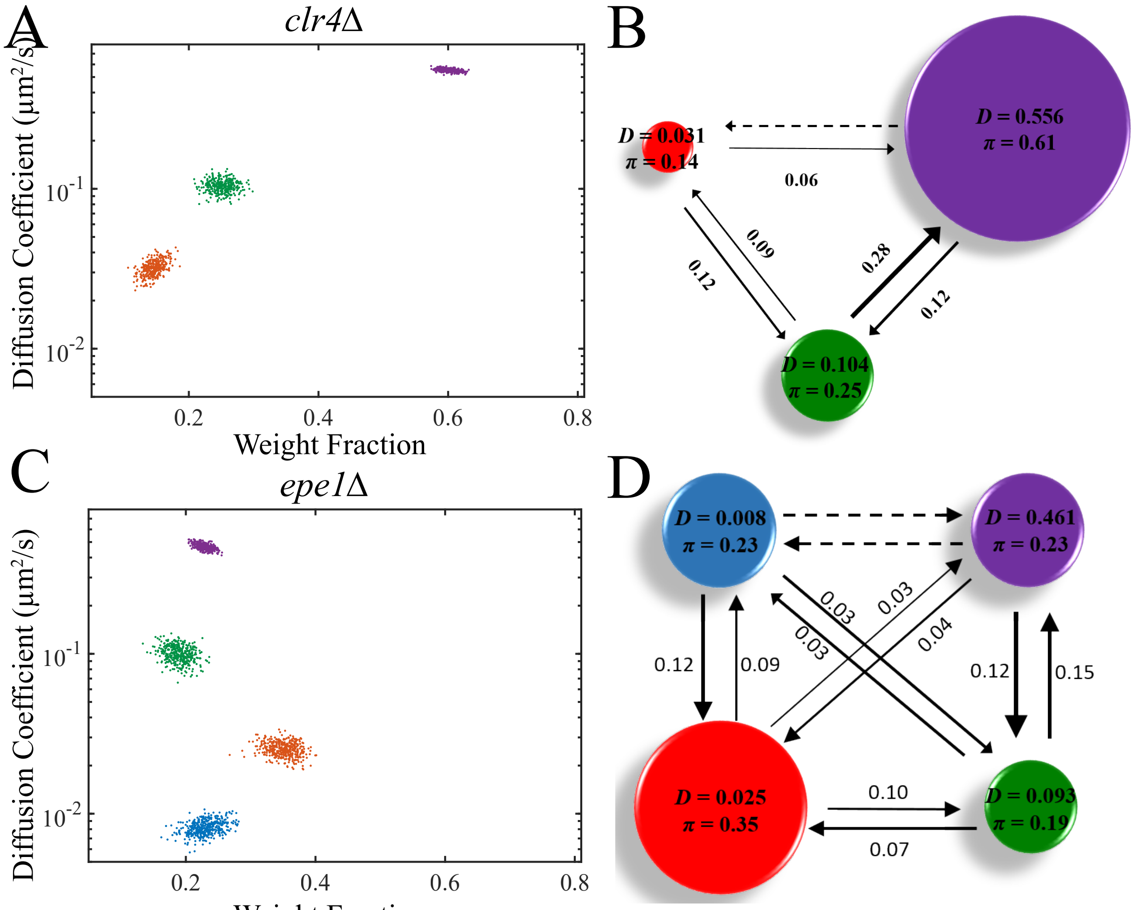


Figure 3.4: Dynamics of PAmCherry-Swi6 in *clr4*Δ and *epe1*Δ cells. A: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for *clr4*Δ. SMAUG identifies three distinct clusters within the data set. **B**: Cartoon depiction of the full SMAUG results for this data set, including transition probabilities.**C**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for *epe1*Δ. SMAUG identifies four distinct clusters within the data set. **D**: Cartoon depiction of the full SMAUG results for this data set. Bubble colors correspond to the term colors in **A** or **C** and bubble sizes represent the weight fractions. Arrows between bubbles indicate the mean of the transition matrix elements for transitions between those terms. Dashed lines indicate transition probabilities that are negligible.

Strain Name	Mutation	Description
*clr4*Δ	PAM-Swi6 + *clr4*Δ	Deletion of lysine histone methyltransferase Clr4 and PAmCherry-Swi6
*epe1*Δ	PAM-Swi6 + *epe1*Δ	Deletion of lysine histone demethylase Epe1 and PAmCherry-Swi6
*mst2*Δ	PAM-Swi6 + *mst2*Δ	Deletion of lysine histone acetyltransferase Mst2 and PAmCherry-Swi6
*clr3*Δ	PAM-Swi6 + *clr3*Δ	Deletion of lysine histone deacetylase Clr3 and PAmCherry-Swi6

Table 3.1: Strain list for the deletion analysis. PAM is the photoactivatable red FP PAmCherry and NG is the green FP mNeonGreen. All mutations are made at the endogenous *ura4* locus. FP and protein name in the description represents gene orientation with FP before the protein indicating an N-terminal fusion and after the protein being C-terminal.

To further understand the effect the H3K9me mark has on the dynamics of Swi6, we deleted the protein responsible for its removal, the lysine histone demethylase Epe1. *epe1*Δ cells should be unable to remove the H3K9me mark once it has been installed and *epe1*Δ is thus a pro-silencing mutation that extends regions of heterochromatin. We imaged this strain and obtained 1,176 trajectories and 12,790 steps. SMAUG analysis returns a 4 state model where the diffusion coefficients and weight fractions are all extremely similar to the dynamics of Swi6 in the wildtype cells (Fig. 3.3A). Mean diffusion coefficients for *epe1*Δ are $D = \{0.008, 0.025, 0.093, 0.475\}$ $\mu m^2/s$ and with weight fractions of $\pi = \{0.23, 0.35, 0.19, 0.23\}$ (Fig. 3.4C). The transition matrix elements are similar but show a slight shift towards a less dynamic system, as most of the transitions are less likely for the same transition than in the Swi6 dynamics in the wildtype cells (Fig. 3.4D). This evidence strongly suggests that the removal of the downstream protein Epe1 has no affect on the dynamics of Swi6. One possible explanation is that there might only a marginal increase in the amount of heterochromatin compared to strains containing Epe1 as the other machinery for epigenetic modification is still present and working to counteract any large-scale increase in heterochromatin formation.

3.5.2 H3K Acetylation Pathway Deletions

To further test how remodeling of heterochromatin affects the dynamics of Swi6, two further strains were created: *mst2*Δ and *clr3*Δ. Both of these proteins are involved in the the acetlyation of histones and thus are important regulators of euchromatin. Mst2 is one of two epigenetic acetyltranserases in the cell but is very specific for the H3K14 residue whereas Clr3 is a more general deacetlyase, though still lysine-specific. Thus, in the *mst2*Δ strain there is less euchromatin as one of the main promoters of histone acetlyation is removed. On the other hand, in the *clr3*Δ strain, there exists more euchromatin as once placed an acetyl group cannot be removed, which encourages euchromatin formation. Figure 3.5 shows the output of the SMAUG analysis for these strains. Both strains return a most probable four state model with coefficient estimates similar to those estimated for the four states identified for the Swi6 dynamics in wildtype cells. The *mst2*Δ data set contained 11,341 steps and the *clr3*Δ data set contained 9,049 steps. In *mst2*Δ, the mean of the diffusion coefficients were $D = \{0.007, 0.018, 0.096, 0.615\}$ $\mu m^2/s$ with mean weight fractions of $\pi = \{0.25, 0.39, 0.22, 0.14\}$ (Fig. 3.5A). In this analysis, State 2 (Fig. 3.5A, red) is increased and State 4 (Fig. 3.5A, purple) is decreased in weight fraction relative to the dynamics of Swi6 in wildtype cells. We hypothesized that State 2 was Swi6 searching histones for the H3K9me mark and, with the deletion of Mst2, there exist more "naked" histones, i.e. histone lacking in both the methyl and acetyl modifications, for Swi6 to search, which would help explain the increase of State 2 while State 1 remains constant. Additionally, there are increased transition probabilities in and out of the most mobile states, including a much larger 4 to 2 transition (Fig. 3.5B, green and purple), supporting the idea that much more Swi6 are moving out of these states and into the searching state (Fig. 3.5B, red) with the increase in the amount of histones available for binding (Fig. 3.5B).

In the case of removing Clr3 however, the amount of acetylated histones will increase and acetlyated histones have been shown to reduce Swi6 binding [76], either through steric interactions or some other method. SMAUG returns a most probable four state

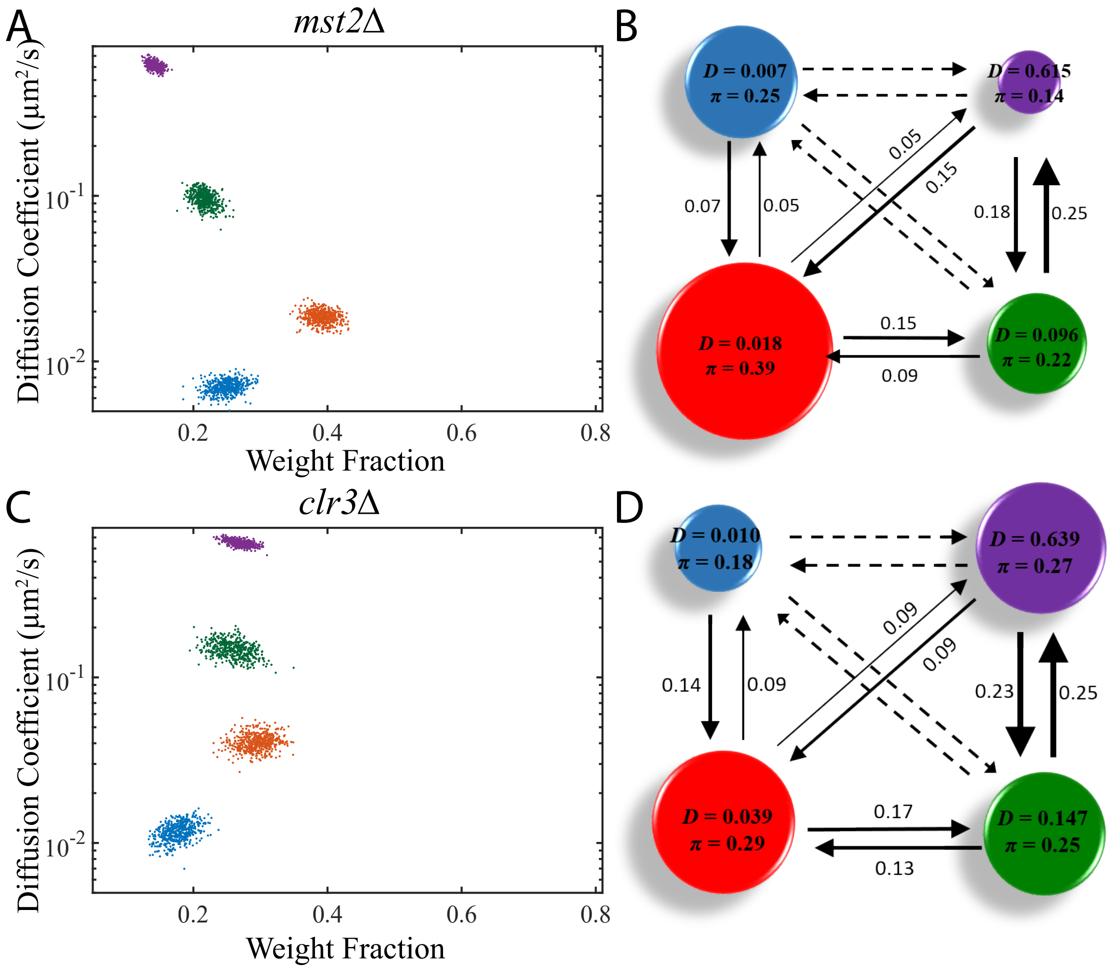


Figure 3.5: Dynamics of PAmCherry-Swi6 in *mst2*Δ and *clr3*Δ cells. A: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis of PAmCherry-Swi6 in *mst2*Δ. SMAUG identifies four distinct clusters within the data set. **B**: Cartoon depiction of the full SMAUG results for this data set, including transition probabilities.**C**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis of PAmCherry-Swi6 in *clr3*Δ. SMAUG identifies four distinct clusters within the data set. **D**: Cartoon depiction of the full SMAUG results for this data set. Bubble colors correspond to the term colors in **A** or **C** and bubble sizes represent the weight fractions. Arrows between bubbles indicate the mean of the transition matrix elements for transitions between those terms. Dashed lines indicate transition probabilities that are negligible.

model with mean diffusion coefficients $D = \{0.010, 0.039, 0.147, 0.639\}$ $\mu m^2/s$ and with weight fractions of $\pi = \{0.18, 0.29, 0.25, 0.27\}$ (Fig. 3.5C). Overall, the diffusion coefficient for each term is slightly elevated and indicates a more mobile system as a whole. The weight fraction of both State 1 and State 2 has decreased (Fig. 3.5C,blue and red) relative

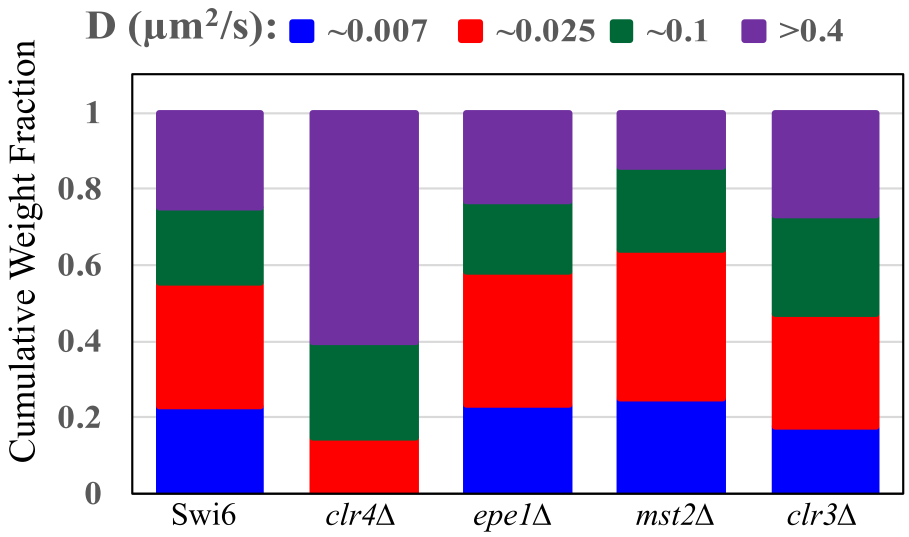


Figure 3.6: Dynamics of PAmCherry-Swi6 with chromatin remodeler mutants. Bar graph showing the amount of weight fraction for the system that each colored state identified for Swi6 in wildtype cells occupies. Purple: most mobile state ($D > 0.4\mu m^2/s$); blue: least mobile state ($D < 0.01\mu m^2/s$)

to the Swi6 State1 and 2 populations in wildtype cells while States 3 and 4 are increased (Fig. 3.5C, green and purple), suggesting that the acetyl marks hamper the protein-target binding, both at histones with the H3K9me mark and at those without. Moreover, all of the transitions between states are increased, indicating a lower dwell time of Swi6 in each state. Perhaps, again, the increased number of acetlyated histones leads to faster Swi6 unbinding (Fig. 3.5D).

Taken together, these results support the idea that the DNA topology itself is, at least in part, a driver of the dynamics seen in Swi6 and therefore influences its role. Anti-silencing mutations such as *clr4*Δ and *clr3*Δ both increase the overall mobility of Swi6, either by removing the target and thereby causing Swi6 to be more diffusive (*clr4*Δ) or by decreasing the populations of Swi6 in the slowest diffusing states and instead increasing the transitions into the more mobile states (*clr3*Δ). Conversely, the results for silencing mutations, such as *epe1*Δ and *mst2*Δ, are more difficult to generalize. On the one hand, *mst2*Δ caused a slight shift towards a less mobile system with more weight fractions in the two least mobile terms and fewer transitions overall. However, another silencing mutation, *epe1*Δ, resulted almost no shift at all in the dynamics of Swi6, though we have hypothesized that this result could be in part due to the extremely low copy number of Epe1 in even wildtype cells, leading to a lesser effect on the chromatin state of the DNA in its absence. Figure 3.6 has an overview of the effect of these mutations on the dynamics of PAmCherry-Swi6.

3.6 Sequence Mutations in Swi6 Reveal Functional Roles

To investigate the role of the various domains of Swi6 in the dynamics of the protein we created several in which Swi6 contains mutations strains that have been shown to affect known biochemical activities of the protein. Table 3.2 has a list of strains used in this section. Because Swi6 plays an important role in the cell we worried that creating mutant copies of this protein might affect the health of the cell. Additionally, we wanted to separate out the effect that a particular sequence mutation has on the dynamics of Swi6 from any other effects on the dynamics stemming from a lack of a completely functional Swi6 in the cell. Thus, we created strains where a fusion of the FP mNeonGreen and Swi6 was added back into the genome at the *leu1* locus (denoted Swi6+ in Table 3.2). To investigate the effect of that Swi6 overexpression from the double copies has on the system, we imaged cells containing a PAmCherry-Swi6 fusion at the native locus and a separate mNeonGreen-Swi6 fusion at the *leu1* locus, called Swi6/Swi6+. We imaged these cells using phase-contrast and found they contained no change in cellular morphology (Fig. 3.7A). We then imaged using SPT and collected a data set of 12,016 steps from 58 cells and performed SMAUG analysis on the data set. SMAUG returned a most probable four state model with mean diffusion coefficients of $D = \{0.010, 0.047, 0.116, 0.662\}$ $\mu m^2/s$ and with weight fractions of $\pi = \{o.36, 0.21, 0.23, 0.21\}$ (Fig. 3.7B). This analysis shows that all terms identified in the analysis for Swi6 alone in wildtype cells are present but that the weight fraction of the slowest diffusing state is increased whereas the amount in the second state is lowered. This result could possibly arise from an over-expression of Swi6

Strain Name	Mutation	Description
Swi6/Swi6+	PAM-Swi6 + NG-Swi6	Additional mNeonGreen-Swi6
CD*	PAM-W104A	PAmCherry-Swi6 with W104A point mutation
CD*/Swi6+	PAM-W104A + NG-Swi6	PAmCherry-Swi6 with W104A point mutation + mNeonGreen-Swi6
ARK*	PAM-K93A/R94A	PAmCherry-Swi6 with double point mutant K93A/R94A
ARK*/Swi6+	PAM-K93A/R94A + NG-Swi6	PAmCherry-Swi6 with double point mutant K93A/R94A + mNeongreen-Swi6
CSD*	PAM-L315E	PAmcherry-Swi6 with L315E point mutation
CSD*/Swi6+	PAM-L315E + NG-Swi6	PAmCherry-Swi6 with L315E point mutation + mNeonGreen-Swi6
F324A	PAM-F324A	PAmCherry-Swi6 with F324A point mutation
F324A/Swi6+	PAM-F324A + NG-Swi6	PAmCherry-Swi6 with F324A point mutation + mNeonGreen-Swi6
KR25A	PAM-KR25A	PAmCherry-Swi6 with 25 K & R residues from the Hinge region mutated to A [72]

Table 3.2: Strain List for the Sequence Mutation experiments. PAM is the photoactivatable red FP PAmCherry and NG is the green FP mNeonGreen. All Swi6 mutations are inserted at the endogenous *ura4* locus. FP and protein name in the description represents gene orientation with FP before the protein indicating an N-terminal fusion and after the protein being C-terminal. All *mneongreen-swi6* genes were inserted at the *leu1* locus.

being a pro-silencing-like mutation, leading to more regions of heterochromatin which, as we found above, tends to create a system with more occupancy in the lowest diffusive states. Such a hypothesis is further strengthened by the transition elements: transitions into State 1 are increased, including relatively high transition probabilities from State 4 (Fig. 3.7C). In any case, the addition of an extra gene copy of a fusion *swi6* does not seem to perturb the dynamics of the system too greatly, by, for example, creating terms unseen previously or by shuttling all extra copies into State 4.

3.6.1 Chromodomain Mutations Alter Swi6 Dynamics to be More Diffusive

Swi6 is made up of the chromodomain (CD) and the chromoshadow domain (CSD). The CD is responsible for recognizing the H3K9me mark and binding to it (Fig. 3.1). To

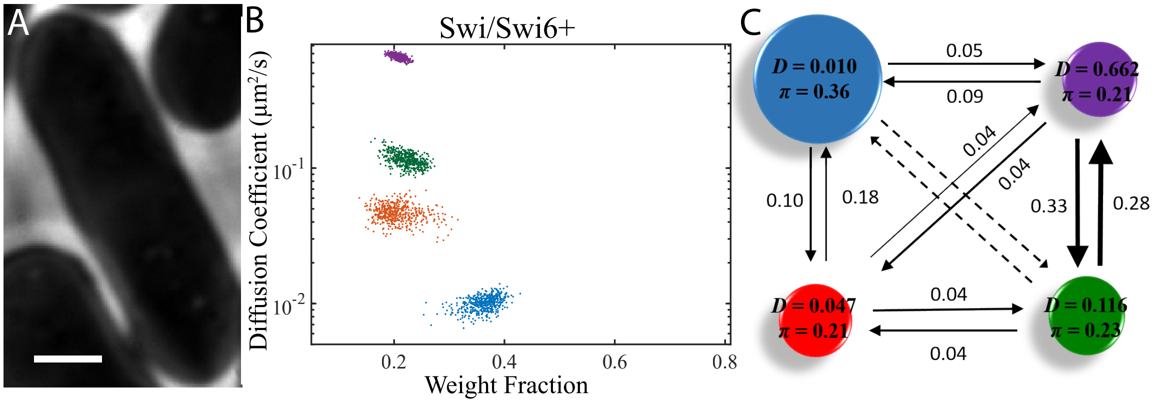


Figure 3.7: Dynamics of the control Swi6/Swi6+ strain. A: Phase-contrast image of a cell containing the PAmCherry-Swi6/mNeonGreen-Swi6+ strain and showing normal morphology. Scale bar = 2 μm. **B**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for Swi6/Swi6+ strain. SMAUG identifies four distinct clusters within the data set. **C**: Cartoon depiction of the full SMAUG results for this data set, including the transition probabilities.

investigate the effect of the chromodomain binding to the H3K9me mark and its role in the dynamics of Swi6, we created a mutant of the Swi6 protein containing a point mutation (W104A). This mutant has been shown *in vitro* to have decreased binding affinity for histones containing the H3K9me [76, 77]. We have labeled this strain CD* as the mutation is in the chromodomain and effects binding of Swi6, the essential function of the CD. We imaged these cells and collected a data set of 9,274 steps and analyzed with eh SMAUG algorithm. SMAUG returns a most probable model with three states, with mean diffusion coefficients $D = \{0.019, 0.104, 0.681\}$ $\mu m^2/s$ and with weight fractions of $\pi = \{0.21, 0.37, 0.42\}$ (Fig. 3.8A). State 1 from the dynamics of Swi6 inside the wildtype cells is not present in the model, similar to the dynamics of PMmCherry-Swi6 in *clr4*Δ, though with a much higher transition probability between the most diffusive states (Fig. 3.8B). This analysis suggests that the W104 residue is required for methyl mark recognition and subsequent binding to the histone, or perhaps that the W to A mutation disrupts this recognition.

We also constructed the CD*/Swi6+ strain which contains the CD point mutant Swi6 (W104A) fused to the PAmCherry tag but also contains a a copy of the unmutated mNeon-

Green-Swi6 fusion. We imaged these cells collecting 12,481 steps. SMAUG analysis of the data supports a four state model as the most probable one (Fig. 3.8C), with mean diffusion coefficients $D = \{0.011, 0.049, 0.192, 0.717\}$ $\mu m^2/s$ and with weight fractions of $\pi = \{0.10, 0.19, 0.29, 0.42\}$. The four states all contain higher diffusion coefficient estimates than the four states identified for the Swi6 in wildtype cells. We hypothesize that the reappearance of the 4th state with a very low diffusion coefficient is due to the heterodimerization of a Swi6 containing the W104A mutant with an mutationless mNeonGreen-Swi6 molecule. The transition matrix shows a much more dynamic system overall, perhaps indicative of the role that binding plays in pushing the system towards the slower mobility states when present (Fig. 3.8D). This analysis suggests that binding to a H3K9me histone may only require a single full length copy of the *swi6* gene. Additionally, even with the extra fusion copy, the amount of weight occupied in State 2 is decreased in both strains, perhaps indicating that the W104A mutation has affected the non-specific binding of histones and/or the search mechanism of Swi6 as well.

Another important aspect of the chromodomain is its role in forming extended, linked regions of heterochromatin. Dimers of Swi6 bound to histones containing the H3K9me mark can interact with other, nearby Swi6 dimers bound to other histones. This interaction occurs through the "ARK loop" (Fig. 3.1A red loops), a 3 residue loop in the CD that, upon binding to the H3K9me histone, opens up into a configuration that is accessible to other Swi6 molecules [71]. We constructed two strains wherein the ARK loop is mutated to AAA, one where the mutant copy is the only copy of Swi6 in the cell and another that contains the AAA mutant plus a copy of mNeonGreen-Swi6. The ARK* strain has a data set containing 13,577 steps. SMAUG analysis returns a model that that has some striking characteristics. SMAUG returns a most probable four state system with mean diffusion coefficients $D = \{0.009, 0.029, 0.135, 0.652\}$ $\mu m^2/s$ and with weight fractions of $\pi = \{0.29, 0.36, 0.15, 0.20\}$ (Fig. 3.9A). While this result seems to resemble the dynamics of Swi6 in wildtype cells at first glance, the striking aspect is the transition matrix en-

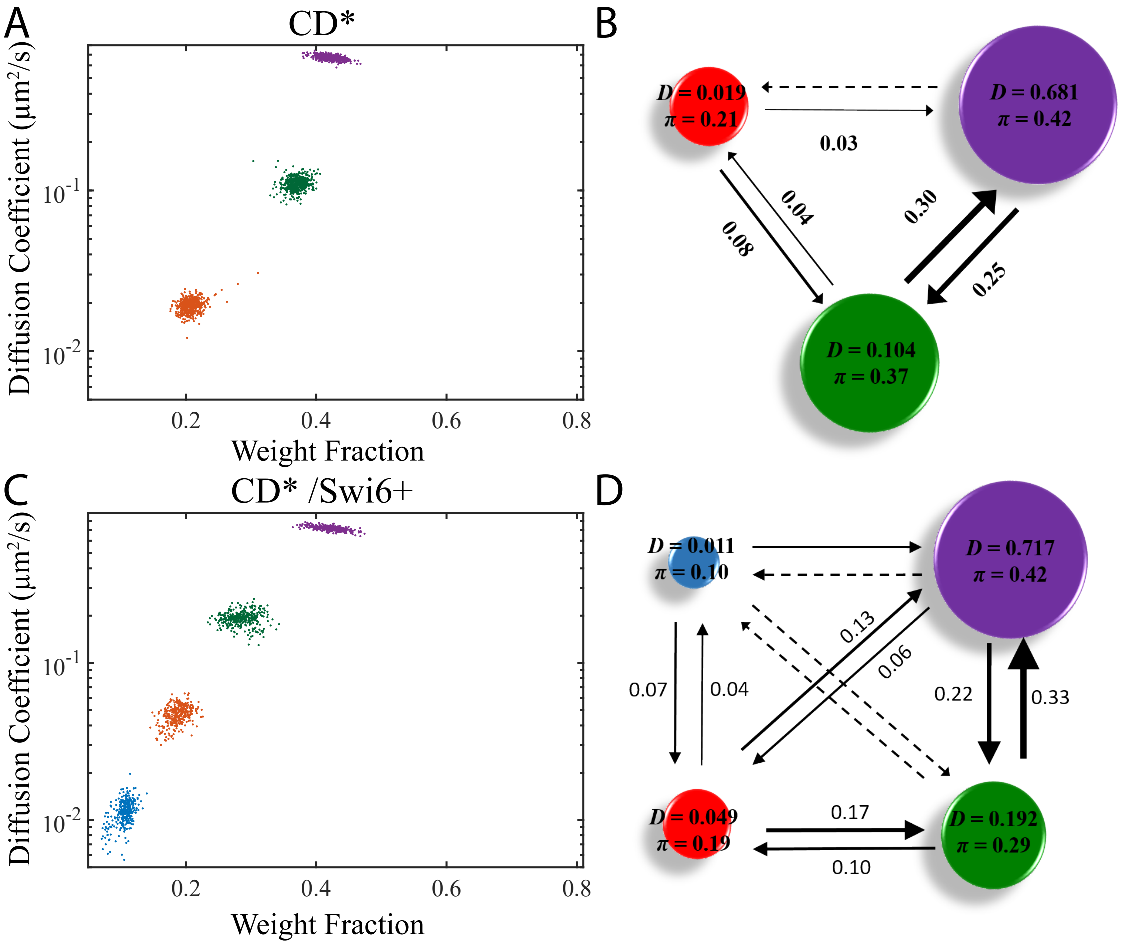


Figure 3.8: Dynamics of PAmCherry-Swi6 in CD* and CD*/Swi6+. A: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for CD*. SMAUG identifies three distinct clusters within the data set. **B**: Cartoon depiction of the full SMAUG results for this data set.**C**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for CD*/Swi6+. SMAUG identifies four distinct clusters within the data set. **D**: Cartoon depiction of the full SMAUG results for this data set.

tries. There are negligible transition probabilities between Sates 2 and 3 and instead the transition path has bypassed State 3 completely in order to move directly to the other intermediate. The transitions between States 4 and 2 have increased in order to compensate for this isolation of State 3 (Fig. 3.9B). This analysis suggests that the interactions of the ARK loop, or at least its accessibility/conformational changes upon binding, are an important aspect of moving between the more diffusive State 3 and the relatively slowly diffusing State 2.

The data set for the ARK*/Swi6+ strain contained 14,007 steps and SMAUG analysis returned as the most probable a four state system with mean diffusion coefficients $D = \{0.007, 0.019, 0.091, 0.707\}$ $\mu m^2/s$ and with weight fractions of $\pi = \{0.09, 0.32, 0.32, 0.27\}$ (Fig. 3.9C). The addition of the extra fusion copy has restored the transitions into and out of State 3 and consequently reduced the minor path transitions, indicating that at least 1 functional copy of the ARK is sufficient for the molecule to function. Additionally, the amount of the system occupying the slowest state as decreased to roughly 10%, similar to the CD*/Swi6+ strain, perhaps as a result of the molecules containing the ARK* mutation being out-competed or less stable while bound and so fewer molecules occupy that state was a fraction of the whole system (Fig. 3.9D).

Taken as a whole, the mutations inside the choromodomain provide a details about the biochemical roles of several of the states identified for the dynamics of Swi6. Mutation of the important recognition residue W104 causes the loss of the bound state but the introduction of extra mutationless copies restores some bound state, indicating that the presence of 1 functional copy of the CD is sufficient for binding to the histone mark. Further, mutation of the residues inside the ARK loop provides the surprising evidence that the presence of at least one functional ARK loop inside the Swi6 dimer is required for transitions through the intermediate states, suggesting that State 3 might be involved with protein-protein interactions. Figure 3.10 has a summary of the dynamics discussed in this section.

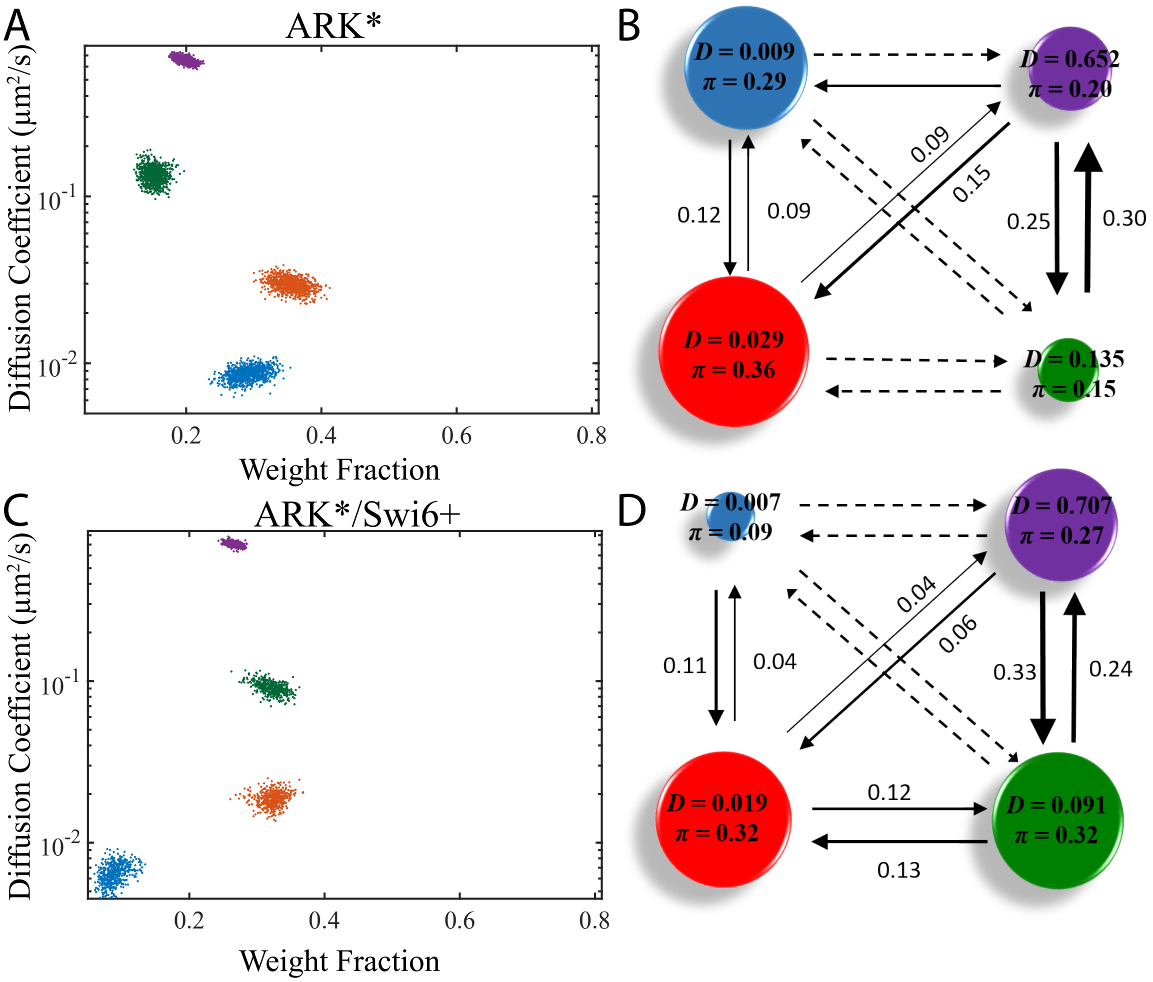


Figure 3.9: Dynamics of PAmCherry-Swi6 in ARK loop mutants. A: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for ARK*. SMAUG identifies four distinct clusters within the data set. **B**: Cartoon depiction of the full SMAUG results for this data set. Interestingly, the transitions between States 2 and 3 have disappeared. **C**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for ARK*/Swi6+. SMAUG identifies four distinct clusters within the data set. **D**: Cartoon depiction of the full SMAUG results for this data set, showing recovered transitions into and out of State 3.

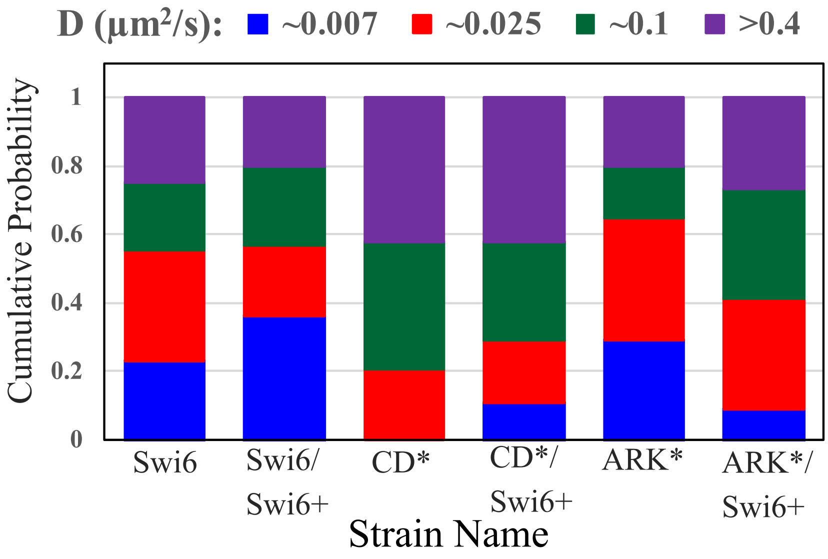


Figure 3.10: Dynamics of PAmCherry-Swi6 with CD mutations. Bar graph showing the amount of weight fraction for the system that each colored state identified for Swi6 in wildtype cells occupies. Purple: most mobile state ($D > 0.4 \mu m^2/s$); blue: least mobile state ($D < 0.01 \mu m^2/s$).

3.6.2 Chromoshadow Domain Mutations Reveal Importance of Dimerization and Identity of Third State

The other main domain of the Swi6 molecule is the chromoshadow domain (CSD). The CSD is involved in protein-protein interactions. It contains the dimerization region as well as mediates other protein interactions important for building the heterochromatin remodeling scaffolds. We constructed several strains that are aimed at understanding the effect that protein interactions has on the dynamics of the molecule. We constructed a strain, CSD*, which contains the point mutation L315E, which has been shown to disrupt the dimerization of two Swi6 molecules [78,79]. We then imaged the CSD* strain and collected a data set of only 3,566 steps from 61 cells. This data set was very hard to image as their was very little amount of activatable PAmCherry-Swi6 in the nucleus. I would guess somewhere in the range of 1-10% the normal amount we see in our experiments. This persists even under very high 405 nm photoactivation pulses, indicating that the mutation caused some issue with the folding of the protein or perhaps is leading to a more rapid degradation inside the cell. We nonetheless still analyzed this data set and SMAUG returned a three state model with mean diffusion coefficients $D = \{0.019, 0.086, 0.639\}$ $\mu m^2/s$ and with weight fractions of $\pi = \{0.19, 0.39, 0.42\}$ (Fig. 3.11A), though the variance of the estimates around these means is much greater than in other strains due to the lack of data. Overall, the mean of the parameter estimates and transition elements in the CSD* strain are very similar to those discovered from the CD* strain: the slowest state is gone and the diffusion estimates are very similar (Fig. 3.11B). Qualitatively, while the diffusion of the system seems similar, the lack of normal protein in the cells indicates that dimerization is important for Swi6 for more than simply binding the H3K9me mark.

We also constructed a strain that contains a Swi6 fused to mNeonGreen along with the CSD* mutant, this strian is labeled CSD*/Swi6+. We then imaged this strain and found that much like the CSD* strain, the cells had very little photo activatable protein, making data collection difficult. We nonetheless collected 2,576 steps from 33 cells and analyzed the data set using the SMAUG algorithm. SMAUG returned a three state model, similar to the CSD* strain, with mean diffusion coefficients of $D = \{0.014, 0.088, 0.762\}$ $\mu m^2/s$ and mean weight fractions of $\pi = \{0.21, 0.32, 0.48\}$ (Fig. 3.11C), though again with a larger variance around these means then in other strains. The transition elements are also very similar to the CD* and CSD* cases (Fig. 3.11D). This analysis makes sense as the CSD* lacks dimerization ability and thus should be insensitive to the presence or absence of the extra fusion mNeonGreen-Swi6 copies and reinforces the conclusion that dimerization is an important factor in the ability of Swi6 to not just bind but to stay present in the cell.

Swi6 acts as a scaffold for the recruitment of other proteins whose biological functions can vary. The CSD of Swi6 is the site at which many of these protein-protein interactions occur, especially important is the residue phenylalanine 324, which has been shown to mediate Swi6 interactions with other DNA associated proteins, such as Cdc18, a replication protein [80]. To investigate the role of protein-protein interactions on the dynamics of Swi6, we constructed a strain, F324A, with this residue mutated into a non-

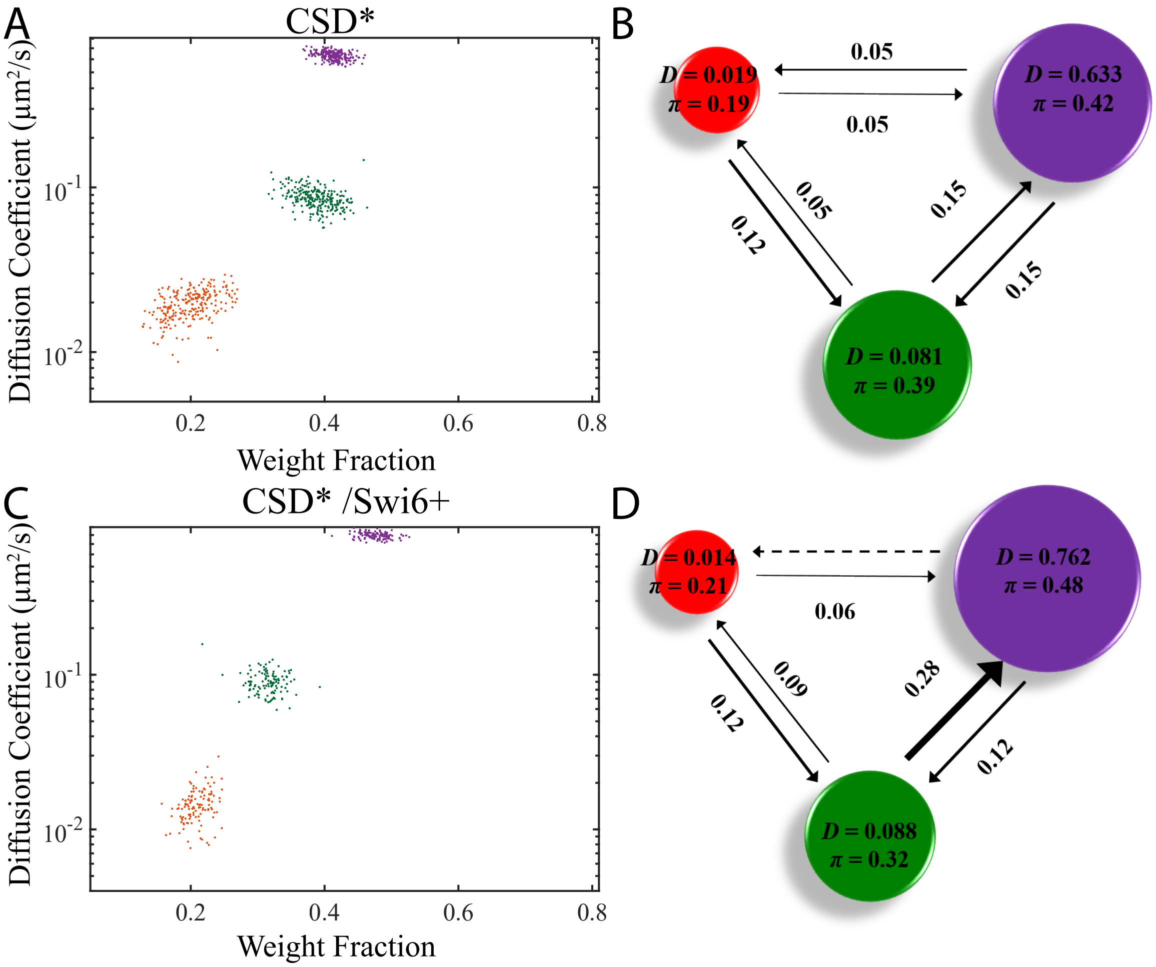


Figure 3.11: Dynamics of PAmCherry-Swi6 in CSD* mutants. **A**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for CSD*. SMAUG identifies three distinct clusters within the data set. **B**: Cartoon depiction of the full SMAUG results for this data set.**C**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for CSD*/Swi6+. SMAUG identifies three distinct clusters within the data set. **D**: Cartoon depiction of the full SMAUG results for this data set.

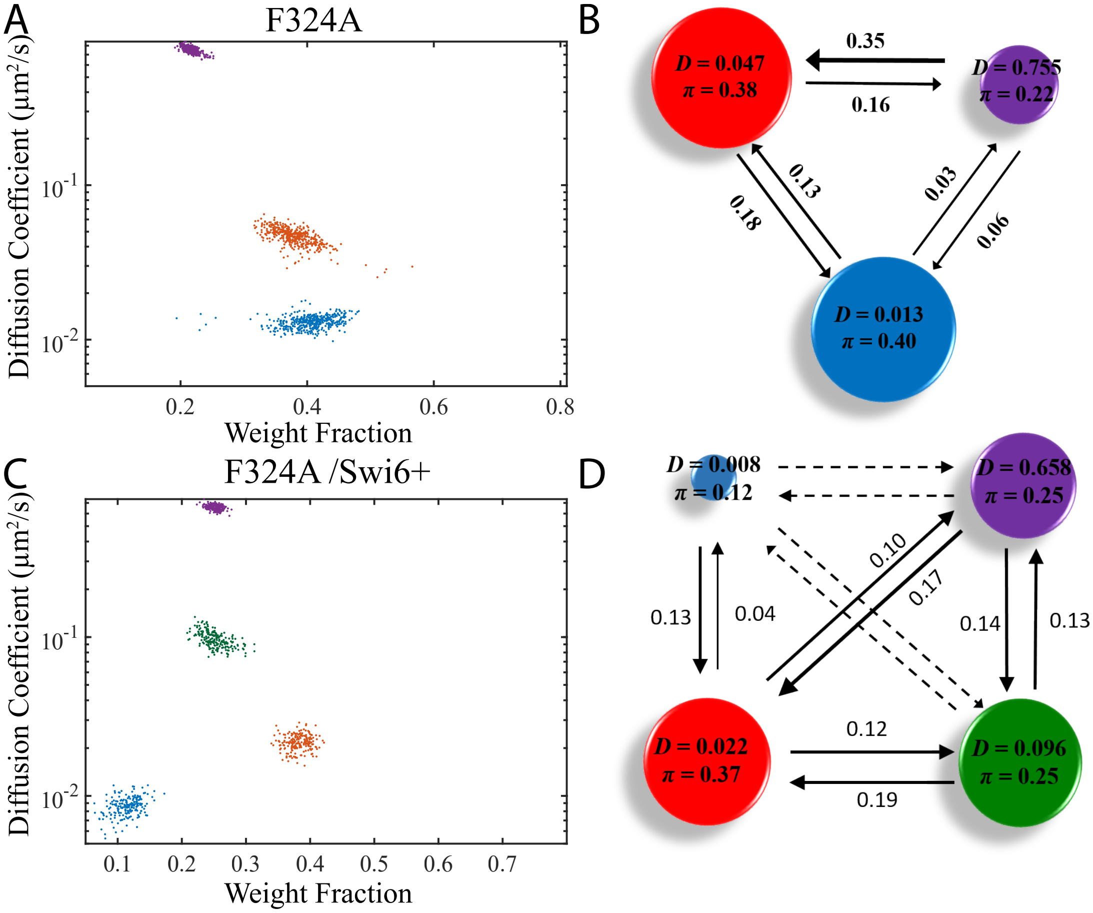


Figure 3.12: Dynamics of Swi6 F324A mutants. A: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for F324A strain. SMAUG identifies three distinct clusters within the data set, interestingly State 3 seems to be the one that has been removed. **B**: Cartoon depiction of the full SMAUG results for this data set.**C**: Diffusion coefficient and weight fraction for F324A/Swi6+. SMAUG identifies four distinct clusters within the data set. **D**: Cartoon depiction of the full SMAUG results for this data set.

functional form. We imaged this strain collecting 10,083 steps from 36 cells. SMAUG analysis returns a most probable three state model with mean diffusion coefficients of $D = \{0.013, 0.047, 0.755\}$ $\mu m^2/s$ and mean weight fractions of $\pi = \{0.40, 0.38, 0.22\}$ (Fig. 3.12A). These identified states are interesting in that they seem to lack any state that overlaps with State 3 from the Swi6 dynamics in wildtype cells (Fig. 3.3B, green), i.e. a state with a diffusion coefficient around 0.1 - 0.2 $\mu m^2/s$, a state present in the analysis of every other strain. Additionally, the two states with the slowest diffusion coefficient contain around 80% of the weight in the system and have diffusion coefficients slightly higher than similar states identified in the analysis of our other cell strains. Transitions between the slowest states is rapid as is the transition from State 4 to State 2 (Fig. 3.12B). We also constructed a Swi6+ version of this strain, called F324a/Swi6+ that contains the extra fusion copy of mNeonGreen-Swi6. We imaged these cells, collecting a data set of 10,657 steps and analyzed those steps using the SMAUG algorithm. SMAUG returns a most probable four state model where the mean diffusion coefficients of $D = \{0.008, 0.022, 0.097, 0.658\}$ $\mu m^2/s$ and mean weight fractions of $\pi = \{0.12, 0.37, 0.26, 0.25\}$ (Fig. 3.12C). The addition of the extra fusion mNeonGreen-Swi6 copy has restored State 3 to the system as well as decreased the diffusion coefficient values for the slowest two states. This analysis resembles very closely the output from the SMAUG method for the ARK*/Swi6+ strain (Fig. 3.9D), except that the transitions between the intermediate states shows no preference for going through State 3 over going straight to State 2 as it does in the ARK*/Swi6+ strain and similar strains (Fig. 3.12D).

Another potential role for Swi6 is that of monitor of silenced genes. If a silenced gene is accidentally transcribed, it has been proposed that Swi6 will bind the mRNA through interaction with the highly-positively charged hinge region, release the H3K9me mark, and transport it for degradation [72] (Fig. 3.1B). To investigate how the hinge region might be influencing Swi6 dynamics we constructed a strain, KR25A, in which 25 positively-charged residues of PAmCherry-Swi6 from the hinge region are mutated to alanine. We imaged this strain and collected 10,038 steps. SMAUG analysis returns a most probable three state model with mean diffusion coefficients of $D = \{0.009, 0.038, 0.416\}$ $\mu m^2/s$ and mean weight fractions of $\pi = \{0.51, 0.37, 0.12\}$ (Fig. 3.13A). Like the F324A strain, this analysis indicates a total lack of State 3, with a majority of the population existing in the slowest, histone-bound form of the molecule. Additionally, the transition matrix indicate very low transition probabilities out of the slowest states, indicating tight binding between the histones and Swi6 (Fig. 3.13B). The absence of State 3 indicates that the hinge region is an important part of the protein-protein interactions, perhaps due to the highly-charged nature of the mutationless hinge or through interaction with mRNA degradation complexes. The large weight fraction of the bound state and the low transition probability out of that state may also indicate that Swi6 binding to mRNA (and thereby releasing the H3K9me mark) constitutes a large amount of the unbinding events for Swi6.

Putting it all together, observing the dynamics of PAmCherry-Swi6 containing mutations inside the chromoshadow domain provides a wealth of details about the biochemical roles of several of the states identified for the dynamics of PAmCherry-Swi6. Mutation of the residue involved in Swi6 dimerization both removes it ability to bind the H3K9me mark and also seems to cause a more rapid degradation of the protein. The addition of a mutationless copy does not alleviate these issues as the mutationless copy cannot bind to the mutated one. Additionally, by mutating the important protein-interaction residue F324 we were able to identify that State 3 is involved in protein-protein interactions that are separate from Swi6's binding to the H3K9me mark and that the addition of mutationless Swi6 was again able to rescue this state, indicating that a single functional interaction face is needed. Finally, by mutating the hinge region into a form that will no longer bind to RNA we were able to discover that the State 3 protein interactions require a functional hinge region, either because the hinge is important in its own right or because this state requires RNA to exist. Figure 3.13C has a summary of the dynamics discussed in this section.

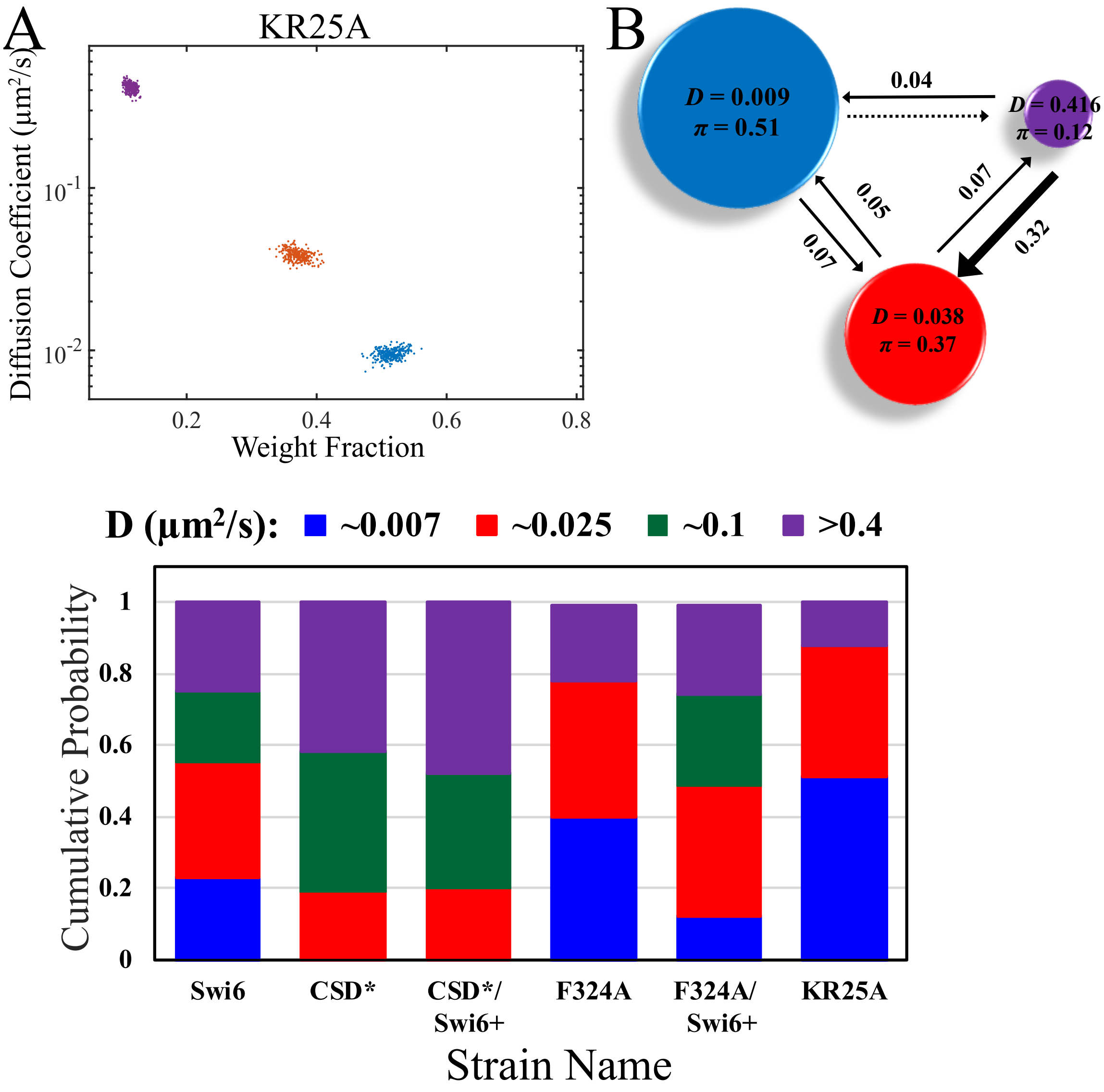


Figure 3.13: Dynamics of PAmCherry-Swi6 with CSD mutations. A: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for KR25A strain. SMAUG identifies three distinct clusters within the data set, with State 3 not present. **B**: Cartoon depiction of the full SMAUG results for this data set. **C**: Bar graph showing the amount of weight fraction for the system that each colored state identified for Swi6 in wildtype cells occupies. Purple: most mobile state ($D > 0.4\mu m^2/s$); blue: least mobile state ($D < 0.01\mu m^2/s$).

Strain Name	Mutation	Description
Clr3PAM	PAM-Clr3 + *swi6*Δ	PAmCherry-Clr3 + *swi6*Δ
Clr3PAM/Swi6+	PAM-Clr3 +*swi6*Δ +NG-Swi6	PAmCherry-Clr3 + *swi6*Δ + mNeonGreen-Swi6
Epe1PAM	Epe1-PAM + *swi6*Δ	Epe1-PAmCherry + *swi6*Δ
Epe1PAM/Swi6+	Epe1-PAM +*swi6*Δ +NG-Swi6	Epe1-PAmCherry + *swi6*Δ + mNeonGreen-Swi6

Table 3.3: Strain list for dynamics of accessory proteins. PAM is the photoactivatable red FP PAmCherry and NG is the green FP mNeonGreen. All mutations are made at the endogenous *ura4* locus except for the mNeonGreen-Swi6 insertions which are inserted at the *leu1* locus. FP and protein name in the description represents gene orientation with FP before the protein indicating an N-terminal fusion and after the protein being C-terminal.

3.7 Dynamics of Accessory Chromatin Proteins

Having a clearer view of the dynamics of Swi6 and the biochemical roles played by each of the identified states, we also decided to investigate the dynamics of the other proteins involved in chromatin remodeling. We constructed strains containing fusions of Clr3 and Epe1 with PAmCherry for single-particle tracking experiments, in both *swi6*Δ and *swi6+* backgrounds for both (Table 3.3). We then imaged these strains in SPT experiments to determine their dynamics. The two strains containing Epe1-PAmCherry showed essentially zero photo activatatable signal inside the cellular nucleus, even under such high activation pulses that they caused the cells to form vesicles, we assume from acute stress. Because of this lack of signal, I abandoned this avenue of investigation but I include it here in this Thesis for the sake of completeness and also so that others may know that the live-cell imaging of Epe1 may be too difficult without inducing high levels of overexpression in the cells.

We then imaged the two strains containing Clr3-PAmCherry fusions. Clr3 is the histone deacetlyase protein that removes the acetly marks from histone, restricting euchromatin regions. These strain are harder to image as there exists less Clr3 in the cells and preliminary data is insufficient for quantitative analysis, but in the future, this strain will allow us to observe the dynamics of Clr3 in similar ways to those discussed above for Swi6 and investigate the hypothesis that Clr3 has only the single function of histone deacetylase and so we suspect relatively simple dynamics representing bound and unbound Clr3. Deviations from this expected result would suggest other roles for Clr3 and can inform further biological experiments.

3.8 Conclusions

The processes that alter the epigenetic state of an organism are complex and contain competing pathways: the establishment and maintenance of epigenetically silenced regions involves a large array of proteins whose actions compete with opposing pathways and whose timing must be tightly controlled. The protein Swi6 is the key protein responsible for the the creation of heterochromatin regions, thus its dynamics are of particular interest to researchers. The studies presented in this chapter have used single-particle tracking to observe this protein, and others, in real time inside living cells. We then utilized the SMAUG algorithm to uncover the hidden states within the data to draw conclusions about the biochemical roles of the identified states. We identified four distinct states in the dynamics of Swi6 inside wildtype cells. Using our mutation studies, we were able to determine that the dynamics displayed by Swi6 are related to both the global compactness of the DNA as well as to specific functional interactions of the molecule. Through these studies, we identified the identities of the diffusive states and the contributions of DNA topology, and using this information we can propose a potential model pathway for the function of Swi6. To begin, dimerization of the protein is a crucial first step for the presence and function of the protein and without dimerization, protein stability and the function of the rest of the proposed pathway is altered. Through our mutation studies, we were able to discover four distinct diffusive states for the Swi6 molecule, corresponding to four identified biochemical roles. In order of decreasing diffusion coefficient they are: freely diffusive, interacting with or binding to DNA, interacting with histones lacking the H3K9me mark and, finally, bound to histones containing the H3K9me mark. Using the transition matrix probabilities we can also conclude that freely diffusive Swi6 proteins must either first bind to DNA or to a histone before moving into the final completely bound state. Thus our model is that freely diffusive Swi6 proteins first bind directly to DNA in the nucleus, which brings the protein close to the histones. Transitions out of the DNA binding state and to the histone binding state was shown to require the conserved ARK loop in the chromodomain, suggesting that interactions between different Swi6 molecules is essential for this transition. From this state, the Swi6 protein can continue to sample histones in the dense heterochromatin region until it encounters one containing the H3K9me mark and binding to it, recruiting further remodeling proteins through the CSD domain. The topology of the DNA can affect the steps in this pathway by encouraging certain transitions over others, as we found that with anti-silencing mutations, the more diffusive states of free diffusion and DNA interaction were favored over histone bound states.

Taken together, the work presented in this chapter shows the power of combining real time measurements of single molecule dynamics and an analysis method such as SMAUG for determining the dynamics and biochemical roles of proteins. Both components are needed to discover the true behavior of a complex system such as this one and to use that information to construct a model for the behavior that fits with the results and that can then be further tested and refined. The results stated in this chapter advance our knowledge of this important system by providing a potential mechanistic model for the interactions of Swi6 while also providing a proof-of-concept that complex systems can be investigated by SPT, leading to meaningful and robust conclusions.

CHAPTER IV

Revealing the Dynamics of a Bacterial Virulence Pathway in *V. cholerae*

The work presented in the chapter is a collaboration between the Biteen Lab and the DiRita lab at Michigan State University

In preparation as a manuscript by Karslake, J.D.*, Demey, L.M.*, Donarski, E.D., DiRita, V.J., and Biteen, J.S. [1]

4.1 Introduction

Since records began to be kept in 1817, eight separate epidemics of the disease cholera have been recorded worldwide [81–83]. While the disease is almost entirely absent from the Unites States and similarly industrialized countries, it remains a global health concern. The disease is found mainly in regions with poor sanitation, regions that are experiencing prolonged violence, such as the recent outbreak in Yemen, or are recovering from a disaster, such as the Haitian outbreak in 2010 after the earthquake. The disease is estimated to

[1] **Author contributions** - J.D.K. and L.M.D. contributed equally to this work. All authors designed the research. J.D.K and E.D.D. preformed the live cell experiments and analyzed the results. L.M.D. constructed the bacterial strains and performed the biochemical assays.

affect as many as 5 million people annually [54]. Cholera is a disease caused by a bacterial infection by the Gram-negative bacterium, *Vibrio cholerae*, usually contracted through the ingestion of contaminated food and water. *V. cholerae* strains are classified by both the agglutination state of the O group 1-specific antiserum directed against the bacterial cell wall, as well by the bacterium's enterotoxigenicity [84]. Within the main toxigenic strain, O1, there are two biotypes, classical and El Tor, each with distinct serotypes. The El Tor biotype is generally associated with milder clinical symptoms than the classical biotype and is responsible for more cases of the disease globally since the mid 1980s [36]. However, the work presented here in this chapter focuses on the O1 classical biotype, serotype Ogawa 395 (shortened to O395), as it has historically caused seven of the eight disease epidemics and as this strain has been extensively characterized genetically. Thus, the study of this biotype could lead to better understanding of the pathogenicity of this organism and help to reduce the severity of possible future outbreaks.

In this chapter, I discuss my investigations of the dynamics involved in a bacterial virulence pathway using super-resolution microscopy and single-particle tracking (SPT) and the model system *V. cholerae*. To begin, I will discuss the relevant biological background and reasons for using single-molecule methods to probe this system. In the following sections, I will present the results of my measurements of the dynamics of this system and use specific genetic mutations to perturb the system in illuminating ways. Finally, I will conclude with some thoughts on the impact and reach of this study.

4.2 Biological Background

For *V. cholerae* to infect humans, it must be ingested and, passing through the gut to the intestines, colonize the surface of the intestinal epithelial cells. Adherence to the host's epithelial cells requires a specialized protein complex called the toxin co-regulated pilus (TCP), a type IV pilus. Upon adherence, host factors and environmental changes induce the production of the virulence pathway and the final product, the cholera toxin (CTX).

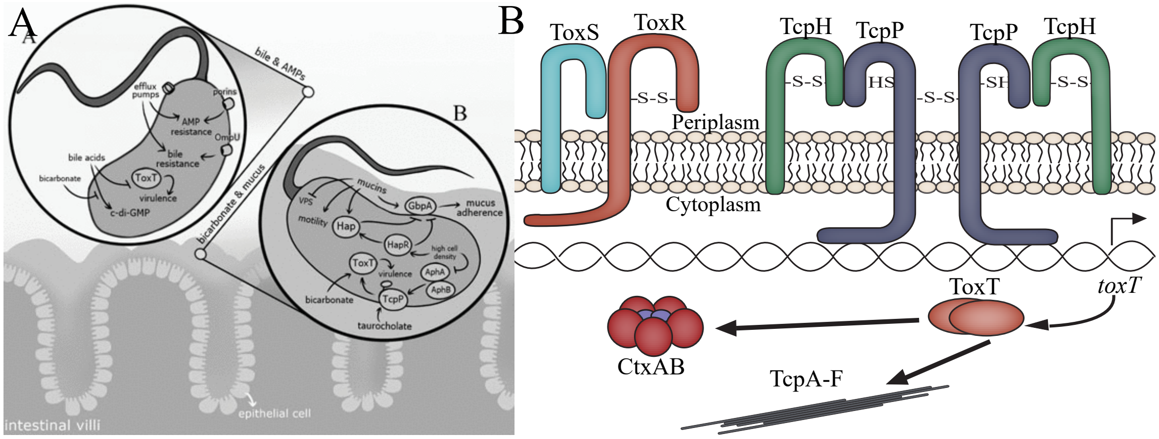


Figure 4.1: The *Vibrio cholerae* Virulence Pathway. A: Schematic of the virulence pathway for *V. cholerae* under different conditions. Adhered to the intestinal wall the bacterium is exposed to mucins and other signals that upregulates the virulence pathway. Figure from [86]. **B**: Schematic of the *V. cholerae* virulence pathway. The membrane-bound protein TcpP binds the DNA directly along with other supporting proteins, leading to the hypothesis that the dynamics of TcpP reflect multiple mobility states. Reproduced from [75].

CTX is a heterohexamer, with one CTxα subunit surrounded by five CTxβ subunits [85] (Fig. 4.1B, bottom left cartoon). This protein is secreted from the bacterium under virulence conditions and it adheres to the host epithelial cells through interactions between the CTxβ subunit and the membrane ganglioside GM_1 and is endocytosed. When inside the host cells the CTxβ subunits are removed leaving the catalytic CTxα subunit exposed. The CTxα subunit then ADP-ribosylates host G-proteins to constitutively activate cyclic-AMP production. Such high levels of cAMP lead to a large increase in the secretion of chloride ions and then water out of the host cells and into the intestinal lumen, causing severe dehydration and death within as little as a few hours if untreated [42] (Fig. 4.1A).

CTX production is regulated by the binding of a regulatory transcription factor ToxT. ToxT is a member of the large AraC/XylS family of proteins and has two domains [87]. The C-terminal domain is known to mediate binding of the ToxT protein to DNA regions called toxboxes, 13-bp degenerate sequence of repeats just upstream of the ToxT activated genes, through a helix-turn-helix motif. The N-terminal domain shares no homology with any other protein domain, as determined by BLAST analysis, and its functional role is unclear, though some studies have suggested that it may be involved in dimerization [88] or recognition of small molecule effectors such as bicarbonate or bile [89].

ToxT production is itself under the regulatory control of another set of proteins, the ToxR regulon, named for the first identified positive regulator [90] (Fig. 4.1B). This regulon is made up of four proteins: ToxR, ToxS, TcpH, and the most important factor TcpP. ToxR and TcpP are bitopic membrane proteins that each contain a cytoplasmic DNA-binding domain, a single transmembrane domain, and a periplasmic domain. The function of the periplasmic domains of these proteins is not totally clear, though they are thought to be involved in protein-protein interaction such as dimerization [91, 92]. The activity of ToxR and TcpP has been shown to require the presence of other proteins, ToxS and TcpH respectively. Both ToxS and TcpH have a single transmembrane helix and a periplasmic domain. The role of these accessory proteins is still unclear, though they are thought to influence stability or promote dimerization, as in cells lacking in TcpH, TcpP is rapidly degraded, though ToxR is not when ToxS is not present [93–95]. ToxR and TcpP bind to the promoter region of the *toxT* gene and initiate gene production under conditions favorable to virulence. The exact mechanism of how this initiation occurs and which signals are important are not entirely clear, but *in vitro* it has been shown that virulence factors are responsive to changes in pH and temperature [94, 96]. As mentioned above, TcpP is the most important protein in the regulon. ToxS and TcpH cannot bind DNA, ToxR by itself cannot induce the production of ToxT, while over-expression of TcpP alone has been shown to activate the production of ToxT [97].

The regulation of this virulence cascade is a complex and carefully controlled process. Using single-particle tracking (SPT), and Single-Molecule Analysis by Unsupervised Gibbs (SMAUG) (Chapter II), I observe the dynamics of this system directly by monitoring individual molecules inside living cells under virulence-inducing conditions and uncover distinct biological states that are hidden or averaged over using other methods. Accurately measuring the dynamics of TcpP under various situations will enable investigations into the role of this unusual membrane-localized mechanism of transcription activation. Previously, our lab used fusions to the photoactivatable fluorescent protein PAmCherry to show that TcpP-PAmCherry diffuses heterogeneously in living cells [26].

4.3 Methods

Strain Construction

Strain construction follows protocol outlined in Reference [98]. In brief, strain construction uses donor strain S-17 and *Vibrio cholerae* strains harboring a pKAS plasmid construct with the mutation desired were selected for on LB plates containing ampicillin and streptomycin (both 100 μg/ml) or TCBS plates containing ampicillin (100 μg/ml). Counter selection for loss of the pKAS construct was done by incubating *V. cholerae* strains harboring the pKAS construct in LB for 2hrs and then 2hrs with 2500 μg/ml streptomycin at 37 °C. 20 μl of this culture was then spread onto LB plates containing 2500 μg/ml of streptomycin and incubated overnight at 37 °C. Streptomycin resistant colonies were screened for the chromosomal mutation of interest via colony PCR and genetic sequencing to validate the exchange.

Microscopy experiments

TcpP-PAmCherry was expressed at the native *tcpP* locus and cells were grown under conditions known to stimulate TcpP-mediated expression of virulence genes [42] (LB rich media at pH 6.5 and 30 °C). Once cells reached mid log-phase, they were diluted into M9 minimal media, and then imaged at room temperature on agarose pads using a 406-nm laser (Coherent Cube 405-100; 102 W/cm^2) for photo-activation and a 561-nm laser (Coherent-Sapphire 561-50; 163 W/cm^2) for imaging. Samples are mounted on an Olympus IX71 inverted epifluorescence microscope with a 100x 1.40 NA oil-immersion objective. The fluorescence emission was filtered with appropriate filters and imaged on a 512 by 512 pixel Photometrics Evolve electron multiplying charge-coupled device (EMCCD) camera and continual images were collected with a 40-ms exposure time per frame. Recorded single-molecule positions were detected and localized as previously described using home-built code [27], and connected into trajectories using the Hungarian algorithm [24].

Data Analysis

Trajectory analysis was performed using home built MATLAB code that uses a Hidden Markov Model embedded inside a Gibbs Sampler to estimate the parameters of interest in a data set, called SMAUG (Chapter II). All trajectories for a given condition are bundled together and analyzed together. Briefly, the SMAUG algorithm takes in a data set of trajectories and estimates the parameters of interest using a Bayesian statistical framework. Data is sorted probabilisticly into terms and then used to refine parameter estimates for that term and then the process is repeated. Over time, this iterative process converges onto the most likely values for the parameter estimates given the input data set. Parameters of interest from the data set are the number of mobility states within the set, the diffusion coefficient, weight fraction, and transition probabilities between the states.

4.4 Dynamics of TcpP in living cells

4.4.1 Strain construction of cells with TcpP-PAmCherry

We began our investigations into the dynamics of TcpP in a strain of *V. cholerae* that contained a genetically encoded fusion of TcpP to the photoactivatable fluorescent protein PAmCherry. Due to overlapping reading frames of *tcpP* and *tcpH*, the entire *tcpH* gene was moved to downstream of the *tcpP-PAmCherry* fusion to avoid an internal PAmCherry in TcpH (strain LD47; Table 4.1). Additionally, a companion strain was created in which the *tcpH* gene was removed entirely (strain LD48; Table 4.1). A list of all the strains used in this section can be found in Table 4.1.

4.4.2 Spurious mutations can decreased protein expression levels

It was later found that all LD47/48 strains carried a spurious mutation in the promoter region of the gene, leading to vastly decreased levels of TcpP and TcpH in the cells. Thus, all strains carrying the LD47/48 as the parent strain were subsequently abandoned as we do not believe the dynamics displayed represent a true picture. I mention these issues here for completeness in reporting but no further work has occurred on these strains.

4.4.3 Biochemical characterization and the dynamics of TcpP

After the discovery of the mutation, new strains were constructed bearing the correct target genetics (LD51 and LD52; Table 4.1). LD51 contains a full length TcpP-PAmCherry fusion with a downstream TcpH and LD52 is LD51 lacking the *tcpH* gene. Biochemical characterization shows that these strains behave as wild-type cells do in that they produce CTxβ at wildtype levels, carry a fusion TcpP-PAmCherry protein that is degraded in cells lacking TcpH, and have wild-type growth and normal cellular morphology under the microscope (Fig. 4.2).

We then imaged these cells in SPT experiments and collected 11,403 steps from 2404 trajectories for LD51. LD52 contained almost no photoactivatable molecules, which aligns with the biochemical data (Fig. 4.2) that TcpP is degraded rapidly in the absence of TcpH. Analysis of the LD51 dataset by SMAUG indicated a most probable interpretation of a $K = 3$-state model with diffusion coefficients of $D_i = \{0.006, 0.044, 0.368\}\mu m^2/s$ and weight fractions of $\pi_i = \{0.18, 0.53, 0.29\}$ (Fig. 4.3). We attribute these identified states to different biological roles of the protein in the cell. Additionally, the transition elements depict a clear path from the state with the highest diffusion coefficient through an intermediate state and from there to the slowest state. This trajectory indicates that the moderate mobility state (red, Figure 4.3C) is a necessary intermediate between fast- and slow-diffusing states (green and blue, respectively, Figure 4.3C). This analysis provides us with a baseline of dynamics which we can then use in conjunction with targeted mutation

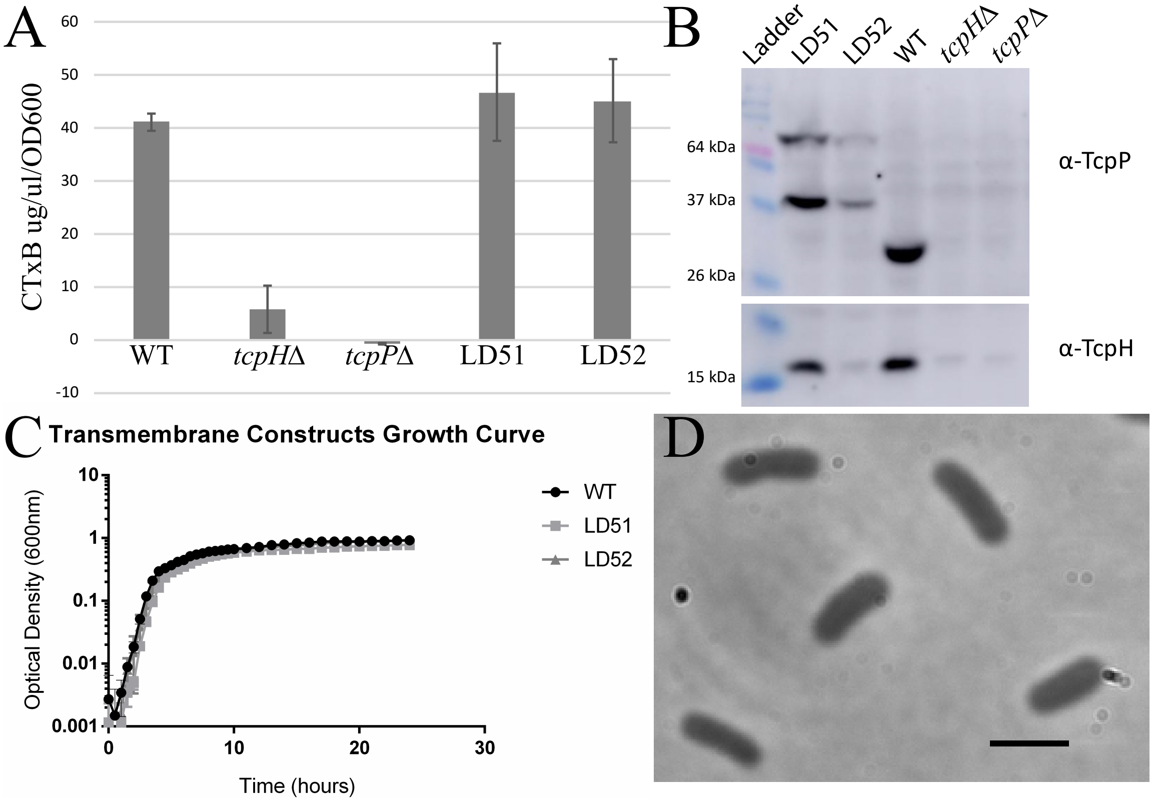


Figure 4.2: Biochemical characterization of cells containing TcpP-PAmCherry. **A**: CTxβ levels in culture supernatants after 24 hrs of incubation in LB, pH 6.5, at 30 °C. **B**: Western blots of whole cell lysates collected after 24 hrs of incubation in LB, pH 6.5, at 30 °C. Western blots were probed with either α-TcpP (Top blot) or α-TcpH (bottom blot). **C**: Growth of WT, LD51 and LD52 in LB, 37 °C. OD600nm values are an average of three biological replicates. **D**: Phase-contrast image of LD51 cells showing normal cellular morphology. Scale bar: 2 μm.

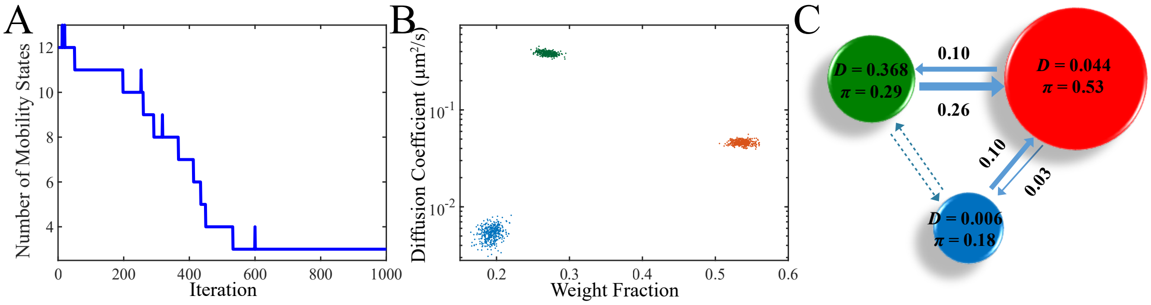
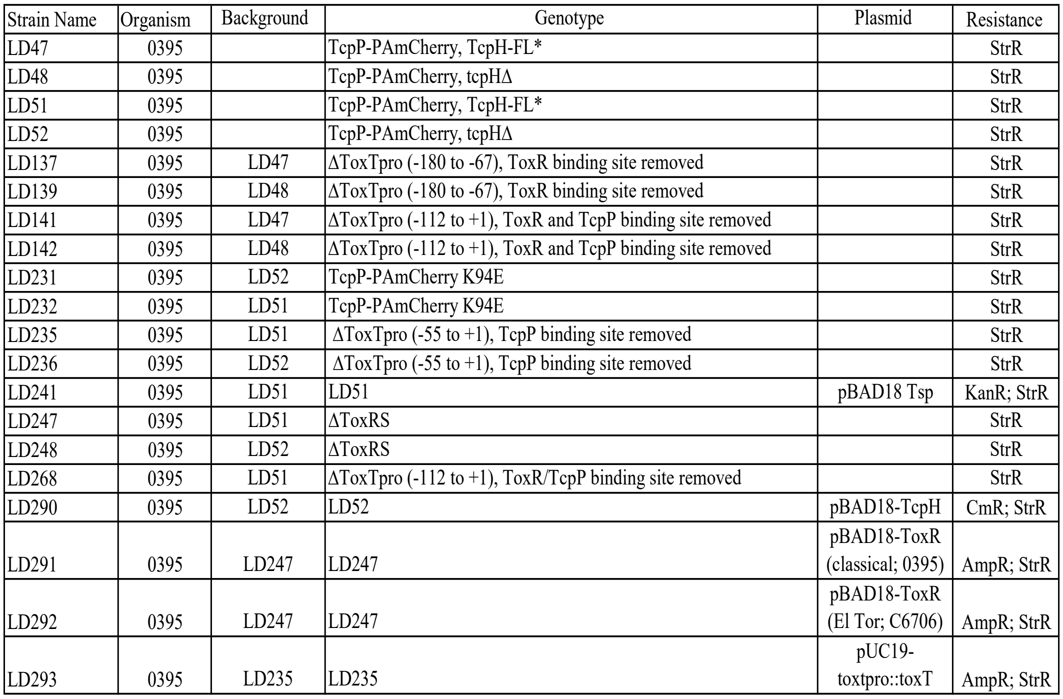


Figure 4.3: Baseline Dynamics of TcpP-PAmCherry. A: SMAUG analysis of the LD51 data set displaying the number of distinct diffusive states in the model vs iteration. The algorithm initializes at a large number of states, K, and converges to a most probable value of $K = 3$. **B**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis. SMAUG identifies three distinct clusters within the dataset. **C**: Cartoon depiction of the full SMAUG results for this dataset, including transition probabilities. Bubble colors correspond to the term colors in **B** and bubble sizes represent the weight fractions. Arrows between bubbles indicate the mean of the transition matrix elements for transitions between those terms. Dashed lines indicate transition probabilities that are negligible. Figure reproduced form [75].

studies to understand what biological role each of the identified states is.

4.4.4 Dynamics of TcpP-PAmCherry with mutations

To that end we constructed several strains with targeted mutations to assist in identifying the biological roles of the identified states from the SMAUG analysis of the dynamics of TcpP in wild type cells, a list of which can be found in Table 4.1. To begin, we constructed a point mutant in the sequence of TcpP to alter the binding affinity of TcpP for the *toxT* promoter (LD232). Residue lysine 94 lies in the domain of TcpP that interacts with the DNA and in LD232, we have introduced a charge-inversion mutation in the protein sequence as lysine has been swapped for glutamic acid, K94E. As K94 lies in the region that interacts with the DNA, we expect this mutation to reduce any states with DNA binding. We then imaged these cells under virulence inducing conditions and collected 24,754 steps. SMAUG analysis returns a most probable three state model with mean with diffusion coefficients of $D = \{0.009, 0.049, 0.469\}\mu m^2/s$ and mean weight fractions of $\pi = \{0.09, 0.54, 0.37\}$ (Fig. 4.4A&B). As expected the slowest state, which we hypothe-



Strain Name	Organism	Background	Genotype	Plasmid	Resistance
LD47	0395		TcpP-PAmCherry, TcpH-FL*		StrR
LD48	0395		TcpP-PAmCherry, tcpHΔ		StrR
LD51	0395		TcpP-PAmCherry, TcpH-FL*		StrR
LD52	0395		TcpP-PAmCherry, tcpHΔ		StrR
LD137	0395	LD47	ΔToxTpro (-180 to -67), ToxR binding site removed		StrR
LD139	0395	LD48	ΔToxTpro (-180 to -67), ToxR binding site removed		StrR
LD141	0395	LD47	ΔToxTpro (-112 to +1), ToxR and TcpP binding site removed		StrR
LD142	0395	LD48	ΔToxTpro (-112 to +1), ToxR and TcpP binding site removed		StrR
LD231	0395	LD52	TcpP-PAmCherry K94E		StrR
LD232	0395	LD51	TcpP-PAmCherry K94E		StrR
LD235	0395	LD51	ΔToxTpro (-55 to +1), TcpP binding site removed		StrR
LD236	0395	LD52	ΔToxTpro (-55 to +1), TcpP binding site removed		StrR
LD241	0395	LD51	LD51	pBAD18 Tsp	KanR; StrR
LD247	0395	LD51	ΔToxRS		StrR
LD248	0395	LD52	ΔToxRS		StrR
LD268	0395	LD51	ΔToxTpro (-112 to +1), ToxR/TcpP binding site removed		StrR
LD290	0395	LD52	LD52	pBAD18-TcpH	CmR; StrR
LD291	0395	LD247	LD247	pBAD18-ToxR (classical; 0395)	AmpR; StrR
LD292	0395	LD247	LD247	pBAD18-ToxR (El Tor; C6706)	AmpR; StrR
LD293	0395	LD235	LD235	pUC19-toxtpro::toxT	AmpR; StrR

Table 4.1: Strains used in TcpP investigations. Table containing information on the strains used in our investigations of TcpP-PAmCherry dynamics. TcpH-FL* indicates full length TcpH moved downstream of TcpP-PAmCherry and start codon changed from ATG to GTG. StrR indicates Streptomycin resistance, AmpR indicates Ampicillin resistance, KanR indicates Kanamycin resistance and CmR indicates Chloramphenicol resistance.

size to be the bound state of the protein binding to the *toxT* promoter region, decreases in weight fraction to roughly half its previous weight. Additionally, the transition out of this proposed bound state has increased dramatically, likely indicating that the binding affinity of the protein is lowered as it unbinds more quickly. The companion strain, LD231, which contains the LD52 background (*tcpH*Δ), again shows very little fluorescence activation.

We then constructed a strain in which the stretch of the DNA that TcpP is known to bind to as been removed from the genome, LD235. Base pairs -55 through +1 upstream from the *toxt* gene (in the *toxT* promoter region) have been removed, leaving no target for the protein to bind. We imaged these cells under virulence-inducing conditions, collecting 25,492 TcpP-PAmCherry steps. SMAUG analysis returns a most probable two state model with mean with diffusion coefficients of $D = \{0.041, 0.336\}\mu m^2/s$ and mean weight fractions of $\pi = \{0.677, 0.322\}$ (Fig. 4.4C). Without the DNA target, the state with the lowest diffusion coefficient from the previous analyses has been lost (Fig. 4.3B, blue). With the loss of the state from the model, the transition elements simply display a increased transition probability of inter-conversion between the leftover states. (Fig. 4.4D). This result supports our hypothesis that the slowest diffusive state was the DNA-bound form. Again, the companion strain lacking TcpH (LD 236 in Table 4.1) shows little to no activatable fluorescence, matching expectations that without TcpH the fusion TcpP is degraded. Similarly, we also constructed another strain that lacks the entire *toxT* promoter region upstream of the gene. Base pairs -112 through +1 are removed, removing both the TcpP binding site and the ToxR binding site. Preliminary investigations of this new strain (LD268; Table 4.1) has only collected 5,288 steps from imaging experiments, insufficient for a quantitative analysis by SMAUG, though we expect to see similar dynamics as LD235.

To investigate the effect that binding to ToxR has on the dynamics of TcpP, we constructed a strain lacking both ToxR and ToxS, LD247, and its companion strain of the same genetics but additionally lacking TcpH, LD248. LD248 again matched our expectation: without TcpH, TcpP-PAmCherry is degraded and the cells lack activatable fluorescence and so no imaging experiments can be preformed. On the other hand, we imaged LD247 and collected 11,646 steps. SMAUG analysis of this prelimiinary data set returns a most probable three state model, with mean with diffusion coefficients of $D = \{0.011, 0.041, 0.449\}\mu m^2/s$ and mean weight fractions of $\pi = \{0.385, 0.514, 0.101\}$ (Fig. 4.4E). Interestingly, the state with the lowest diffusion coefficient increases in weight fraction while the state with the highest diffusion coefficient decreases significantly when compared to the TcpP dynamics in the wild type cells. Also of note, there exists for TcpP-PAmCherry in LD247 a transition directly from the state with the highest diffusion coefficient to the lowest (Fig. 4.4F). This difference between strain LD51 and strain LD247 suggests that the highest diffusive state is TcpPH bound to ToxRS (Fig. 2.6, at least in part, and that this binding helps shuttle TcpP through the intermediate state and not directly to the bound state, perhaps as part of a regulatory function for ToxRS in the virulence pathway [42]. More experiments are needed to obtain good statistics for this strain.

4.4.5 Future work

The role of ToxR in TcpP dynamics will be further investigated by imaging two new strains, LD 291 and LD292, which contain plasmids carrying the *toxR* gene from both the classical and El Tor biotypes, respectively (Table 4.1). These proteins will help uncover what role the ToxR is playing in the dynamics of TcpP while also helping to uncover if there are differences between the biotypes. However, these strains, along with LD290 and LD293, are new and growth conditions are still being optimized. Despite containing the antibiotic resistance gene on the plasmids listed, the growth of these strains in media with the antibiotic added has been stunted and cellular morphology under the microscope is badly changed as a majority of the cells are vastly oversized, indicating that, while the cells can grow, they are not healthy. Our lab has not used these antibiotics before in *V. cholerae*, and if the growth defects continue we will ask our collaborators to swap out the ampicillin/chloramphenicol resistance genes for kanamycin, which we know to not cause growth issues in this system [26]. I have included these strains here both for completeness and also to emphasize that this project is ongoing.

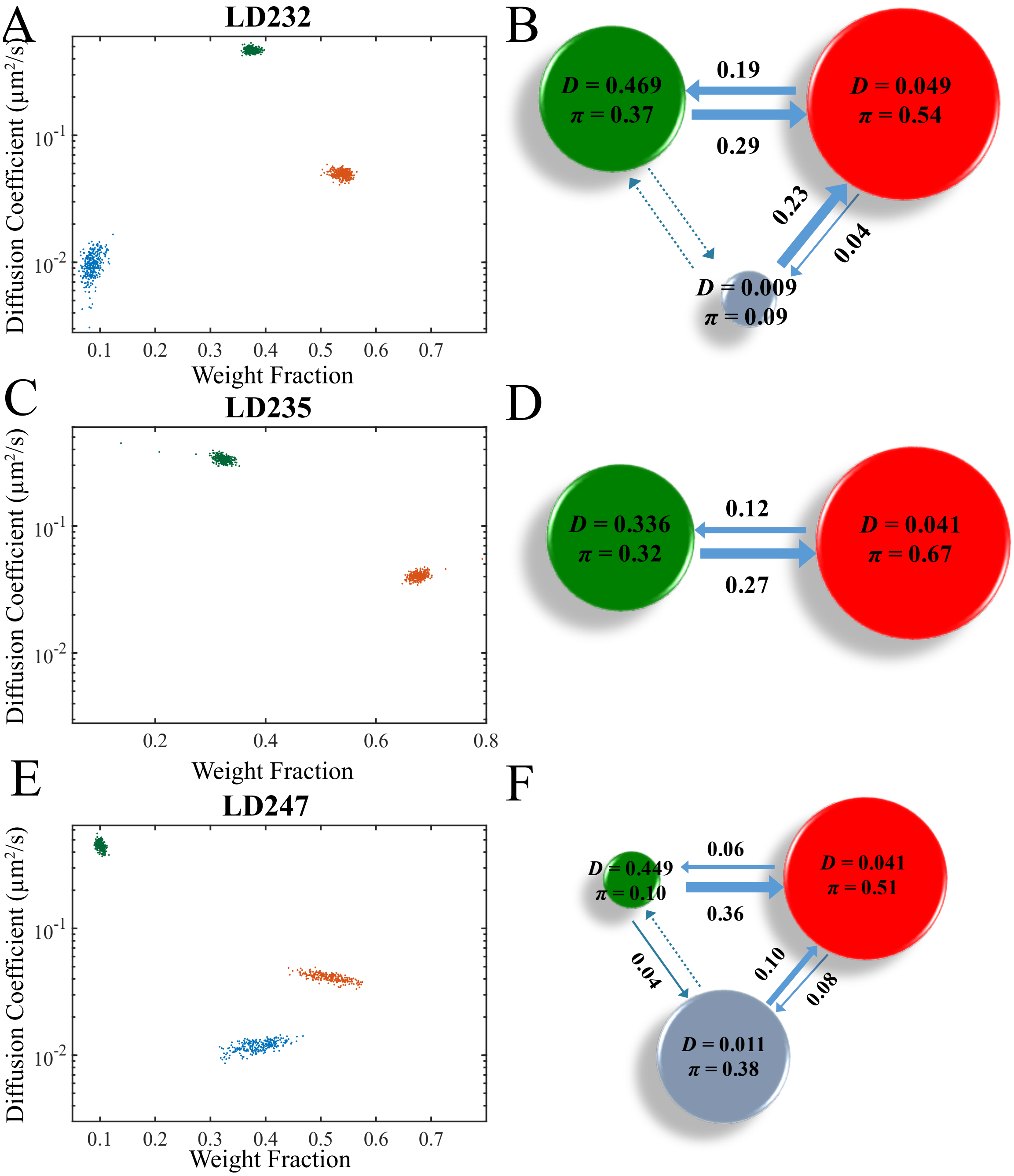


Figure 4.4

Figure 4.4 *(previous page)*: Dynamics of TcpP-PAmCherry under mutations. A: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for LD232. SMAUG identifies three distinct clusters within the data set. **B**: Cartoon depiction of the full SMAUG results for this data set.**C**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for LD235. SMAUG identifies only two distinct clusters within the data set. **D**: Cartoon depiction of the full SMAUG results for this data set. **E**: Diffusion coefficient and weight fraction estimates from the output of the SMAUG analysis for LD247. SMAUG identifies only two distinct clusters within the data set. **F**: Cartoon depiction of the full SMAUG results for this data set.

4.5 Conclusions

The timing and organization of bacterial virulence is a complex problem. Cells must combine various competing signals from their environment about whether or not to induce the virulence cascade, into a coherent, single action for the cell. The timing of this action also needs to be precise if the bacterial community is going to grow and expand. In the case of *Vibrio cholerae*, this process is ultimately under the control of TcpP. The work presented in this chapter has utilized SPT techniques to observe the dynamics of TcpP inside living cells under virulence-inducing conditions. We then used the SMAUG algorithm in order to find the hidden states that exist within our tracking data and uncover biological function. It was found that TcpP exists under virulence inducing conditions in three distinct diffusive states. Then, using the information from a series of mutation studies, we can propose a model of dynamics of the system. TcpP exists mainly in the intermediate state, which we believe is a monitoring state. However, this state is not one that directly binds DNA, as mutations in the DNA binding region of TcpP did not affect this state. It is only from the intermediate state that TcpP can bind to the ToxT promoter region and initiate the virulence cascade. The slowest diffusive state is this state of TcpP bound to the DNA, presumably at the *toxT* promoter region, though that will be investigated in future studies. Furthermore, ToxR is sequestering a portion of the TcpP molecules away from the intermediate state and into the most diffusive state, which is a freely diffusive state for the protein complex of at least TcpP and ToxR, though it can be assumed TcpH and ToxS are likely present as well. Combining the knowledge that the presence of ToxR helps increase the production of CTX with the data presented here, suggests that the role of moderator of ToxT production is preformed by ToxR, as it can both sequester TcpP away from, and enhance the binding of, TcpP to the *toxT* promoter region. Further investigations into this system are planned, but overall the work presented in this chapter lays a solid foundation for understanding this system and provides information that will guide future experiments. As discussed in previous chapters, the combination of biochemical techniques, single-particle tracking experiments and SMAUG analysis is a powerful tool for investigating complex biological systems.

CHAPTER V

Conclusions and Future Directions

5.1 Introduction

The overarching goal of this Thesis was to expand the scope and quality of single-particle tracking experiments and analysis. In the preceding chapters, I presented the theory and construction of a new analysis method for single-particle tracking (SPT) data and my investigations into the dynamics of several complex biological systems using this new method. In this final chapter, I will review the conclusions and impacts that are drawn from each of the chapters and discuss some possible future directions for these projects. Finally, I will wrap up with some overarching conclusions.

5.2 SMAUG

In Chapter 2, I presented the rationale, theory, validation and some applications of a new analysis method for SPT data that we titled Single-Molecule Analysis by Unsupervised Gibbs (SMAUG) [75]. By changing the analysis framework from a curve-fitting based method into a non-parametric Bayesian method, the conclusions drawn gain an increased measure of mathematical rigor while removing possible supervisory biases. This method was shown to be both precise and accurate in estimating a variety of of parameters of interest and, when applied to live-cell tracking data, SMAUG is capable of uncovering the dynamics of a system. Using SMAUG we uncovered a three state dynamical system for bacterial virulence in *Vibrio cholerae* and found the timing and dynamical changes involved in B-cell activation by antigen stimulation. Both of these biological results provide new information for the field and show the utility of the SMAUG algorithm.

A method like SMAUG is required if SPT experiments are going to continue to be applied to ever more complex systems. SMAUG is an excellent first step in expanding the scope and quality of information gained from SPT data and further work will push this ability even further. The future directions I envision for this project are split into short- and long-term directions. In the short-term, the SMAUG code itself could be enhanced to provide better functionality and speed. For example, the algorithm itself has no method for self-determining when the analysis should be completed, instead the algorithm is simply allowed to run for an extremely long duration and then stopped at a pre-defined number of iterations to ensure that the sampler has converged on the most probable values. This setup takes extra time and computing power. One method by which the code could be altered to directly probe convergence is through the introduction of independent Markov chains analyzing the same data set separately. Then, every so often, the values of the chains are compared and the variances between chains (inter) and inside each chain (intra) calculated and compared to some cutoff criteria. Gelman *et al.* have a calculation and method for determining a cutoff in Reference [99]. However, adding addition independent chains will increase the time required for analysis and so optimizing the code for speed would need to be a secondary priority. Recently, a paper describing a method for using GPUs for accelerating a Gibbs sampler was published and would greatly assist in reducing the time required for analysis [100]. Such improvements to the SMAUG code itself, along with others such as implementing GUIs or other ease-of-use improvements, are worthwhile short-term directions for this project.

A more long-term direction for this project would be to switch our entire analysis pipeline over to a Bayesian framework. The SMAUG code works on the very end of our analysis of SPT data: it will estimate the most probable parameter values for a given input data set. However, if the quality of the input data is poor, then the estimates will not be accurate. We currently use curve-fitting methods for localizing single-molecules to sub-diffraction positions and we use an energy minimization algorithm for connecting those localizations into trajectories. Both our fitting and tracking steps could, in theory, instead be performed using a Bayesian method, improving the amount and quality of data that is fed to SMAUG while reducing points of user interaction and thus potential bias. Some of the work for this has already be attempted, though until now these have mostly not been implemented as the cost in computing time did not outweigh the benefits of such methods [10]. However, now with increases in computing power and decreases and cost, coupled with new GPU-accelerated analysis methods, these Bayesian based algorithm might be more readily useable. A fully Bayesian approach to SPT analysis would expand the scope and quality of SPT experiments.

5.3 Epigenetic Silencing in Yeast

In Chapter 3, I presented results from an investigation into the dynamics of Swi6, the key protein in epigenetic silencing, in the yeast model system *Schizosaccharomyces pombe*. We found that the dynamics involved are complex and rely on an interplay between the DNA topology and the protein's affinity for its target, its ability to dimerize and to interact with other proteins. We expected this system to be complex from the start and thus knew we would require an analysis method like SMAUG to draw accurate conclusions. Combining SMAUG analysis with genetic knockouts and functional mutations allowed us to investigate the complexities of the system and interrogate the roles that each of the identified states plays inside the cell. Using all of these results, we then proposed model of behavior and interactions for the protein that we can use future experiments to refine and correct. This work advances knowledge in this field by providing a much more refined look at the processes involved and their relation to each other. Such information on the dynamics involved with individual molecules can only be accessed by single-molecules methods and therefore, while many of these biological roles had been proposed previously, their relationship to each other and the order in which they occur remained unclear.

This project is ongoing as we continue to test more of our hypotheses and refine our model using targeted mutations. While we are close to wrapping up the first stage of the exploration into this system, there are many future directions that will be explored, in the near and long term. In the short term, this project will continue to look at the dynamics of Swi6 inside living cells in a variety of conditions and genetic backgrounds. Several strains are being constructed currently that we think will shed more light on the roles of the identified states. One such strain is a double point mutant that has been shown *in vitro* to increase the amount of oligomerization between neighboring Swi6 molecules [78]. Additionally, a triple knockout mutant, *mst2*Δ*clr3*Δ*gcn5*Δ, is being constructed. This triple knockout strain will completely abolish both the histone acetlyation and histone methylation pathways, providing a system in which "naked" histones are the only histones present. Both of these strains, and others that will follow, will continue to provide evidence for the biological function of the various identified states present in Swi6.

In the longer term, this project will expand to observing epigenetic changes as they happen in real time. Using microfluidics chambers, our collaborators in the Ragunathan lab have shown that they can image the same yeast cell as it grows and divides for up to a week. Thus we envision watching the same cell over time to see how its epigenetic state changes with generational time, either using single-molecule methods or more conventional bulk fluorescence techniques. The dynamics of Swi6-PAmCherry fusions inside the cell could provide a real-time readout of the epigenetic state of that cell. Alternatively, genes for fluorescent proteins could be encoded into silenced regions and the presence/absence of fluorescent signal used as a readout of the epigenetic state of the cell. Investigations into the timing or cell-cycle dependence of epigenetic changes or of the effect of stress on epigenetic patterning could be investigated with such a setup. In addition, this project could also investigate the dynamics of not just Swi6, but any of the other proteins involved in the silencing or anti-silencing pathways, such as we did in Chapter 3 for the Clr3-PAmCherry strain. Each of the proteins involved in the epigenetic silencing or anti-silencing pathways is a potential target for SPT studies and analysis, using the analysis and methods outlined in Chapter 3 as a guide. The field of epigenetics is relatively new but growing quickly and so there are many unanswered questions that can be investigated by this approach and just as many potential future directions for this study to take.

5.4 Dynamics of Proteins Involved in Bacterial Virulence

In Chapter 4, I presented work from our further investigations into the dynamics of bacterial virulence using the model system *Vibrio cholerae*, the causative agent of the disease cholera. We used a TcpP-PAmCherry fusion to track the molecule in real time and analyzed the trajectories with SMAUG. We found that the system is complex under virulence-inducing conditions and used targeted mutations to discover biological roles for the identified diffusion states. Our lab has been working on observing the dynamics of TcpP inside *V. cholerae* for a long time. Only recently have we constructed strains in which the protein fusion is inserted into the genome and lacks any non-intended mutations. Using these strains, along with other mutations, we have been able to identify the biochemical roles of the distinct diffusive states identified from the TcpP dynamics in the wildtype cells.

Using the work presented in Chapter 4, others can continue the investigation of this system, having a solid baseline of results from which to build a more complete model for the binding of TcpP. One future plan is to observe TcpP dynamics while the cells are in conditions that have been proposed as factors for virulence activation, such as high levels of bile in the media or the presence of quorum sensing auto-inducers [55]. By watching TcpP dynamics under these conditions the effect of these molecules can be directly observed. Another avenue of interest would be to fuse ToxR with a label and watch its dynamics under both virulence-inducing and non-inducing conditions, as ToxR is not degraded like TcpP is under non-inducing conditions. This information could help illuminate both the cause and the timing of the switch from a non-virulent state to the virulent state. In addition, investigating the differences between the biotypes would be very illuminating. If TcpP dynamics in El Tor are sufficiently different from the classical strain, this difference could help explain the milder symptoms that El Tor presents in people with the disease and lead to better understanding of how to combat the more deadly strains. For any and all of these proposed ideas, the work presented in Chapter 4 serves as both a guide in how to conduct the experimental design and as a basis to which further studies can be compared.

5.5 Overarching Conclusions

Throughout this Thesis, I have shown that single-molecule methods are powerful tools for biological investigations as they provide the unique advantage of directly observing molecules in real time inside living cells. By capturing, measuring and analyzing the motion of single molecules, we can directly probe biological functions as they occur. This advantage can be wasted however if care is not taken to analyze the resulting data in a mathematically rigorous way. The work presented here shows that while biological systems can be complex, a wealth of information can be gained when the analysis is done carefully and thoughtfully. The work presented in this Thesis includes the theory, application and validation of a novel method of analysis for single-particle tracking data, pushing the state of the art for analysis into a more rigorous mathematical regime. Additionally, this Thesis includes many applications of this method to biological systems from which novel information is gained that provide a mechanistic look into the behavior of the system while also providing a basis for further investigations. The novel combination of single-molecule dynamics studies, biochemical and molecular genetic techniques and advanced mathematical analysis methods provide researchers with a potent tool for investigations into a myriad of outstanding questions. The work presented here lays the ground work for those future studies, where investigators use sophisticated biophysical methods to answer outstanding questions in cell biology.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Wollman Adam J. M., Nudd Richard, Hedlund Erik G., Leake Mark C., "From Animaculum to single molecules: 300 years of the light microscope", *Open Biology* **5**, 150019 (2015). DOI: 10.1098/rsob.150019.

[2] K. Pogliano, E. Harry, R. Losick, "Visualization of the subcellular location of sporulation proteins in Bacillus subtilis using immunofluorescence microscopy.", *Molecular microbiology* **18**, 459 (1995). DOI: 10.1111/j.1365-2958.1995.mmi_18030459.x.

[3] J. Maddock, L. Shapiro, "Polar location of the chemoreceptor complex in the Escherichia coli cell", *Science* **259**, 1717 (1993). DOI: 10.1126/science.8456299.

[4] D. E. Wolf, *Methods in Cell Biology* (Elsevier, 2013), vol. 114, pp. 69–97.

[5] O. Shimomura, F. H. Johnson, Y. Saiga, "Extraction, Purification and Properties of Aequorin, a Bioluminescent Protein from the Luminous Hydromedusan, Aequorea", *Journal of Cellular and Comparative Physiology* **59**, 223 (1962). DOI: 10.1002/jcp.1030590302.

[6] R. Y. Tsien, "The Green Fluorescent Protein", *Annual Review of Biochemistry* **67**, 509 (1998). DOI: 10.1146/annurev.biochem.67.1.509.

[7] A. N. Kapanidis, S. Uphoff, M. Stracy, "Understanding Protein Mobility in Bacteria by Tracking Single Molecules", *Journal of Molecular Biology* **430**, 4443 (2018). DOI: 10.1016/j.jmb.2018.05.002.

[8] E. A. Rodriguez, *et al.*, "The Growing and Glowing Toolbox of Fluorescent and Photoactive Proteins", *Trends in Biochemical Sciences* **42**, 111 (2017). DOI: 10.1016/j.tibs.2016.09.010.

[9] G. B. Airy, "On the Diffraction of an Object-glass with Circular Aperture", *Transactions of the Cambridge Philosophical Society* **5**, 283 (1835).

[10] A. Lee, K. Tsekouras, C. Calderon, C. Bustamante, S. Presse, "Unraveling the Thousand Word Picture: An Introduction to Super-Resolution Data Analysis", *Chemical reviews* **117**, 7276 (2017). DOI: 10.1021/acs.chemrev.6b00729.

[11] F. V. Subach, *et al.*, "Photoactivatable mCherry for high-resolution two-color fluorescence microscopy", *Nature Methods* **6**, 153 (2009). DOI: 10.1038/nmeth.1298.

[12] W. E. Moerner, L. Kador, "Optical detection and spectroscopy of single molecules in a solid", *Physical Review Letters* **62**, 2535 (1989). DOI: 10.1103/PhysRevLett.62.2535.

[13] E. Brooks Shera, N. K. Seitzinger, L. M. Davis, R. A. Keller, S. A. Soper, “Detection of single fluorescent molecules”, *Chemical Physics Letters* **174**, 553 (1990). DOI: 10.1016/0009-2614(90)85485-U.

[14] H. H. Tuson, J. S. Biteen, “Unveiling the Inner Workings of Live Bacteria Using Super-Resolution Microscopy”, *Analytical Chemistry* **87**, 42 (2015). DOI: 10.1021/ac5041346.

[15] D. O. Holland, M. E. Johnson, “Stoichiometric balance of protein copy numbers is measurable and functionally significant in a protein-protein interaction network for yeast endocytosis”, *PLoS Computational Biology* **14** (2018). DOI: 10.1371/journal.pcbi.1006022.

[16] I. E. G. Morrison, C. M. Anderson, G. N. Georgiou, R. J. Cherry, “Measuring diffusion coefficients of labelled particles on cell surfaces by digital fluorescence microscopy”, *Biochemical Society Transactions* **18**, 938 (1990). DOI: 10.1042/bst0180938.

[17] C. M. Anderson, G. N. Georgiou, I. E. G. Morrison, G. V. W. Stevenson, R. J. Cherry, “Tracking of cell surface receptors by fluorescence digital imaging microscopy using a charge-coupled device camera” **101**, 12 (1992).

[18] E. Betzig, *et al.*, “Imaging Intracellular Fluorescent Proteins at Nanometer Resolution”, *Science* **313**, 1642 (2006). DOI: 10.1126/science.1127344.

[19] S. T. Hess, T. P. K. Girirajan, M. D. Mason, “Ultra-High Resolution Imaging by Fluorescence Photoactivation Localization Microscopy”, *Biophysical Journal* **91**, 4258 (2006). DOI: 10.1529/biophysj.106.091116.

[20] M. J. Rust, M. Bates, X. Zhuang, “Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)”, *Nature Methods* **3**, 793 (2006). DOI: 10.1038/nmeth929.

[21] J. S. Biteen, “Extending the tools of single-molecule fluorescence imaging to problems in microbiology”, *Molecular Microbiology* **85**, 1 (2012). DOI: 10.1111/j.1365-2958.2012.08089.x.

[22] A. Kusumi, T. A. Tsunoyama, K. M. Hirosawa, R. S. Kasai, T. K. Fujiwara, “Tracking single molecules at work in living cells”, *Nat Chem Biol* **10**, 524 (2014).

[23] A. Gahlmann, W. E. Moerner, “Exploring bacterial cell biology with single-molecule tracking and super-resolution imaging”, *Nature Reviews Microbiology* **12**, 9 (2014). DOI: 10.1038/nrmicro3154.

[24] J. Munkres, “Algorithms for the Assignment and Transportation Problems”, *Journal of the Society for Industrial and Applied Mathematics* **5**, 32 (1957). DOI: 10.1137/0105003.

[25] P. R. Selvin, T. Ha, *Single-Molecule Techniques: A Laboratory Manual* (Cold Spring Harbor Labratory Press, Cold Spring Harbor, New York, 2008).

[26] B. L. Haas, J. S. Matson, V. J. DiRita, J. S. Biteen, "Single-molecule tracking in live Vibrio cholerae reveals that ToxR recruits the membrane-bound virulence regulator TcpP to the toxT promoter", *Molecular microbiology* **96**, 4 (2015). DOI: 10.1111/mmi.12834.

[27] D. J. Rowland, J. S. Biteen, "Measuring molecular motions inside single cells with improved analysis of single-particle trajectories", *Chemical Physics Letters* **674**, 173 (2017). DOI: 10.1016/j.cplett.2017.02.052.

[28] D. Ernst, J. Koehler, "Measuring a diffusion coefficient by single-particle tracking: statistical analysis of experimental mean squared displacement curves", *Physical Chemistry Chemical Physics* **15**, 845 (2013). DOI: 10.1039/c2cp43433d.

[29] H. Qian, M. P. Sheetz, E. L. Elson, "Single particle tracking. Analysis of diffusion and flow in two-dimensional systems.", *Biophysical Journal* **60**, 910 (1991).

[30] A. Robson, K. Burrage, M. C. Leake, "Inferring diffusion in single live cells at the single-molecule level.", *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20120029 (2013). DOI: 10.1098/rstb.2012.0029.

[31] M. J. Saxton, K. Jacobson, "Single-particle tracking: applications to membrane dynamics", *Annual Review of Biophysics and Biomolecular Structure* **26**, 373 (1997). DOI: 10.1146/annurev.biophys.26.1.373.

[32] M. A. Deverall, *et al.*, "Membrane lateral mobility obstructed by polymer-tethered lipids studied at the single molecule level", *Biophysical Journal* **88**, 1875 (2005). DOI: 10.1529/biophysj.104.050559.

[33] F. Persson, M. Linden, C. Unoson, J. Elf, "Extracting intracellular diffusive states and transition rates from single-molecule tracking data", *Nature Methods* **10**, 265 (2013). DOI: 10.1038/NMETH.2367.

[34] P. K. Koo, M. Weitzman, C. R. Sabanaygam, K. L. v. Golen, S. G. J. Mochrie, "Extracting Diffusive States of Rho GTPase in Live Cells: Towards In Vivo Biochemistry", *Plos Computational Biology* **11**, e1004297 (2015). DOI: 10.1371/journal.pcbi.1004297.

[35] K. E. Hines, "A Primer on Bayesian Inference for Biophysical Systems", *Biophysical journal* **108**, 2103 (2015). DOI: 10.1016/j.bpj.2015.03.042.

[36] S. Sharma, "Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy", *Annual Review of Astronomy and Astrophysics* **55**, 213 (2017). DOI: 10.1146/annurev-astro-082214-122339.

[37] I. Sgouralis, S. Presse, "An Introduction to Infinite HMMs for Single-Molecule Data Analysis", *Biophysical journal* **112**, 2021 (2017). DOI: 10.1016/j.bpj.2017.04.027.

[38] K. Hines, J. Bankston, R. Aldrich, "Analyzing Single-Molecule Time Series via Nonparametric Bayesian Inference", *Biophysical Journal* **108**, 540 (2015). DOI: 10.1016/j.bpj.2014.12.016.

[39] J. W. Yoon, A. Bruckbauer, W. J. Fitzgerald, D. Klenerman, "Bayesian inference for improved single molecule fluorescence tracking.", *Biophysical journal* **94**, 4932 (2008). DOI: 10.1529/biophysj.107.116285.

[40] M. E. Beheiry, *et al.*, "A Primer on the Bayesian Approach to High-Density Single-Molecule Trajectories Analysis", *Biophysical journal* **110**, 1209 (2016). DOI: 10.1016/j.bpj.2016.01.018.

[41] A. J. Berglund, "Statistics of camera-based single-particle tracking", *Physical Review E* **82**, 011917 (2010). DOI: 10.1103/PhysRevE.82.011917.

[42] J. S. Matson, J. H. Withey, V. J. DiRita, "Regulatory Networks Controlling Vibrio cholerae Virulence Gene Expression", *Infection and Immunity* **75**, 5542 (2007). DOI: 10.1128/IAI.01094-07.

[43] M. B. Stone, S. A. Shelby, M. F. Nunez, K. Wisser, S. L. Veatch, "Protein sorting by lipid phase-like domains supports emergent signaling function in B lymphocyte plasma membranes", *Elife* **6**, e19891 (2017). DOI: 10.7554/eLife.19891.

[44] E. Edwald, M. B. Stone, E. M. Gray, J. Wu, S. L. Veatch, "Oxygen Depletion Speeds and Simplifies Diffusion in HeLa Cells", *Biophysical journal* **107**, 1873 (2014). DOI: 10.1016/j.bpj.2014.08.023.

[45] S. Geman, D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721 (1984). DOI: 10.1109/TPAMI.1984.4767596.

[46] A. E. Gelfand, A. F. M. Smith, "Sampling-Based Approaches to Calculating Marginal Densities", *Journal of the American Statistical Association* **85**, 398 (1990). DOI: 10.2307/2289776.

[47] L. Tierney, "Markov-Chains for Exploring Posterior Distributions", *Annals of Statistics* **22**, 1701 (1994). DOI: 10.1214/aos/1176325750.

[48] M. Linden, V. Curic, E. Amselem, J. Elf, "Pointwise error estimates in localization microscopy", *Nature Communications* **8**, 15115 (2017). DOI: 10.1038/ncomms15115.

[49] C. E. Rasmussen, "The infinite Gaussian mixture model", *Advances in Neural Information Processing Systems* **12**, 554 (2000).

[50] K. S. Karunatilaka, E. A. Cameron, E. C. Martens, N. M. Koropatkin, J. S. Biteen, "Superresolution Imaging Captures Carbohydrate Utilization Dynamics in Human Gut Symbionts", *mBio* **5**, e02172 (2014). DOI: 10.1128/mBio.02172-14.

[51] N. L. Hjort, C. Holmes, P. Muller, S. Walker, *Bayesian Nonparametrics* (Cambridge University Press, New York, 2010).

[52] J. Sethuraman, "A Constructive Definition of Dirichlet Priors", *Statistica Sinica* **4**, 639 (1994).

[53] S. G. Walker, “Sampling the Dirichlet mixture model with slices”, *Communications in Statistics-Simulation and Computation* **36**, 45 (2007). DOI: 10.1080/03610910601096262.

[54] M. Ali, A. R. Nelson, A. L. Lopez, D. A. Sack, “Updated Global Burden of Cholera in Endemic Countries.”, *PLoS Neglected Tropical Diseases* **9.6** (2015).

[55] M. Yang, *et al.*, “Bile salt-induced intermolecular disulfide bond formation activates Vibrio cholerae virulence”, *Proceedings of the National Academy of Sciences* **110**, 2348 (2013). DOI: 10.1073/pnas.1218039110.

[56] V. Seda, M. Mraz, “B-cell receptor signalling and its crosstalk with other pathways in normal and malignant cells”, *European journal of haematology* **94**, 193 (2015). DOI: 10.1111/ejh.12427.

[57] N. Monnier, *et al.*, “Bayesian Approach to MSD-Based Analysis of Particle Motion in Live Cells”, *Biophysical Journal* **103**, 616 (2012). DOI: 10.1016/j.bpj.2012.06.029.

[58] D. E. Gottschling, “Summary: EpigeneticsâĂŤfrom Phenomenon to Field”, *Cold Spring Harbor Symposia on Quantitative Biology* **69**, 507 (2004). DOI: 10.1101/sqb.2004.69.507.

[59] D. Moazed, “Mechanisms for the Inheritance of Chromatin States”, *Cell* **146**, 510 (2011). DOI: 10.1016/j.cell.2011.07.013.

[60] E. Heitz, “Das Heterochromatin der Moose”, *Jahrb Wiss Botanik* **69**, 762 (1928).

[61] H. J. Muller, “Types of visible variations induced by X-rays inDrosophila”, *Journal of Genetics* **22**, 299 (1930). DOI: 10.1007/BF02984195.

[62] E. J. Richards, S. C. R. Elgin, “Epigenetic Codes for Heterochromatin Formation and Silencing: Rounding up the Usual Suspects”, *Cell* **108**, 489 (2002). DOI: 10.1016/S0092-8674(02)00644-X.

[63] S. I. S. Grewal, D. Moazed, “Heterochromatin and Epigenetic Control of Gene Expression”, *Science* **301**, 798 (2003). DOI: 10.1126/science.1086887.

[64] D. Moazed, “Common Themes in Mechanisms of Gene Silencing”, *Molecular Cell* **8**, 489 (2001). DOI: 10.1016/S1097-2765(01)00340-9.

[65] K. Luger, A. W. MÃďder, R. K. Richmond, D. F. Sargent, T. J. Richmond, “Crystal structure of the nucleosome core particle at 2.8 ÃĔ resolution”, *Nature* **389**, 251 (1997). DOI: 10.1038/38444.

[66] T. Jenuwein, C. D. Allis, “Translating the Histone Code”, *Science* **293**, 1074 (2001). DOI: 10.1126/science.1063127.

[67] T. Kouzarides, “Chromatin Modifications and Their Function”, *Cell* **128**, 693 (2007). DOI: 10.1016/j.cell.2007.02.005.

[68] J.-i. Nakayama, J. C. Rice, B. D. Strahl, C. D. Allis, S. I. S. Grewal, “Role of Histone H3 Lysine 9 Methylation in Epigenetic Control of Heterochromatin Assembly”, *Science* **292**, 110 (2001). DOI: 10.1126/science.1060118.

[69] S. Rea, *et al.*, “Regulation of chromatin structure by site-specific histone H3 methyltransferases”, *Nature* **406**, 593 (2000). DOI: 10.1038/35020506.

[70] S. I. S. Grewal, S. Jia, “Heterochromatin revisited”, *Nature Reviews Genetics* **8**, 35 (2007). DOI: 10.1038/nrg2008.

[71] D. Canzio, *et al.*, “A conformational switch in HP1 releases auto-inhibition to drive heterochromatin assembly”, *Nature* **496**, 377 (2013). DOI: 10.1038/nature12032.

[72] C. Keller, *et al.*, “HP1swi6 Mediates the Recognition and Destruction of Heterochromatic RNA Transcripts”, *Molecular Cell* **47**, 215 (2012). DOI: 10.1016/j.molcel.2012.05.009.

[73] T. Cheutin, *et al.*, “Maintenance of Stable Heterochromatin Domains by Dynamic HP1 Binding”, *Science* **299**, 721 (2003). DOI: 10.1126/science.1078572.

[74] “Yeast Extract with Supplements (YES)”, *Cold Spring Harbor Protocols* **2016**, pdb.rec091355 (2016). DOI: 10.1101/pdb.rec091355.

[75] J. D. Karslake, *et al.*, “SMAUG: Analyzing single-molecule tracks with nonparametric Bayesian statistics”, *bioRxiv* p. 578567 (2019). DOI: 10.1101/578567.

[76] L. C. Bryan, *et al.*, “Single-molecule kinetic analysis of HP1-chromatin binding reveals a dynamic network of histone modification and DNA interactions”, *Nucleic Acids Research* **45**, 10504 (2017). DOI: 10.1093/nar/gkx697.

[77] B. Xhemalce, T. Kouzarides, “A chromodomain switch mediated by histone H3 Lys 4 acetylation regulates heterochromatin assembly”, *Genes & Development* **24**, 647 (2010). DOI: 10.1101/gad.1881710.

[78] D. Canzio, *et al.*, “Chromodomain-Mediated Oligomerization of HP1 Suggests a Nucleosome-Bridging Mechanism for Heterochromatin Assembly”, *Molecular Cell* **41**, 67 (2011). DOI: 10.1016/j.molcel.2010.12.016.

[79] D. L. Mendez, *et al.*, “The HP1a disordered C terminus and chromo shadow domain cooperate to select target peptide partners”, *Chembiochem: A European Journal of Chemical Biology* **12**, 1084 (2011). DOI: 10.1002/cbic.201000598.

[80] P.-C. Li, L. Chretien, J. CÃťtÃľ, T. J. Kelly, S. L. Forsburg, “S. pombe replication protein Cdc18 (Cdc6) interacts with Swi6 (HP1) heterochromatin protein”, *Cell Cycle* **10**, 323 (2011). DOI: 10.4161/cc.10.2.14552.

[81] D. A. Cravioto, “Independent Panel of Experts on the Cholera Outbreak in Haiti” p. 32 (2011).

[82] D. A. Sack, R. B. Sack, C.-L. Chaignat, "Getting Serious about Cholera", *New England Journal of Medicine* **355**, 649 (2006). DOI: 10.1056/NEJMp068144.

[83] R. C. Charles, E. T. Ryan, "Cholera in the 21st century", *Current Opinion in Infectious Diseases* **24**, 472 (2011). DOI: 10.1097/QCO.0b013e32834a88af.

[84] R. A. Finkelstein, *Medical Microbiology*, S. Baron, ed. (University of Texas Medical Branch at Galveston, Galveston (TX), 1996), fourth edn.

[85] D. A. Herrington, *et al.*, "Toxin, toxin-coregulated pili, and the toxR regulon are essential for Vibrio cholerae pathogenesis in humans", *The Journal of Experimental Medicine* **168**, 1487 (1988).

[86] J. G. Conner, J. K. Teschler, C. J. Jones, F. H. Yildiz, "Staying Alive: Vibrio choleraeâĂŹs Cycle of Environmental Survival, Transmission, and Dissemination", *Microbiology Spectrum* **4** (2016). DOI: 10.1128/microbiolspec.VMBF-0015-2015.

[87] D. E. Higgins, E. Nazareno, V. J. DiRita, "The virulence gene activator ToxT from Vibrio cholerae is a member of the AraC family of transcriptional activators.", *Journal of Bacteriology* **174**, 6974 (1992).

[88] B. M. Childers, *et al.*, "Identification of residues critical for the function of the Vibrio cholerae virulence regulator ToxT by scanning alanine mutagenesis", *Journal of Molecular Biology* **367**, 1413 (2007). DOI: 10.1016/j.jmb.2007.01.061.

[89] D. T. Hung, E. A. Shakhnovich, E. Pierson, J. J. Mekalanos, "Small-molecule inhibitor of Vibrio cholerae virulence and intestinal colonization", *Science (New York, N.Y.)* **310**, 670 (2005). DOI: 10.1126/science.1116739.

[90] K. M. Peterson, J. J. Mekalanos, "Characterization of the Vibrio cholerae ToxR regulon: identification of novel genes involved in intestinal colonization", *Infection and Immunity* **56**, 2822 (1988).

[91] V. J. DiRita, J. J. Mekalanos, "Periplasmic interaction between two membrane regulatory proteins, ToxR and ToxS, results in signal transduction and transcriptional activation", *Cell* **64**, 29 (1991). DOI: 10.1016/0092-8674(91)90206-E.

[92] F. Hennecke, A. MÃijller, R. Meister, A. Strelow, S. Behrens, "A ToxR-based two-hybrid system for the detection of periplasmic and cytoplasmic proteinâĂŞprotein interactions in Escherichia coli: minimal requirements for specific DNA binding and transcriptional activation", *Protein Engineering, Design and Selection* **18**, 477 (2005). DOI: 10.1093/protein/gzi053.

[93] V. J. DiRita, M. Neely, R. K. Taylor, P. M. Bruss, "Differential expression of the ToxR regulon in classical and E1 Tor biotypes of Vibrio cholerae is due to biotype-specific control over toxT expression", *Proceedings of the National Academy of Sciences* **93**, 7991 (1996). DOI: 10.1073/pnas.93.15.7991.

[94] J. S. Matson, V. J. DiRita, “Degradation of the membrane-localized virulence activator TcpP by the YaeL protease in Vibrio cholerae”, *Proceedings of the National Academy of Sciences of the United States of America* **102**, 16403 (2005). DOI: 10.1073/pnas.0505818102.

[95] N. A. Beck, E. S. Krukonis, V. J. DiRita, “TcpH influences virulence gene expression in Vibrio cholerae by inhibiting degradation of the transcription activator TcpP”, *Journal of Bacteriology* **186**, 8309 (2004). DOI: 10.1128/JB.186.24.8309-8316.2004.

[96] W. P. Teoh, J. S. Matson, V. J. DiRita, “Regulated intramembrane proteolysis of the virulence activator TcpP in Vibrio cholerae is initiated by the tail-specific protease (Tsp)”, *Molecular Microbiology* **97**, 822 (2015). DOI: 10.1111/mmi.13069.

[97] T. J. Goss, C. P. Seaborn, M. D. Gray, E. S. Krukonis, “Identification of the TcpP-binding site in the toxT promoter of Vibrio cholerae and the role of ToxR in TcpP-mediated activation”, *Infection and Immunity* **78**, 4122 (2010). DOI: 10.1128/IAI.00566-10.

[98] K. Skorupski, R. K. Taylor, “Positive selection vectors for allelic exchange”, *Gene* **169**, 47 (1996). DOI: 10.1016/0378-1119(95)00793-8.

[99] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, *Bayesian Data Analysis* (Chapman and Hall/CRC, 2004), second edn.

[100] A. Terenin, S. Dong, D. Draper, “GPU-accelerated Gibbs sampling: a case study of the Horseshoe Probit model”, *Statistics and Computing* **29**, 301 (2019). DOI: 10.1007/s11222-018-9809-3.