ON AND OFF: FLUORESCENTLY TAGGING DNA AND COUNTING PROTEIN STOICHIOMETRY IN THE CELL

Matthew Joseph Satusky

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry.

Chapel Hill
2020

Approved by:

Dorothy Erie

Matthew Redinbo

Joanna Atkin

Keith Weninger

Tom Kunkel

Kerry Bloom

ABSTRACT

Matthew Joseph Satusky: ON AND OFF: FLUORESCENTLY TAGGING DNA AND COUNTING PROTEIN STOICHIOMETRY IN THE CELL
(Under the direction of Dorothy Erie)

Protein interaction with DNA is vital to cellular function and genome stability; protein-mediated processes include gene regulation and DNA damage repair. These processes have been associated with conformational changes in DNA structures and recruitment of multiple proteins to a given site. Fluorescence microscopy is a powerful tool for measuring these interactions both *in vitro* and *in vivo*. In this thesis, I focus on two limitations of fluorescence microscopy: attachment of fluorescent dyes to oligonucleotides and analysis of fluorescence images in live cells.

Oligonucleotides modified with fluorescent dyes are useful for techniques such as Förster resonance energy transmission (FRET) that measure conformational changes in DNA. When end-labeling is not an option, the dye must be placed internally through phosphoramidite backbone labeling or a modified thymine. Phosphoramidite dyes are sequence-independent, but incorporation may create a gap in the backbone. Cyanine dyes incorporated in this manner are anchored in two places, limiting the dipole orientation freedom and possibly affecting FRET values. When sequence specificity is necessary, a thymine in an appropriate location for FRET pairing may not exist. In this work, we characterize a new, cost-efficient method of sequence-independent labeling of the DNA backbone using the common phosphorothioate modification

Stoichiometry of proteins in an interaction is a fundamental characteristic of biochemical functions, however it remains difficult to determine in the nucleus of a live cell in real time. Stochastic photobleaching of fluorescently labeled proteins provides a means of counting molecules, as the loss of fluorescence is quantized and proceeds in a stepwise manner. This powerful method for determining stoichiometry is limited by the high background associated with nuclear proteins in live cells. Here we describe the creation of software pipeline that processes microscope movies to identify cells, locate fluorescent foci, extract pixel intensity information, and quantify photobleaching steps for fluorescently labeled nuclear proteins.

Sometimes science is more art than science, Morty.

-Rick Sanchez

ACKNOWLEDGEMENTS

First and foremost: Thank you, Dorothy, for inviting me to the Island of Misfit Toys. If you didn't take a chance on me, I surely wouldn't have gotten to the end. Also, thank you for giving me the freedom to create a project out of thin air and the guidance to cobble it all together. You have helped me grow into a much more complete scientist. I would also like to thank the rest of my committee, especially Keith Wenninger, who provided valuable technical feedback, and Eric Brustad, who left us too early (for San Diego).

I need to thank my parents, Greg and Mary Jo, and my brother, Mike. You supported me through a decade and a half of college and I couldn't have done it without you. Thank you for listening to me ramble at lengths in mostly scientific jargon, and for always apologizing before asking me when I was going to graduate.

To the Erie Lab: Hunter, thanks for being eerily on the same page as me for basically everything and slumming it with me in the teaching lab. You have been a great friend and roommate throughout. Sarah, thank you for commiserating and tolerating my many intensely curious jokes. I would not have been able to finish without our postdocs, Jackie and Sharonda. Jackie, you get more work done than anyone I've ever met. Sharonda, you are truly an amazing scientist (and scope whisperer). Thank you both; you are sources of inspiration and awe. Thank you Emily, for the bench-side chats, and Caitlin, for making me feel like I am still cool (almost). And Nolan, thanks for being a great GM and a cloning wizard. Also, thank you to Jake and the previous lab members who I never met but wrote code that I incorporated.

Stef, thank you for the midday coffee breaks and sanity checks, along with all of the scientific talks. You were someone I could lean on in tough times. Erinn, thanks for listening to me complain and for keeping me laughing with lewd humor. Parth, our barstool philosophy and science conversations kept some of my cynicism at bay, so thank you for that. Andrew, thanks for being a great roommate and for teaching me so much about game development. Thank you to those who taught me techniques and who collaborated with me: Danielle, for listening to Tycho in the Gen room, and Amy, for purifying protein ridiculously fast. Finally, thank you to everyone I was lucky enough to be friends with during this time, especially my intramural street hockey teammates for bringing the cup home twice.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

24mer	Oligonucleotide containing 24 nucleotides
Å	Angstrom
°C	Degrees Celsius
η^2	Eta-squared
µl	Microliter
µm	Micron
Σ	Sum
χ^2	Chi-squared
AAAS	Association for the Advancement of Science
ATP	Adenosine triphosphate
AU	Arbitrary units
BIDBE	*N,N′*-bis(α-iodoacetyl)-2-2′-dithiobis(ethylamine)
BLAST	Basic local-alignment search tools
bPEG-silane	biotin-poly(ethylene glycol)-silane
BSA	Bovine serum albumen
Ca^{2+}	Calcium ion
$CaCl_2$	Calcium chloride
CCD	Charge-coupled device
CK	Chung-Kennedy algorithm
CURE	Course-based undergraduate research experience
dmGen	*Drosophila melanogaster* structure-specific endonuclease Gen
DMSO	Dimethyl sulfoxide

DNA	Deoxyribonucleic acid
dT	Deoxythymine
DTT	Dithiothreitol
E	FRET efficiency
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
ESI	Electrospray ionization
eGFP	Enhanced green fluorescent protein
emCCD	Electron-multiplying charge-coupled device
FRET	Förster resonance energy transmission
GB	Gigabytes
g/mol	Grams per mole
HCl	Hydrochloride
HJ	Holliday junction
HPLC	High performance liquid chromatography
hr	Hour
I_A	Acceptor fluorescence intensity
I_D	Donor fluorescence intensity
IPTG	Isopropyl-β-D-thiogalactoside
KCl	Potassium chloride
kDa	Kilodaltons
KOAc	Potassium acetate
KOH	Potassium hydroxide

L	Liters
LacI	*E. coli lac* repressor protein
LacI-eGFP	Fusion protein of LacI and eGFP
LacO	*E. coli lac* repressor protein DNA binding site
LacO[#]	Yeast strain containing # copies of the LacO binding site
LCAS	Laboratory Course Assessment Survey
LC-MS	Liquid chromatography/mass spectrometry
m	Image width
MB	Megabytes
mg	Milligram
$MgCl_2$	Magnesium chloride
mg/ml	Milligrams per milliliter
min	Minute
mL	Milliliters
mM	Millimolar
mm	Millimeter
mPEG-silane	methoxy-poly(ethylene glycol)-silane
ms	Milliseconds
MW	Molecular weight
m/z	Mass-to-charge ratio
N	Number of histogram bins
n	Image height
n	Sample size

NA	Numerical aperture
NaCl	Sodium chloride
NaOH	Sodium hydroxide
nM	Nanomolar
nm	Nanometers
nmol	Nanomole
nt	Nucleotides
Oligo	Oligonucleotide
^{32}P	Phosphorus-32 radioactive isotope
PAGE	Polyacrylamide gel electrophoresis
PCAST	President's Council of Advisors on Science and Technology
PCR	Polymerase chain reaction
pH	Potential for hydrogen
PIFE	Protein-induced fluorescence enhancement
PNG	Portable Network Graphics
PS	Phosphorothioate
QEP	Quality enhancement plan
QTOF	Quadrupole time-of-flight
r	Distance between donor and acceptor dye in a FRET experiment
R_0	Förster radius
R^2	Coefficient of determination
RCF	Relative centrifugal force
RNA	Ribonucleic acid

s	Seconds
SACSCOC	Southern Association of Colleges and Schools Commission on Colleges
SD	Standard deviation
SEM	Standard error of the mean
smFRET	Single-molecule Förster resonance energy transmission
SNR	Signal-to-noise ratio
STEM	Science, technology, engineering, and mathematics
TA	Teaching assistant
Taq	*Thermus aquaticus*
TBE	Tris-borate-EDTA
TCEP	Tris(2-carboxyethyl)phosphine
TIFF	Tagged Image File Format
TIRF	Total internal reflection fluorescence
T_m	Melting temperature
Tris	tris(hydroxymethyl)aminomethane
UI	User interface
U/ml	Units per milliliter
UNC	University of North Carolina at Chapel Hill
URM	Underrepresented minority
UV/Vis	Ultraviolet and visual spectrum
V	Volts
w/v	Weight-by-volume

CHAPTER 1: MONITORING PROTEIN-DNA INTERACTIONS IN REAL TIME THROUGH FLUORESCENCE MICROSCOPY

1.1 Introduction

Maintenance of genomic stability is critical to the survival of any living cell. Errors during replication can lead to misincorporation of nucleotides or insertion-deletion loops[1,2]. Exogenous sources of damage, such as ultraviolet radiation or chemical mutagens, can lead to breakage of one or both strands. The cell is equipped with repair mechanisms for these damage scenarios, but sometimes the functions of these proteins are interrupted, leading to increases in genetic mutations[1] and tumorigenesis[3,4]. Furthermore, some chemotherapeutics work through activation of these repair cascades and decreased pathway function would lead to decreased efficacy[1,5]. It is therefore useful to understand the underlying mechanisms of repair pathways. One particularly useful technique for observing repair proteins and their interactions with DNA is fluorescence microscopy. The goal of my research has been to improve fluorescence methods both experimentally and computationally.

The following sections outline the general methodology for fluorescence imaging as it pertains to observing protein-DNA complexes, beginning with a review of different imaging techniques and a brief introduction to *in vitro* and *in vivo* uses. I then introduce labeling methods for nucleic acids *in vitro* and proteins in live cells. Finally, I discuss the general procedure of analyzing microscope images. Later chapters will cover labeling and image analysis in greater depth.

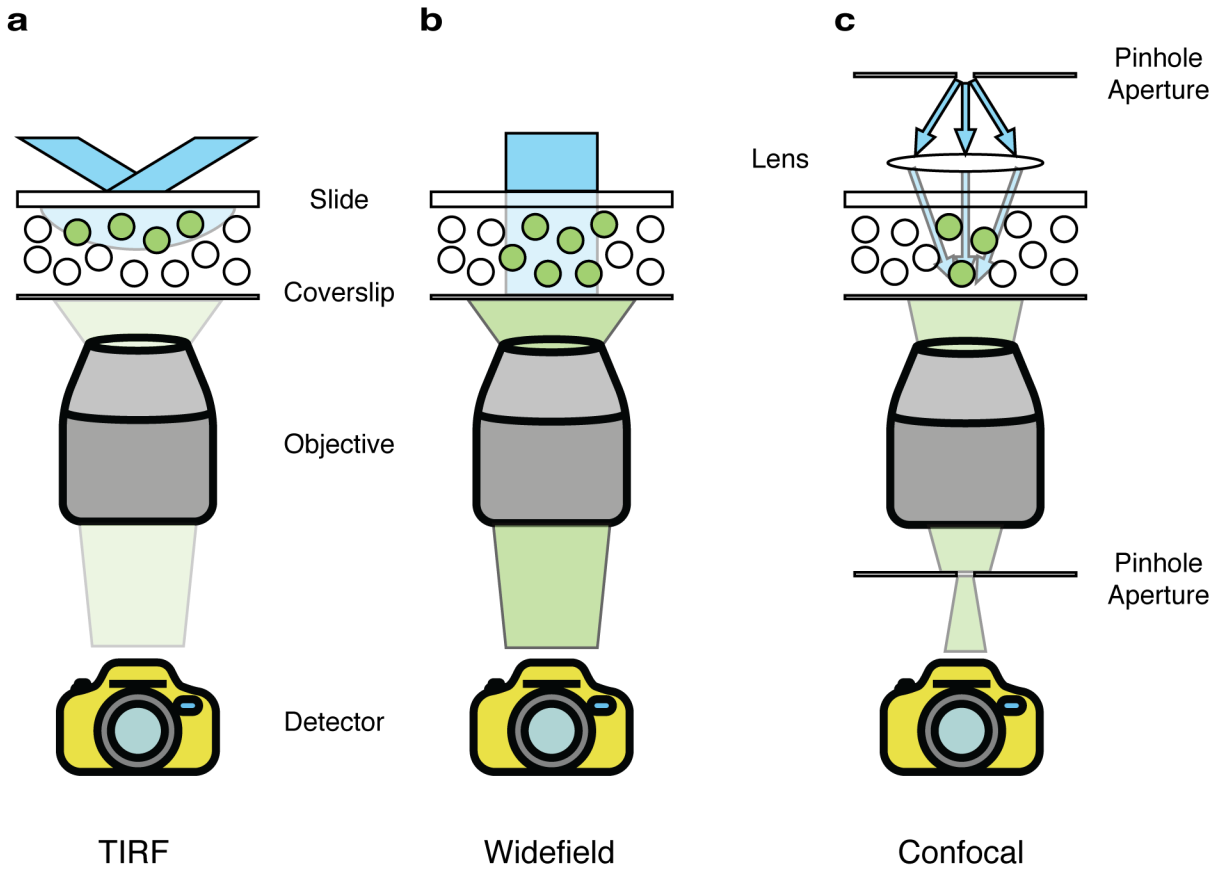

Figure 1.1. Fluorescence imaging techniques. (a) TIRF imaging uses angled illumination such that the excitation beam is reflected off of the slide surface, causing an evanescent wave to propagate through the sample at the surface of the slide, which produces low background. (b) Widefield microscopy illuminates the entire thickness of the sample, creating high signal but also high background. (c) Confocal microscopy uses two pinhole apertures, one to limit the illumination to a single focal plane and one to limit detection of diffused light, increasing the spatial resolution and decreasing the background.

1.2 Observing Protein-DNA Complexes with Fluorescence Microscopy

1.2.1 Imaging techniques

Fluorescence microscopy can be used to monitor protein-DNA complexes both *in vitro*[6,7] and *in vivo*[8]. For *in vitro* experiments, total internal reflection fluorescence (TIRF) microscopy can be utilized to reduce background noise[9]. This technique uses illumination that strikes the slide at an angle such that the excitation beam is completely reflected, rather than passing through the sample. As a result, an evanescent wave propagates through the sample, but the evanescent field decays exponentially, penetrating only ~300 nm beyond the surface and exciting only those fluorophores close to the surface (Fig. 1.1a). This technique eliminates the majority of the signal contribution of molecules away from the slide surface, creating a high signal-to-noise ratio (SNR)[10].

TIRF imaging can be used for *in vivo* experiments as well, but the applications are limited to molecules in or on the membrane. The short decay length of the evanescent wave that makes TIRF ideal for *in vitro* applications prohibits the excitation of molecules beyond the surface, as the wave penetrates only approximately 250 nm. When examining complexes that reside in the nucleus, a method must be used that allows deeper excitation penetration. Although widefield microscopy can be used, the background is much higher and molecules that are closer together may not be differentiable due to the excitation of the whole sample; the width of illumination causes excitation of fluorophores outside of the focal plane and diffracted emission is collected (Fig. 1.1b). Alternatively, confocal microscopy can be employed to reduce the impact of some of these confounds. A confocal microscope uses two pinhole apertures: one for point excitation, which reduces off-target illumination, and one for emission capture, which removes the bulk of diffracted signal by limiting out of plane photons that can reach the

objective (Fig. 1.1c). Both of these improvements are important when trying to quantify fluorescence intensity: analysis of a variety of experiments is typically limited by the SNR and any contribution of nearby molecules that are within the resolution limit may contribute to the signal. The drawback of a confocal setup is that the apertures limit the strength of the captured signal[11].

1.2.2 In vitro fluorescence imaging of protein-DNA interactions

While bulk quantification can give insight into the behavior of a population as a whole, single-molecule techniques allow for the capture of transient changes that otherwise would be masked[1]. The ability to quantitatively (or semi-quantitatively) observe these changes will help elucidate structure-function relationships and the underlying mechanisms of an interaction. One powerful technique for capturing rare or short-lived interactions within protein-DNA complexes is measuring Förster resonance energy transmission (FRET) between fluorophores. FRET occurs when a donor fluorophore transfers energy in a non-radiative fashion to an acceptor fluorophore. This energy transfer is dependent on the intermolecular distance and the Förster radius (R_0) for that particular dye pair, the latter of which includes, among properties, the relative orientations of their dipole moments. If at least one of the dyes has full rotational freedom and the R_0 for the dye pair is known, the distance between the dyes can be calculated from the FRET efficiency[12].

In most single-molecule FRET (smFRET) experiments, one component is affixed to a slide, the surface of which has been modified; when imaging protein-DNA complexes, the DNA is usually attached. This attachment is typically achieved through biotinylation of both the surface and DNA, using a streptavidin bridge between the two[1,12]. Immobilizing the substrate allows interactions to be located and monitored using a CCD camera. Organic fluorescent dyes

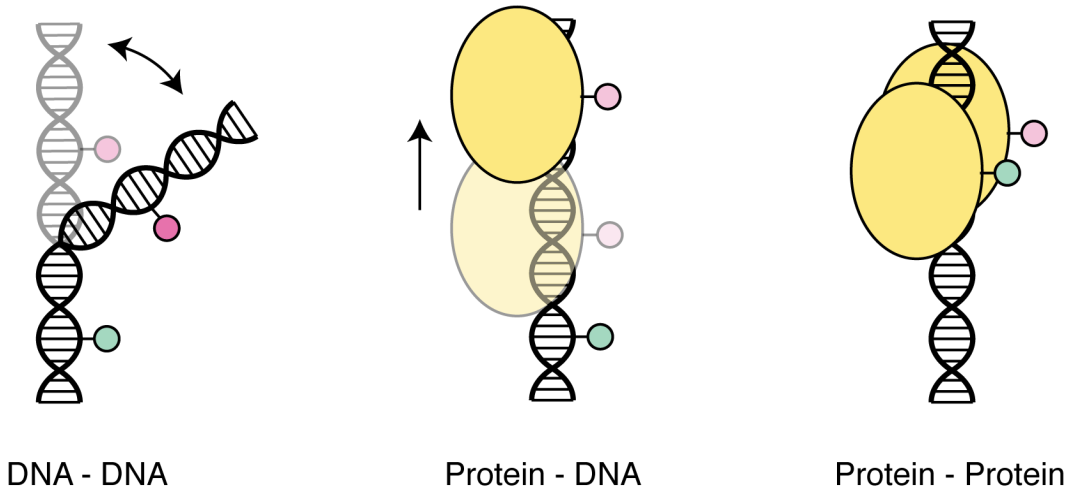

Figure 1.2. Labeling sites determine the information gathered by a FRET experiment. DNA-DNA labeling schemes show conformational change in the DNA, while protein-DNA labeling shows the position of the protein relative to the labeling site on the DNA. Protein-protein can be between different subunits (shown) or at different sites within a single subunit.

are used for *in vitro* smFRET experiments because they tend to be the brightest and most photostable fluorophores, which translates to a higher SNR and longer lifetimes for observation. There are three common labeling schemes for studying protein-DNA interactions: DNA-DNA, which shows conformational changes in the DNA; protein-DNA, which provides information about binding location and protein translocation; and protein-protein, which can illustrate binding of different subunits or conformational changes within the protein (Fig. 1.2)[7,13,14].

1.2.3 In vivo imaging of protein-DNA complexes

While *in vitro* experiments provide invaluable information about the formation, mechanism, and stoichiometry of protein-DNA complexes, these experiments involve removing the components of the reaction from the cellular context. It is important, therefore, to investigate these complexes within their natural cellular environment. Labeling DNA-interacting proteins inside of living cells allows for monitoring of the formation of protein-DNA complexes. These complexes tend to appear as foci, or bright spots of concentrated and relatively stationary protein

that form at a binding site. Fluorescent foci can provide information about the stoichiometry of one or more proteins and colocalization of different proteins.

1.3 Labeling for Fluorescent Imaging

1.3.1 Labeling DNA with fluorescent dyes

DNA synthesis has become cost-efficient and widely commercially available. Most vendors offer oligonucleotides modified with a variety of chemical moieties, including a number of attachment chemistries and dyes. There are currently two available avenues for dye labeling of DNA (or RNA): inclusion of a modified thymine residue (dT) with an organic linking arm containing a reactive group that extends from the 5-methyl group of the base into the major groove or incorporating a linker group into the sugar-phosphate backbone using phosphoramadite chemistry. In this latter method, the fluorophore mimics a 5' phosphate and 3' hydroxyl. There are several considerations to account for when designing a labeling scheme for a DNA substrate.

Because FRET efficiency is indirectly proportional to r^6, where r is the intermolecular distance between the dye pair, labeling site placement is crucial for measuring subtle changes in conformation; placing dyes too close or far from each other decreases the sensitivity of detection[15]. One common backbone incorporation involves anchoring cyanine dyes with two linker arms, which limits the rotational freedom of the dipole moment. If two dyes are attached in this fashion, the dipoles can be aligned in a manner that will yield unpredictable FRET signal with changes in conformation, due to changes in the relative postion of the donor and acceptor dipoles[16]. Recently, a phosphoramidite linker has become available that contains a single linking point (an amine group), which mitigates the dipole concerns, but phosphoramidite insertion in

the backbone may still raise concerns about secondary structure effects as there is a resultant gap in the backbone.

Modified dT residues also contain only a single attachment point to the dye, which makes them an attractive alternative to backbone labeling. These linkers can contain several chemical moieties that allow for the use of a wide variety of available dyes. The caveat of this option is that there must be a thymine in the sequence at the labeling site; when sequence-specificity is a concern, the site must be amenable to an existing dT. Chapter 3 describes a new method for labeling oligonucleotides using a common and cost-efficient modification.

1.3.2 Labeling proteins in vivo

Proteins can be labeled within a living cell through several methods. A widely used method is to use a fluorescent fusion protein, which involves insertion of the fluorescent protein's gene at one end of the gene for the protein of interest, forcing the cellular machinery to create a covalently-linked fluorescent protein. The relative ease of DNA manipulation, combined with the variety of fluorophore colors and lack of further labeling steps, make fusion proteins an attractive option[17–20]. However, fluorescent proteins are much dimmer than organic dyes and there are potential functional effects of attaching large fluorescent proteins[18,20].

Alternatively, proteins can be labeled with organic dyes. These dyes can be incorporated through fusion proteins that contain an active site designed to react with organic probes[21,22], through chelation of reactive dye moieties to peptide tags[23], or through the use of unnatural amino acid incorporation[24]. These methods allow for the use of brighter organic dyes, which

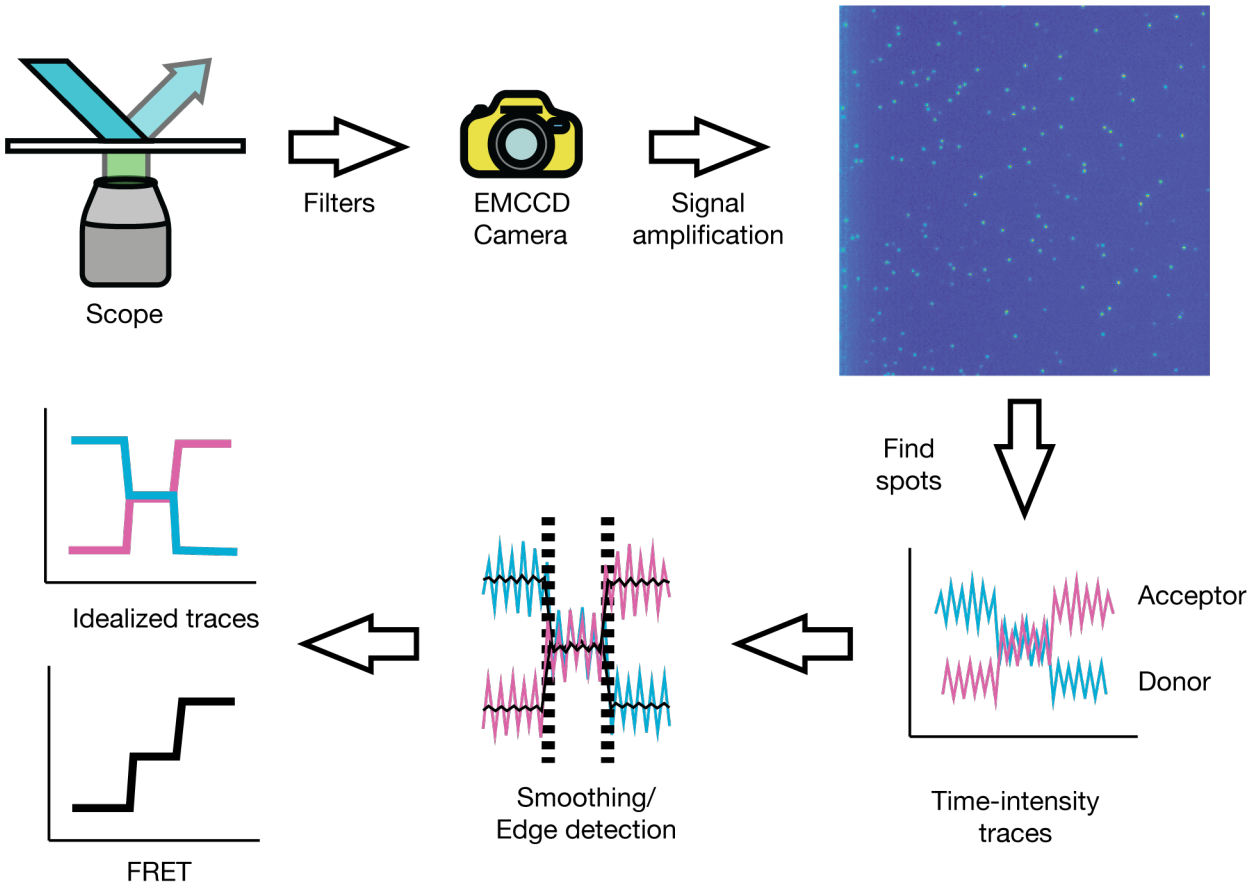

Figure 1.3. Workflow for an smFRET experiment. Individual molecules appear as isolated dots in a TIRF image. Data is collected in two channels, one for donor and one for acceptor, which are aligned to track the donor/acceptor pair together. The spots are located algorithmically using a threshold and time-intensity traces for the donor and acceptor signal are extracted over the length of the movie. Traces are smoothed using a non-linear windowed smoothing algorithm and edges are detected. Finally, FRET efficiency is calculated for the dye pair.

yield a higher SNR, but have accompanying drawbacks. Organic dyes tend to be hydrophobic, which leads to solubility and cell permeability issues, as well as cytoxicity[25,26]. Unnatural amino acids are sometimes difficult to synthesize and can lead to expression difficulties[27,28].

1.4 Analysis of Fluorescence Microscopy Images

1.4.1 Spot picking for generation of time-intensity traces

Fluorescence microscopy experiments that monitor protein-DNA complexes both *in vitro* and *in vivo* (when complexes appear as foci) result in generation of a movie where complexes appear as bright spots (Fig. 1.3). The first analysis step that must occur is the identification of these spots. This step is either performed by hand[29,30] or algorithmically by finding local maxima[7,31,32]. Once spots are identified, pixel intensities can be extracted, background-corrected, and summed to give the total intensity of the spot. Performed over the course of the experiment, the plot of fluorescence against time is a time-intensity trace and is used for analysis of FRET and stoichiometry, as described later. It should be noted that this process is significantly more difficult for *in vivo* experiments, because cells create inherently more background due to refraction and autofluorescence of intracellular molecules. While spot can be fitted to Gaussian distributions to determine pixel intensity contributions, this process has been shown to give nearly identical results as pixel summing at a significant computational cost[33].

1.4.2 Smoothing of time-intensity traces

Cameras used to record the low number of photons generated by fluorophore emission must be extremely sensitive and often include signal amplification. Shot noise from low photon counts, pixel read error, and the physical process of electronic signal amplification result in

systematic noise in time-intensity traces[34]. This noise must be removed to extract real changes in the intensity that correspond with state changes and is accomplished primarily by two methods: edge-preserving nonlinear windowed smoothing or reconstitution of the underlying signal through statistics.

A popular edge-preserving smoothing algorithm was developed by Chung and Kennedy in 1991 to analyze patch-clamp data[35]. In this method, the smoothed value of a point is determined by the moving averages the forward and reverse context. The contributions of the upstream and downstream averages, however, are weighted by the standard deviation of windows ahead or behind the point, thereby reducing the contribution of context regions that contain an intensity level transition. Alternatively, probability densities and statistical information in the data can be used to algorithmically recapitulate an idealized trace of the underlying signal[36,37]. This method, which includes the commonly used Hidden-Markov algorithm[38,39], uses robust statistical modeling that results in a completely smoothed trace, ideal for automation of later analysis steps. However, these methods are computationally more taxing than windowed smoothing, become less accurate when SNR decreases, and often require some knowledge of the underlying kinetic model or expected number of states[37,40].

1.4.3 FRET efficiency state analysis

Using the smoothed time-intensity traces, intensity levels and transitions between them can be identified. For FRET experiments, the smoothed donor and acceptor traces can be converted to FRET efficiencies using Eq. 1, where E is the FRET efficiency, I_A is the acceptor intensity, and I_D is the donor intensity.

Eq. 1.1. $$E = \frac{I_A}{I_A + I_D}$$

With a known R_0, E can be converted to the distance between fluorophores, r, by Eq. 2, which can then be used to infer conformational or relative location states. Mechanistic models can be established using the amount of time spent in each state[6].

Eq. 1.2. $$E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6}$$

1.4.4 Measuring stoichiometry with fluorescence microscopy

A useful option for stoichiometric measurement *in vivo* takes advantage of stochastic photobleaching of fluorescent molecules, or the irreversible loss of fluorescence through molecular rearrangement. Continual excitation results in a stepwise reduction in intensity as fluorophores bleach over the course of the experiment. Photobleaching of fluorescent foci has been used to evaluate stoichiometry of protein complexes *in vivo*[30,41,42]; however, it has traditionally been applied to prokaryotes or membrane proteins in eukaryotes, where TIRF can be utilized to increase the SNR. Imaging eukaryotic proteins within the nucleus presents a significant challenge because TIRF excitation cannot penetrate deeply enough and eukaryotic cells contain organelles, increasing potential sources of light scattering[43], decreasing the SNR. While it may be possible to reduce background when imaging within nuclei using reduced labeling efficiency, this would not be useful for quantifying stoichiometry as unlabeled populations would not be observed.

1.5 Present Work

The bulk of this thesis aims to improve upon methodology for fluorescent microscopy, both *in vitro* and *in vivo*. In Chapter 2, I describe a new approach to computational analysis of stochastic photobleaching events to determine stoichiometry *in vivo*. Chapter 3 is a description of a method of labeling oligonucleotides in a cost-efficient and sequence independent manner using a previously described bifunctional linker. Chapter 4 describes our efforts to overhaul the upper-level undergraduate biochemistry laboratory curriculum to bring graduate level research experience to undergraduate students. Finally, Appendix A contains a user guide to the software described in Chapter 2.

CHAPTER 2: CREATING AN ANALYSIS PIPELINE FOR DETERMINING STOICHIOMETRY IN LIVE YEAST THROUGH STOCHASTIC PHOTOBLEACHING

The data used for development of this software were collected by Dr. Jacquelyn Bower. The code for the smoothing and edge-detection algorithms was taken from software written by former lab members Dr. Jacob Gauer, Dr. Vanessa DeRocco, and Cherie Lanyi, with edits for computational efficiency and customized output for this specific application.

2.1 Introduction

Stoichiometry is fundamental to chemical reactions, but its measurement remains difficult in biological complexes, especially within living cells[44]. Fluorescence microscopy is an attractive option for stoichiometric quantification, as labeling methods can minimally perturb the target systems and advances in detection technology allow for observation on a single-molecule level[19]. Despite the current technological sophistication, hurdles to accurate quantification of stoichiometry remain in both imaging methods and data analysis. Living cells have inherently higher levels of background noise than *in vitro* imaging experiments. This effect is particularly pronounced in eukaryotic cells, which tend to be larger, have more complex internal cellular structure (serving as sources of light scatter)[43,45], and have higher levels of autofluorescence[46]. Imaging nuclear proteins in *Saccharomyces cerevisiae*, for example, during the majority of the cell cycle requires photons emitted by a fluorophore to traverse the nuclear membrane, cytosol (containing organelles), cellular membrane, and cell wall. Traversing media with differing refractive indices results in diffraction at each interface. For this project, we will be enzymatically digesting the cell wall of the yeast to create spheroplasts in an attempt to reduce the background.

2.1.1 Measuring stoichiometry in living cells

For standard microscope setups, there are two main approaches for quantifying stoichiometry using fluorescence: ratiometric comparison to an internal standard and counting stochastic photobleaching events. Ratiometric analysis involves measuring the intensity of fluorescent foci, or volumes of concentrated fluorophores, from both a target of unknown quantity and a fluorescently labeled standard[47,48]. Since the standard has a known quantity, comparison to the target can reveal the number of molecules composing the experimental focus. Alternatively, photobleaching of dye molecules can be used to calculate stoichiometry[47,48]. When provided sufficient excitation over time, fluorophores will undergo irreversible chemical rearrangement, resulting in the loss of fluorescence. When monitoring the fluorescence intensity of a dye population over time, these events will appear as a step-like decline, with the height of each step corresponding to the loss of fluorescence of a single fluorophore. The primary challenge of this method is identifying those steps when the signal-to-noise ratio (SNR) is low. The focus of this chapter will be using photobleaching step quantification to determine stoichiometry.

A fluorescent standard must be employed for ratiometric comparison, to establish the intensity of a single fluorophore, or to assess analytical reliability[41,47,48]. Several common standards exist, including the yeast centromeric protein Cse4p[47,49,50], the bacterial flagellar motor protein MotB[30,41], and a number of membrane-bound protein complexes[33,51,52]. To ensure reliability of stoichiometric analysis, it is important to corroborate molecule counts using some other biochemical method[53]. Our long-term goal is to quantify DNA mismatch repair foci in cells. Studies from our lab using atomic force microscopy have demonstrated that DNA mismatch repair foci can be comprised of recognition protein counts ranging from a single

protein to possibly 10 or more bound to a lesion site[54]. Whether complexes on the larger end of that spectrum form *in vivo* or are anomalies created by *in vitro* conditions remains to be seen. One potential confound for *in vivo* analysis is that smaller complexes may not produce detectable foci and would therefore be excluded; this would be indistinguishable from lack of formation of small complexes. Therefore, it is important to first establish the limit of detection for our system.

Fluorescently labeled *E. coli lac* repressor protein (LacI-eGFP) has been used as a chromosomal marker in yeast, where repeat arrays of its DNA binding sequence (LacO) were inserted into the genome. This is an intriguing standard because yeast contains no endogenous LacI, so all copies of the protein would theoretically be fused with eGFP; LacI binds tightly with the LacO site[55], creating ideal foci for analysis; and the number of LacO sites can be controlled through genetic manipulation, creating custom array sizes. Using this system, a stoichiometric "ruler" can be created within the nucleus using decreasing LacO repeat numbers; the limit of detection can then be determined by the minimum number of LacO sites necessary to produce detectable foci.

2.1.2 Extracting photobleaching steps from noisy time-intensity traces

There are currently several methods for distinguishing a photobleaching event in a time-intensity trace from background noise. A windowed-smoothing algorithm can be employed to reduce noise in plateau regions between event transitions but should contain an edge-preserving mechanism to prevent smoothing over transitions. The simplest approach is to use a windowed-median method, which preserves edges better than using the mean[56]; however, when applied to *in vivo* stoichiometry applications, the method developed by Chung and Kennedy in 1991 for patch-clamp analysis[35] was shown to create standard deviations (SD) ~30% closer to raw values

than windowed-median smoothing[57]. In brief, the Chung-Kennedy (CK) algorithm calculates the mean for regions before and after the data point and uses the SDs of those regions to weight the contribution of the respective mean to the smoothed value. Input parameters for CK smoothing (such as window size, edge width, and equation variables) can be adjusted to optimize the performance[7], but the drawback is that it requires user input and may include bias.

A second approach is to use probability modeling to determine idealized traces containing discrete intensity states. This approach often assumes a hidden Markov model exists behind the noise[38,58]. Briefly, a Markov model is defined by states, each with an "emission probability," and "transition probabilities" that describe the probability of transitioning from one state to the next[59]. This method is useful for applications where a number of expected states and the underlying kinetics of the system are understood but become less functional when these characteristics are unknown[60]. In situations where the SNR is low, such that state transitions are on the order of background noise, it becomes difficult to algorithmically resolve the number of intensity states.

A third option is to use intrinsic statistical characteristics of the data to determine state changes. Algorithms can employ a sliding Student's t-test[61] or χ^2 test[62] to determine state transitions, but require user-defined parameters and handle situational differences in data with mixed results. The χ^2 algorithm, for example, requires the number of steps as a parameter, which may be prohibitive depending on what is known about the target; the Student's t-test algorithm tends to return unreliable results when steps that occur rapidly[60]. To avoid user bias, Kalafut and Visscher developed an automated algorithm that uses no user-defined parameters[60], but only derived statistical characteristics of the data to determine non-uniform step sizes. While promising, this algorithm yielded error >25% below a SNR of 1.5.

In this approach, we first employ a CK smoothing algorithm with two sets of forwards and backwards windows, one each for the mean and SD[7]. Multiple window sizes allow for fine tuning of the algorithm. Once the trace is smoothed, we test for state transitions by two methods: we convolve the trace with a Gaussian derivative kernel[63] and use the SDs calculated by the CK filtering windows to determine when a point is between two plateau regions[7]. Once transition locations are defined, statistical analysis between the surrounding regions determines the degree of confidence; this procedure will be described in more detail later.

2.1.3 Analysis of photobleaching steps to determine stoichiometry in vivo

Currently, there are two approaches to finding stoichiometry from photobleaching steps. The first approach is to directly count the number of steps; this procedure is most applicable for low molecule counts[47], as noise and the probability of multiple coincidental bleaching events increases as the fluorophore count increases[33,36,48]. The alternative method is to determine the size of a single photobleaching step and calculate the number of molecules present at the beginning of the experiment using the total loss of fluorescence intensity[30,41,47]. In either method, the size of a single photobleaching step must be established (for direct counting, multiple bleaches must be distinguished from a single event).

The simplest approach is to smooth the time-intensity trace with an edge-preserving filter[35,47,48] and choose steps manually[30,64]. The obvious downside of this method is that the definition of what passes for a real step is determined by the user, which introduces bias and variability between users. Recent advances in automated step-finding algorithms have applied a more robust statistical approach which uses probability density analysis and other statistical modeling to determine an idealized trace of the underlying signal. This analysis becomes

significantly less accurate and computationally taxing in situations where the signal-to-noise ratio is low[31,36,38,51,62]. The physical limitations of *in vivo* experiments where TIRF imaging is not an option lead to high noise that is on the same order as the signal, which makes it exceedingly difficult to accurately extract real photobleaching steps through an automated process, particularly when that noise is stepwise in nature.

Our approach is similar to previous ones in that we look for plateaus between bleaching events, but rather than attempting to remove background steps through thresholding, as is standard for manual step identification, we collect all steps from background and bleaching events. We can then characterize background steps by collecting data from control sources, at which point we can isolate and remove the subset of steps within experimental data that belongs to the noise. This procedure allows us to employ statistical methods for background removal with high confidence, a hallmark of automated probability-based algorithms, while operating under noise levels that might otherwise be prohibitive.

Here, I describe the development of a computational pipeline that allows for bulk analysis of photobleaching data, starting from identifying fluorescent foci in microscope movies of live yeast or purified protein *in vitro*, through time-intensity trace generation, and finally to step identification and photobleaching event quantification. The philosophy behind this pipeline is to minimize the processing time from image analysis through final data analysis. This goal is achieved through limiting the amount of user interaction, reducing the amount of data carried over from each step, and maximizing computational efficiency. There are currently still steps that require manual input, such as the determination of filter/edge-detection parameters, and as a rule of thumb we must manually review trace analysis output to ensure the data look “real” before proceeding to final analysis steps.

However, future directions (described later) will seek to reduce this interaction to a minimum. Secondary goals of the algorithm are to maximize memory-usage efficiency for use on local machines, optimizing file handling for use on the research computing cluster, and creating user-friendly interfaces for steps where input is required. All code for this pipeline was written in Matlab (Mathworks).

2.2 *In Vivo* Focus Identification and Time-Intensity Trace Extraction Pipeline

2.2.1 Yeast imaging

To determine stoichiometry of a protein-mediated function *in vivo*, a yeast strain is generated with the target protein tagged with a florescent fusion-protein. Living cells are converted to spheroplasts by digesting the cell wall with Zymolyase (Zymo Research, Irvine, CA) in an attempt to reduce refraction. Spheroplasts are deposited on plasma-cleaned quartz slides, which promotes flattening of the cells on the surface, and imaged on a microscope with laser illumination of a wavelength appropriate for the absorption spectrum of the fusion protein. Images are recorded until complete photobleaching is apparent and saved as a TIFF image stack.

2.2.2 Cell identification and numbering

The first illuminated frame of the movie (i.e. the first image captured after the laser is turned on) is automatically calculated by finding the maximum mean frame intensity within the first 50 frames of the movie, which consistently chooses to the correct frame when corroborated

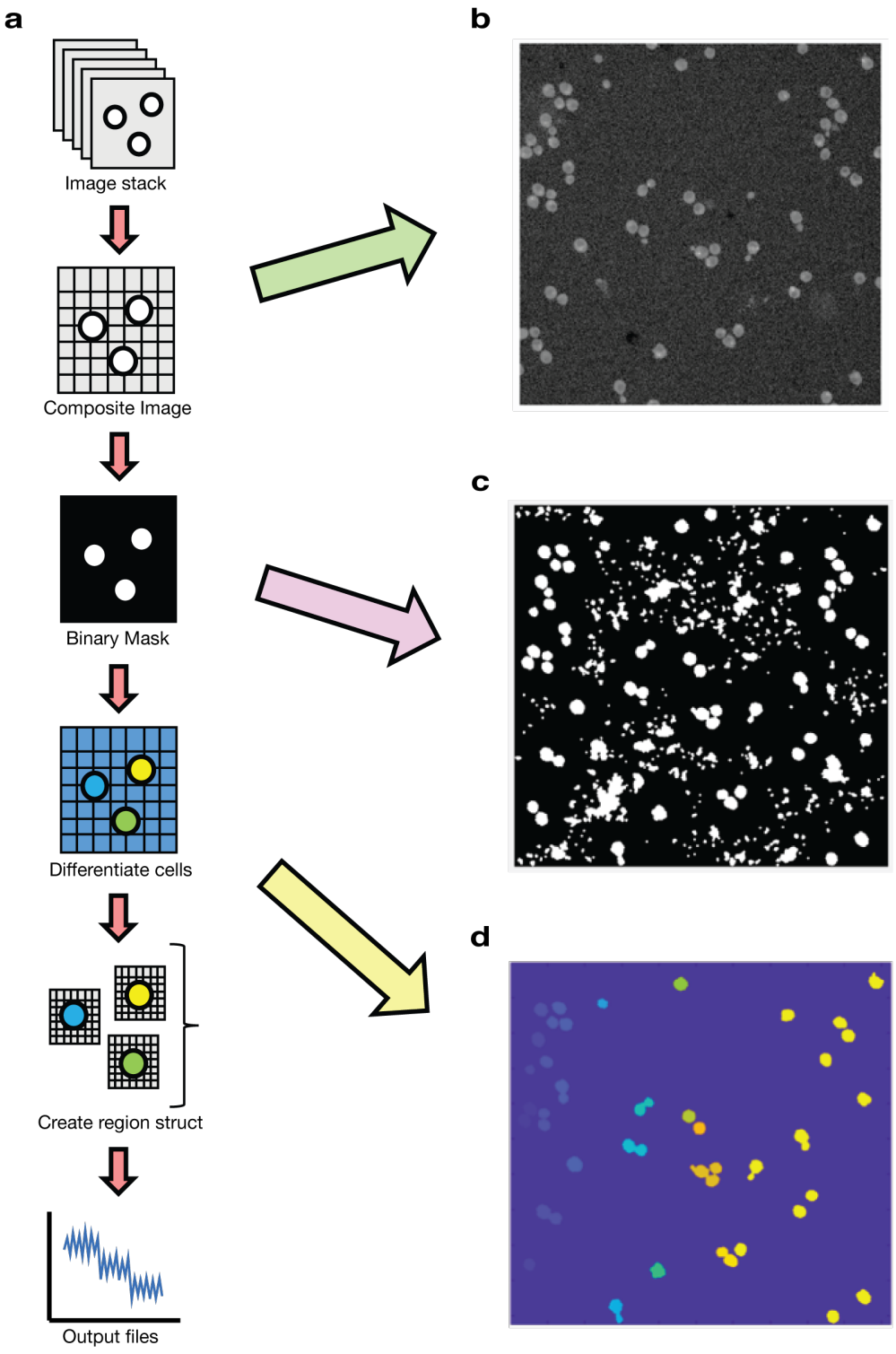

Figure 2.1. Cell finding algorithm overview. (a) A composite image is created from the sum of the first ten frames and background corrected. A binary mask is created over the identified cells, which are then numbered sequentially. All pertinent data is extracted for each cell and time-intensity trace files are output. (b) Example composite image. (c) Example binary mask, prior to background removal. (d) Example output of numbered cells.

through manual frame-by-frame inspection. Composite images for pre- and post-photobleaching are created by summing the first 10 illuminated frames and final 10 frames of the movie, respectively. The post-bleaching image is subtracted from the initial composite to increase contrast between cells and slide background (Fig. 2.1b).

A watershed algorithm set for a gradient background is applied to create a binary mask (Fig. 2.1c), which is then eroded and dilated with a "disk" structural element to remove small regions identified in the background but not associated with cells. The mask is then queried for connected regions using an 8-directional filter. The standard deviation of the pixel intensities within each connected component is calculated and any regions with a standard deviation below an arbitrary threshold of 1.5x the standard deviation of the entire image are discarded. This threshold was determined through trial and error and consistently removes regions manually identified as background without removing visible cells, which have much higher standard deviations independent of the presence of fluorescent foci. Any regions on the border of the image are also discarded. The remaining regions are cells, which are numbered sequentially, and a map is generated where the pixel values within each cell are equal to the number generated for that particular cell (Fig. 2.1d).

2.2.3 Fluorescent focus detection and data architecture

Each cell found in the composite image is subjected to thresholding to determine whether a fluorescent focus is present. The threshold is set by default to 3 standard deviations above the median of the cell, although this can be adjusted as necessary. If any pixels are found to have intensity values above threshold, the cell undergoes further processing; any cells without pixels above threshold are removed from the dataset, reducing the computational time and power for

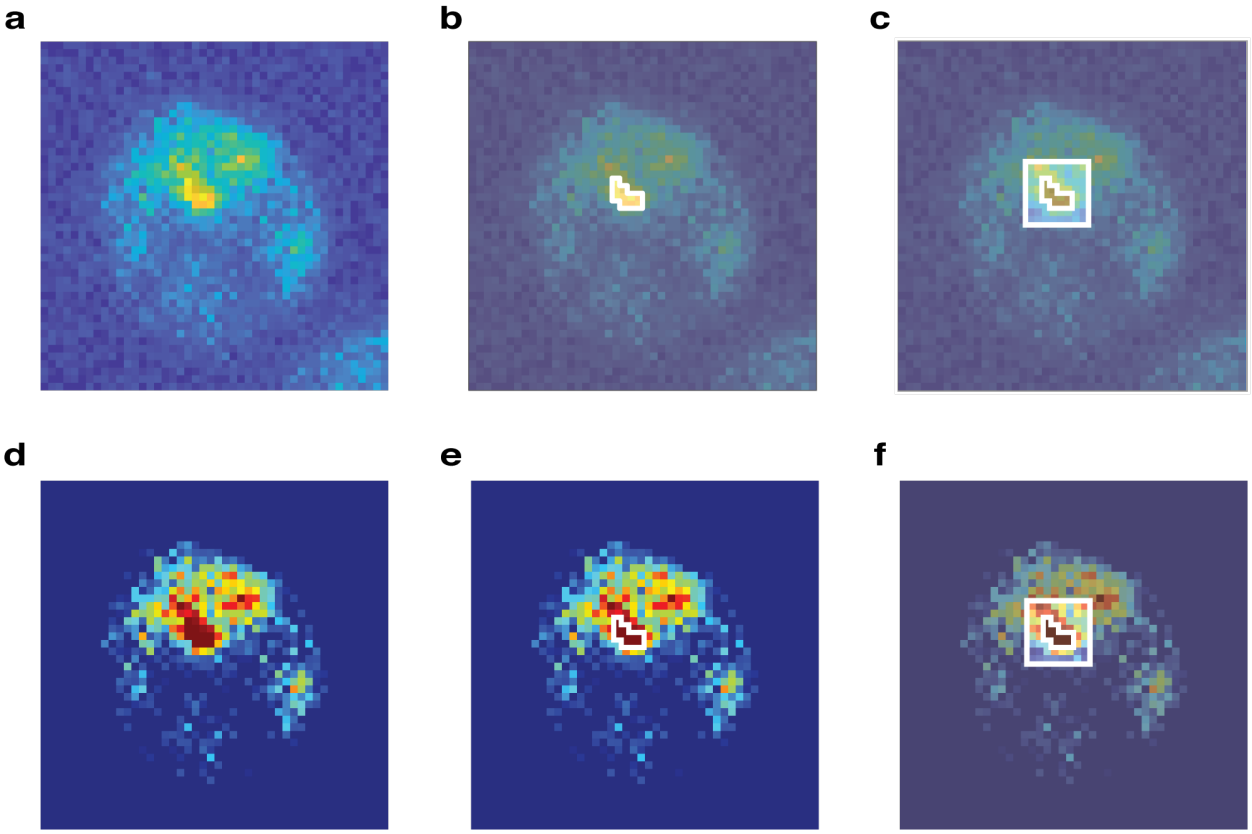

Figure 2.2. Focus and background pixel selection in an example cell. (a) Fluorescence intensity map of a single cell with a defined focus from the composite image. (b) Focus pixels found above threshold are highlighted. (c) A two-pixel border around the spot, including any pixels within the bounds not detected as a focus, is taken for backgrounding (highlighted). (d) Threshold mapping for the same cell is shown, scaled from 0-3 standard deviations above the cell median. (e) The focus pixels chosen highlighted on the threshold map. (f) The background pixels highlighted on the threshold map.

downstream analysis. A new binary mask is created for pixels above threshold in each cell with a detected focus (Fig. 2.2b,e) and the mask is tested for connected components to determine the number of foci within the cell. Each focus is then determined to have a minimum number of pixels (set by the user) or it is removed from further processing.

The user has the option to process foci by two methods. The default method is to find the maximum intensity value within the focus and draw a 3-by-3-pixel box around the point of maximum intensity (the "box method"; Fig. 2.3a, center); alternatively, the entire list of pixels found above threshold is used (Fig. 2.3a, right). In both methods, the algorithm ranks the pixels by intensity value and takes the top number of pixels set by the user, with a default value of 4 pixels. A rectangular box around each focus, extending two pixels in all directions beyond the bounds of the focus, is taken for frame-by-frame background correction of each individual spot, excluding pixels contained by any foci or outside of the cell (Fig. 2.2c,f).

A "struct" object is created for the movie, where each index position corresponds to the cell number found in the cell map. Those indices for cells without pixels above threshold or a focus of sufficient pixel size remain empty to reduce the memory footprint while maintaining the integrity of the numbering system from the cell map. Each index in the struct contains arrays of cell pixel indices and binary indications of whether each pixel is part of the background calculation or a focus. The binary designations are structured as "cell" arrays, where each pixel contains an array where the column number is the number of the focus (if multiple spots are discovered in a single cell). Each pixel's binary array can then be converted to an integer, which allows them to be saved as a foci map and parsed during later processing.

Only retaining data from those cells that have fluorescent foci in them provides a significant computational advantage. Early iterations of the pipeline processed the entire image

for each frame of the movie; this procedure was feasible for smaller movies that were used in initial benchmarking (512 x 512 images, 1000 frames, ~500 MB file size), but when used to analyze confocal images (1024 x 1024 images, 2000 frames, ~4 GB file size) the program would crash the Matlab environment. Taking advantage of vectorization and indexing only those pixels in cells containing foci results in a dramatic increase in computational efficiency; the initial program could process a 500 MB movie in approximately 3 minutes, while the updated script can process a 4 GB movie in 30 seconds.

2.2.4 Generation of time-intensity trace information

Photobleaching analysis requires the fluorescence intensity of a focus to be monitored throughout the course of a movie; the plot of intensity against time is referred to as a time-intensity trace. As the algorithm generates a trace, the movie frames are iterated through and all calculations are done frame-by-frame, adding the results to the respective struct index corresponding to each cell. Intensity values for all pixels within each cell with at least one focus are saved for reproduction of movies cropped around the cell, while excluding extraneous data from regions of bare slide and cells that lack a focus (reducing computational time and power). For each focus, the pixel intensities for the selected number of focus pixels and the surrounding background pixels are recorded. The median background intensity is calculated and subtracted from each focus pixel, after which the corrected focus intensities are summed. This process results in two matrices within the struct, one each for the corrected foci and their respective background median, where each row corresponds to a separate focus and each column is a frame. As a result, each row in the corrected foci matrix is a time-intensity trace for that particular focus.

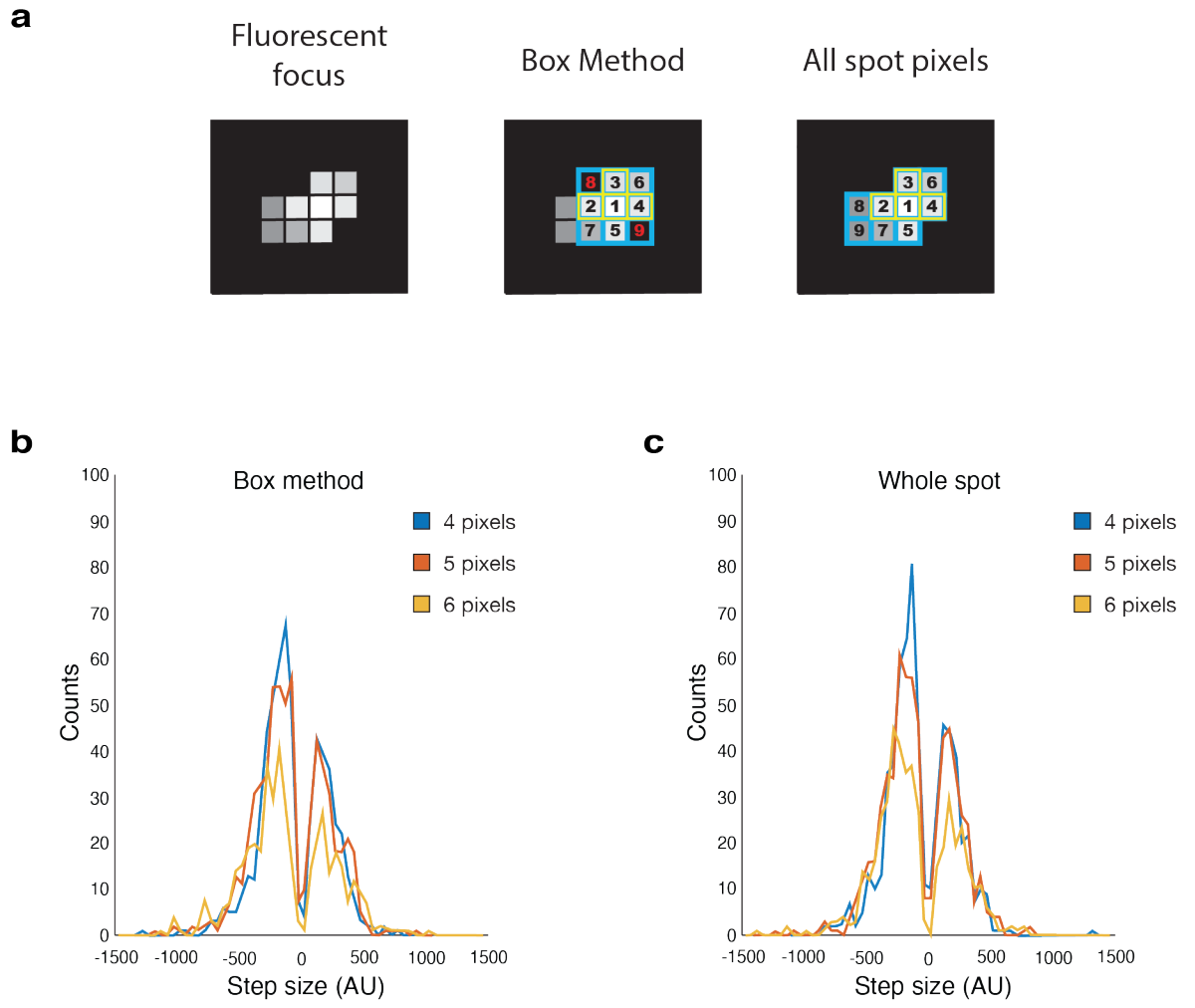

Figure 2.3. Pixel selection methods. (a) A focus contains pixels above the threshold (left; lighter pixels signifying higher intensity). Pixels can be selected using two methods. In the "box method" (center), a 3x3 box (blue outline) is drawn around the pixel with maximum intensity and the included pixels are ranked in descending order of intensity (shown by pixel numbers). Alternatively, all pixels above threshold can be ranked (right, blue outline). In this example, the top four pixels are selected for analysis (yellow highlight). In this case the box contains two pixels not included in the focus (black with red numbers), and these would be excluded outright. (b) Step size distribution comparison of the box method using 4-6 pixels for analysis. (c) Step size distribution when taking the top overall 4-6 pixels from the entire spot. The box method and top overall pixel distributions look identical to each other and the number of pixels taken does not seem to significantly effect step size.

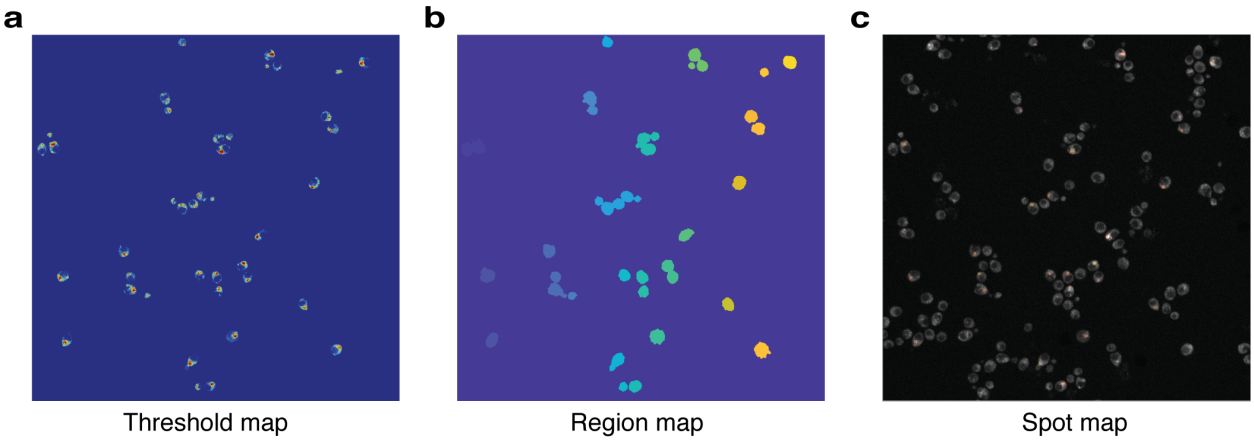

Figure 2.4. Output image maps. All image maps are output in both PNG and binary format. (a) The thresholds are scaled from 0-3 standard deviations above the median of each respective cell. (b) The region map uses the default colormap for the cell numbers. (c). The spot map is the contrast-adjusted composite image with the chosen background boxes overlaid in color. This allows the spot pixels to not be obscured. Overlapping regions of background are displayed as different color than either of the individual regions.

2.2.5 Output files

Image maps are generated for spots found, cell numbering, and threshold values. Each map is output in both PNG and binary format (with the ".erie" file extension), for user viewing and use during the transition analysis step later in the pipeline, respectively. The cell number map is a direct output of the region map, using the default "parula" colormap for the PNG file (Fig. 2.4b). To identify the chosen spots, the integer-converted background pixel map is superimposed on the composite image, using an arbitrary colormap for contrast, since any background overlap between foci within a cell will have a different value than either of the individual background designations (Fig. 2.4c). Finally, the threshold map is a heatmap scaled from 0 to 3 standard deviations above the median of each respective cell, with all values below

the median are set to 0 (Fig. 2.4a). This map allows the user to easily see how large the spot is and how high the surrounding pixels are relative to the spot.

The Chung-Kennedy and Gaussian kernel smoothing and edge-detection algorithms were developed during prior single-molecule Förster-resonance energy transmission (FRET) work in our lab[7,65]. To avoid writing redundant code, time-intensity traces are saved in the same format as in the FRET analysis pipeline, modified to exclude the differentiation of "donor" and "acceptor" data. The binary output file is a 16-bit integer matrix, where the first row consists of frame numbers and each subsequent row is the corrected frame-by-frame focus intensity for each spot identified. This format allows the data to be fed into a modified version of the batch trace analysis portion of the FRET pipeline.

2.3 Image Background Selection Application

The initial analysis of yeast images is automated and based on the detection of fluorescent foci in each cell. To correct for any systemic background interference from the camera, laser, or cellular autofluorescence, data must be collected from locations not including a fluorescent focus. To this end, I created an application that performs the same analysis as the cell image analysis pipeline in a standalone environment. In lieu of region identification and thresholding, pixels are taken using the coordinates of user mouse clicks on the composite image. If the coordinate is close to the interface between a cell and the slide background, the algorithm determines which region comprises the most pixels within the nine-pixel box around the click coordinate (using the cell number map, where the slide background has a value of 0) and only takes spot and background pixels from that associated region, so as not to skew the background correction. The trace output from this application is the same format as those generated from the

automated script, but there are no associated map outputs. The filenames automatically reflect that these traces are background. This procedure not only allows us to determine if background noise generates "steps" in our analysis, but also allows us to determine the source of the noise (i.e. camera noise or cellular autofluorescence).

2.4 Adapted Script for Analysis of *In Vitro* Control Imaging

While the generation of time-intensity traces and localized backgrounding remains the same for *in vitro* experiments, the identification of foci to be analyzed from *in vivo* images is dependent on first identifying the cells. Thus, the same script cannot be used to identify fluorophores deposited directly onto a slide. Furthermore, the software currently used by the lab[7] to analyze FRET is able to handle neither the TIFF format of the images captured by the imaging suite used by the confocal microscope setup nor the resolution and single-channel nature of the acquired images.

To accommodate these images, the initial methods of the analysis script were modified to locate molecules in a fashion similar to the FRET analysis software[32,66]. A composite image is created as previously described, at which point the image is block-processed using 16-pixel-by-16-pixel regions. The median and standard deviation of each block is calculated, creating a new matrix for each with dimensions of $m/16$ x $n/16$, where m and n are the width and height of the image, respectively. These dimensions are then expanded to the original image dimensions using a bicubic extrapolation. For each pixel coordinate in the composite image, a threshold is created using the corresponding values from the standard deviation and median matrices; the default threshold is 5x the standard deviation over the median, but the multiplier is user-adjustable. Only those pixels found above this threshold that are also the maximum value in the surrounding 3-by-

3-pixel window are identified for further processing. To reduce contribution of intensity from neighboring molecules, identified maxima within 5 pixels of another are eliminated.

Following identification of maximum spot pixels, the script continues in the same manner as the background identification application. The 3-by-3-pixel box around each local maximum in the composite image are sorted by intensity and the coordinates of the top intensity pixels are converted to linear indexing and recorded into a structural element with the same format as the *in vivo* processing script. The number of pixels taken is adjustable and should be the same as the number taken for the other analysis methods to maintain background fidelity; by default, 4 pixels are recorded. Background pixels are taken from a 2-pixel border around the box and conversion to time-intensity traces occurs using the same algorithm as the *in vivo* pipeline.

2.5 Filter Sliders Application

In contrast to TIRF imaging, time-intensity traces generated from *in vivo* images have an inherently low signal-to-noise ratio. Parameters for the Chung-Kennedy smoothing algorithm and both the Chung-Kennedy and Gaussian kernel edge-detection methods must therefore be tuned to ensure the accuracy of detected step intensity levels and edges. The current FRET pipeline does not include a means of real-time testing of these parameters; users are required to enter parameters, run the batch trace analysis script on some or all of the traces, review the results using a separate application, and then refine and rerun until desired results are achieved.

There are two major drawbacks beyond the recognizable inefficiency of this method. First, the effects of changing each parameter are not entirely obvious, leading to a mostly blind guess-and-check characteristic to the process. Second, there is a tendency of the user to refine parameters using the best-looking trace from the dataset, one that may not be representative of

the set as a whole. To improve ease and accuracy, I created a Matlab application that allows the user to adjust the filter parameters and monitor the effects in real-time. The user interface allows for display of up to three traces, selected manually by the user or at random. Each parameter has an associated value slider that, when changed, updates the displayed traces to reflect all changes in smoothing and detected edges. To add to the robustness of the statistical analysis, I included a Wilcoxon Rank-Sum in addition to the Student's t-test already being performed on the regions around each transition. The p-values used to accept a transition are also included in the adjustable sliders.

To add additional confidence, I implemented a scoring system for the detected transitions. One point is allotted for each of the following criteria, up to a maximum of six points. Transitions receive one point each for detection by the Chung-Kennedy algorithm and through Gaussian kernel convolution. Any overlap of the 95% confidence intervals for regions flanking a transition are determined using both smoothed and raw data, each worth a point in the event there is no overlap. Finally, a point is awarded if the smoothed surrounding regions pass each of the Student's t or Wilcoxon Rank-Sum checks. The p-value of the statistical tests required for positive scoring can be adjusted separately of those included in the edge-detection, so that less-confident edges can be included in the detection method and filtered later. This capability is particularly useful when the data are noisy and/or the changes in intensity are small. In addition, I included a minimum score filter to remove those edges detected with lower confidence. It is important to note that edges with scores lower than this cutoff may still appear; when edges are removed, the edge-detection algorithm then creates average regions between the remaining edges. This procedure may change the scoring of those edges, because the 95% confidence intervals and statistical tests will be affected. If the score threshold is applied again, it

could further remove transitions that previously were above threshold, again potentially altering the scores of the remaining transitions. This iteration could conceivably repeat *ad absurdum* until no transitions remain; therefore, the score cutoff is only applied once when the transitions are initially detected. Each smoothed trace can be saved individually in the same format as the output of the batch analysis software if the user wishes to view the results in downstream applications. Alternatively, the user can save the current values of all parameters into a configuration text file that is automatically read by the modified batch analysis algorithm.

2.6 Step Analysis Application

A key step to determining stoichiometry is to establish the loss in fluorescence intensity associated with the photobleaching of a single fluorophore. If one can confidently determine this value, then the number of fluorophores initially present can be calculated from the total intensity of the spot. Alternatively, the number of photobleaching steps can be counted directly, but the intensity loss from the bleaching of a single fluorophore is required for differentiating events where multiple bleaches occur simultaneously.

2.6.1 Step size determination

The smoothing and transition-finding algorithms provide edge locations with mean values for the regions between edges (Fig. 2.5). When these mean values are plotted, the transitions between different intensity levels become clearly visible. We can calculate the difference between the mean intensity levels around each transition to define a step size; this method is, however, inherently noisy because there is error in the calculation of each regional mean. The uncertainty in the value of the regional means results in steps of different magnitude

when the trace contains transitions of fluorescence recovery (Fig. 2.5a). Alternatively, we take an approach that draws a series of horizontal lines to define "step levels." Beginning with the highest mean intensity region, the program checks for any regions whose standard error of the mean (SEM) overlaps with the SEM of this region (Fig. 2.5b). Any overlapping states are statistically indistinguishable and are therefore considered to be part of the same step level. A weighted mean for the step level is calculated, weighted by the SEM and number of points of each constituent region (Fig. 2.5c). We then move to the next highest region that is not already included in a defined step level and repeat this process. After iterating through all regions in this manner, we finally adjust any regions that overlap with multiple steps by determining the step level to which the mean of the region is closest, recalculating the weighted means of the steps as necessary. This process is repeated three times to ensure these multi-overlap regions are classified to the appropriate step level.

2.6.2 Step size data analysis

Now that steps are defined by changes from one step level to another, rather than the raw differences between regional means, step sizes can be binned for analysis. Typical time-intensity traces from *in vivo* images contain steps both increasing and decreasing in fluorescence intensity due to noise or fluorophore blinking, resulting in two distributions centered around 0 intensity. For real fluorescent foci, the negative (down) step population is larger than that of the positive (up) steps. The up-steps represent a combination of blinks, rebinding events, and background noise; the down-steps are comprised of blinks, dissociation events, background noise, and photobleaching events.

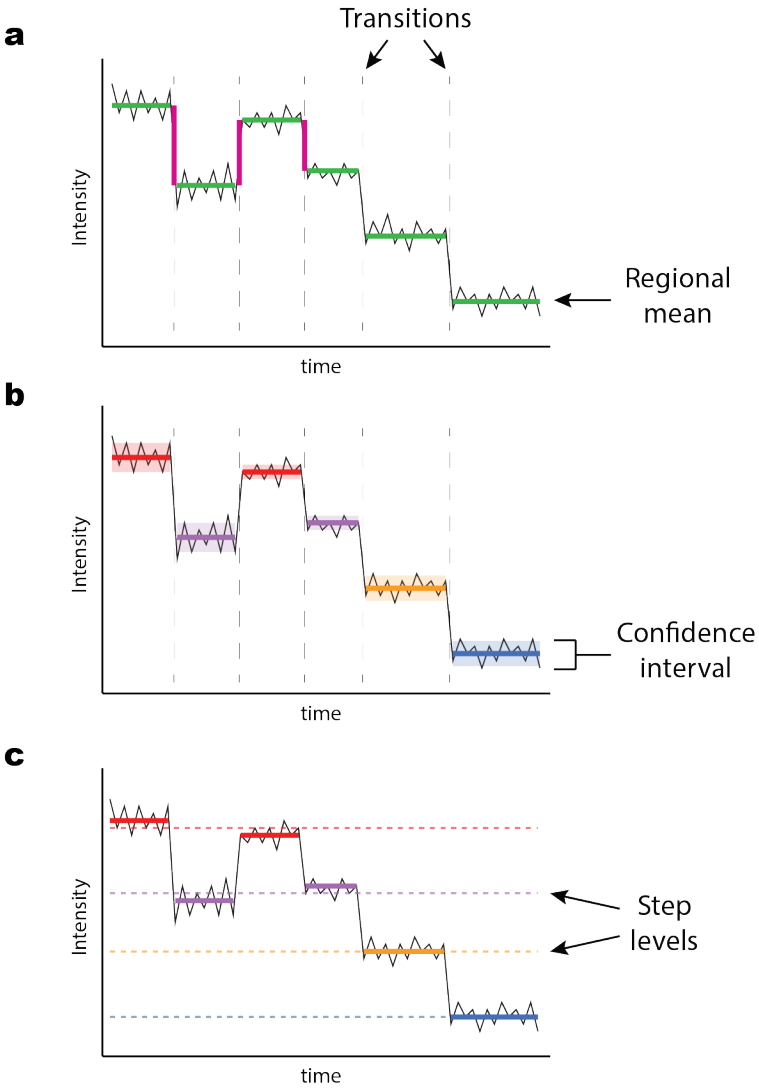

Figure 2.5. Conversion of regional means to "step levels." (a) Using the smoothed time-intensity trace (solid black line), means (green lines) are calculated for the regions between transitions (vertical dashed lines). Using the difference between consecutive regional means as a step size would lead to three different magnitudes for the first three transitions (highlighted in pink) in this example trace. (b) A 95% confidence interval for each region is calculated around the mean. Regions with overlapping confidence intervals can be grouped together (shown by matching colors). (c) Weighted means are calculated for each region group (horizontal dashed lines). Each line represents a "step level;" a step is then defined as a transition from one step level to another and the weighted means are used to calculate the magnitude of the step.

We can remove known blinks from the population by finding any down-step followed immediately by an up-step of the same magnitude. This logic may theoretically include quick dissociation and rebinding, but both are functionally the same for our analysis as we are looking to count molecules through photobleaching and not glean binding information. Dissociation would appear identical to bleaching events (in our particular case, it would appear as two simultaneous bleaches since the LacI-eGFP binds as a dimer), but any subsequent rebinding would remove these events from the net reduction in intensity over the course of the movie. Background noise should create equivalent populations of up- and down-steps.

Based on these assumptions, we can identify photobleaching step populations by summing the positive and negative populations, creating a corrected population of net down-steps. We can fit the plot of histogram bin counts as a function of bin center value to a combined series of Gaussian curves or binomial distributions. After removing blinks and background population curves, the up-steps should contain only rebinding events; the down-steps should contain a prominent peak at the size of a single photobleach, with sequential peaks of decreasing height at multiples of the single step size. Importantly, the size of a single photobleaching step above the background is found using an entire dataset for increased statistical power. The calculated step size can then be applied to either of the two counting methods. The individual photobleaching steps can be counted directly, discerning simultaneous bleaches from single events, or the total intensity drop over the course of a movie can be divided by the size of a single step to calculate the number of molecules initially present.

2.7 Benchmarking with Initial Data

Testing of the analysis pipeline was performed on a test dataset from a LacO4 strain (containing four LacO sequence repeats). At full occupancy, foci in this strain will be comprised of eight LacI-eGFP molecules. Notably in this case, binding/unbinding events will be comprised of two molecules, assuming both fluorophores are in the "on" state.

2.7.1 Determining optimal pixel count for time-intensity trace generation

The optimum number of pixels to use in time-intensity trace generation needs to be determined before downstream analysis can take place. Inclusion of a pixel results in the collection of the background noise intrinsic to that pixel. Pixels that are around the edge of a focus are likely to contain more diffused signal from the focus fluorophores than those pixels at the center. The result is a lower SNR in pixels around the periphery of the focus; excluding these noisier pixels from analysis should increase the SNR of the included data.

Using both the box method and ranking all focus pixels, an inclusion range of 1 to 9 pixels (the maximum available when using the box method) were compared to determine the optimal number; previous TIRF analysis optimization from the Weninger lab suggested that after 4 pixels the returns were diminished (personal communication). Higher pixel counts (7-9) yielded greater noise, especially for the box method; when a focus is asymmetric and the pixel of maximum intensity is near the edge, pixels outside of the focus would fall within the box and included (Fig. 2.3a, center). Using data from 1 and 2 pixels resulted in inability to detect loss of fluorescence due to photobleaching from stepwise background noise. The dataset was ultimately tested for 4- to 6-pixel inclusion and propagated through the completion of the step analysis pipeline. The step-size data appear to have identical distributions for equivalent pixel counts

using both methods (Fig. 2.3b,c). Within both methods, the increase in pixel count appears to affect the data only in that it shifts the calculated step sizes to larger intensities; this effect appears to be true for both the identified background and apparent photobleaching populations.

The reduction in cumulative step counts with increasing pixel count (Fig. 2.3b,c) is due to a reduction in the number of foci analyzed when using higher pixel counts. To illustrate why foci with fewer pixels above threshold than the number of pixels taken for analysis must be rejected, we can consider a hypothetical focus containing 4 pixels above threshold. A 3x3 box including the 4-pixel focus contains 5 non-focus pixels; taking the 5 pixels with highest intensity, for example, would include one non-focus pixel. Using the alternative method of taking the most intense pixels from the entire focus, taking 5 pixels from a 4-pixel focus is impossible, so including a focus with only 4 pixels in a dataset of 5-pixel foci would not be appropriate as it contains intensity information from a different number of pixels.

2.7.2 Background steps are distinguishable from "real" steps

Analysis of traces collected through the background selection application indicate that background steps appear smaller than "real" steps and fit to a single Gaussian for each population (positive and negative), with small standard deviations and peak locations for up- and down-steps that have absolute values within error of each other (Fig. 2.6a). Interestingly, these background steps appear in spots selected from both regions of empty slide (Fig. 2.6a, blue, n = 20,464) and in cells where foci don't exist (Fig. 2.6a, red, n = 18,262), suggesting that the noise results from the microscope/camera system, as opposed to cellular autofluorescence. This result is consistently reproducible across imaging experiments and collection days, giving us high confidence that the noise is systemic. When comparing the background distributions to

intensities from cellular foci across all strains, prominent peaks are present in both the positive and negative distributions at higher intensity magnitudes than those seen in background traces (Fig. 2.6a, green, n = 3,140). Shoulders on the major peaks (both positive and negative) appear to be of similar size to the background steps, suggesting we can distinguish between "real" and background steps within the foci data.

2.7.3 A single photobleaching step size can be calculated

The step size data were fit to multiple Gaussian distributions to determine the size of a single photobleaching step. The distributions fit to two and three Gaussians for the positive and negative intensity changes, respectively. Populations of down- and up-steps with centers close to those measured as background were found (Fig. 2.6b, purple, -101 AU and Fig. 2.6b, red, 127 AU for LacO4, compared to approximately +/- 100 AU in background traces) and removed. The remaining population is comprised of three distributions. Two of these distributions approximately mirror each other in the negative and positive domains, centered at -222 AU (Fig. 2.6b, green) and 250 AU (Fig. 2.6b, orange), respectively. The final peak is found at -450 AU (Fig. 2.6b, cyan), which is close to double the intensity loss of the other negative peak. From these observations, it is likely that the two peaks of smaller intensity (-222 AU and 250 AU) correspond to intensity changes associated with single fluorophores. The positive distribution would be comprised of fluorescence recovery following blinks, which explains why it contains fewer events; the negative peak would contain both blinks and bleaching events. A negative peak at twice the step-size corresponds to multiple bleaches or dissociation of a single LacI-eGFP dimer. It is reasonable that this peak has a lower amplitude as these events would occur less frequently.

With a single-molecule step size value, we can iterate back through the individual foci to determine the number of molecules in each focus using Eq. 3, where *N* is the total number of bins.

Eq. 3. $$number\ of\ steps = \frac{\sum_{i=1}^{N}(bin\ center_i * bin\ counts_i)}{single\ step\ size}$$

Specifically, we multiply the counts of each bin by the central value of the bin and sum all the products. This process removes up-steps – consisting of blinks, binding/dissociation events, and background steps – leaving only the contribution of photobleaching steps. This net total can be divided by the calculated single-molecule photobleaching step size, giving the total number of molecules present at the start of the movie in a particular focus. Alternatively, we can apply the same analysis used for a single focus to the entire dataset, creating a net loss of fluorescence across the entire set of foci. Dividing this net loss by both the size of a single photobleaching step and the total number of foci would provide an average number of molecules for all foci.

The yeast strain used for preliminary analysis contains 4 LacO sites; each LacO site can accommodate a single LacI-eGFP dimer containing two fluorophores for a total of 8 fluorophores (assuming all are in an "on" state). Using the focus-wise calculation method, the distribution of the number of molecules in the foci of this movie displays a major peak at approximately 6 molecules, which represents 75% occupancy by molecules in the "on" state. Peaks at 2 molecules (25% occupancy) and 8 molecules (100% occupancy) are also present, but to a lesser extent (Fig 2.4e). Observing across all foci gives an average of 6.2 molecules, which is close to the main peak in the per-trace analysis. The observed distribution contains reasonable molecule counts (at 100% occupancy and lower) and is an interesting preliminary result in

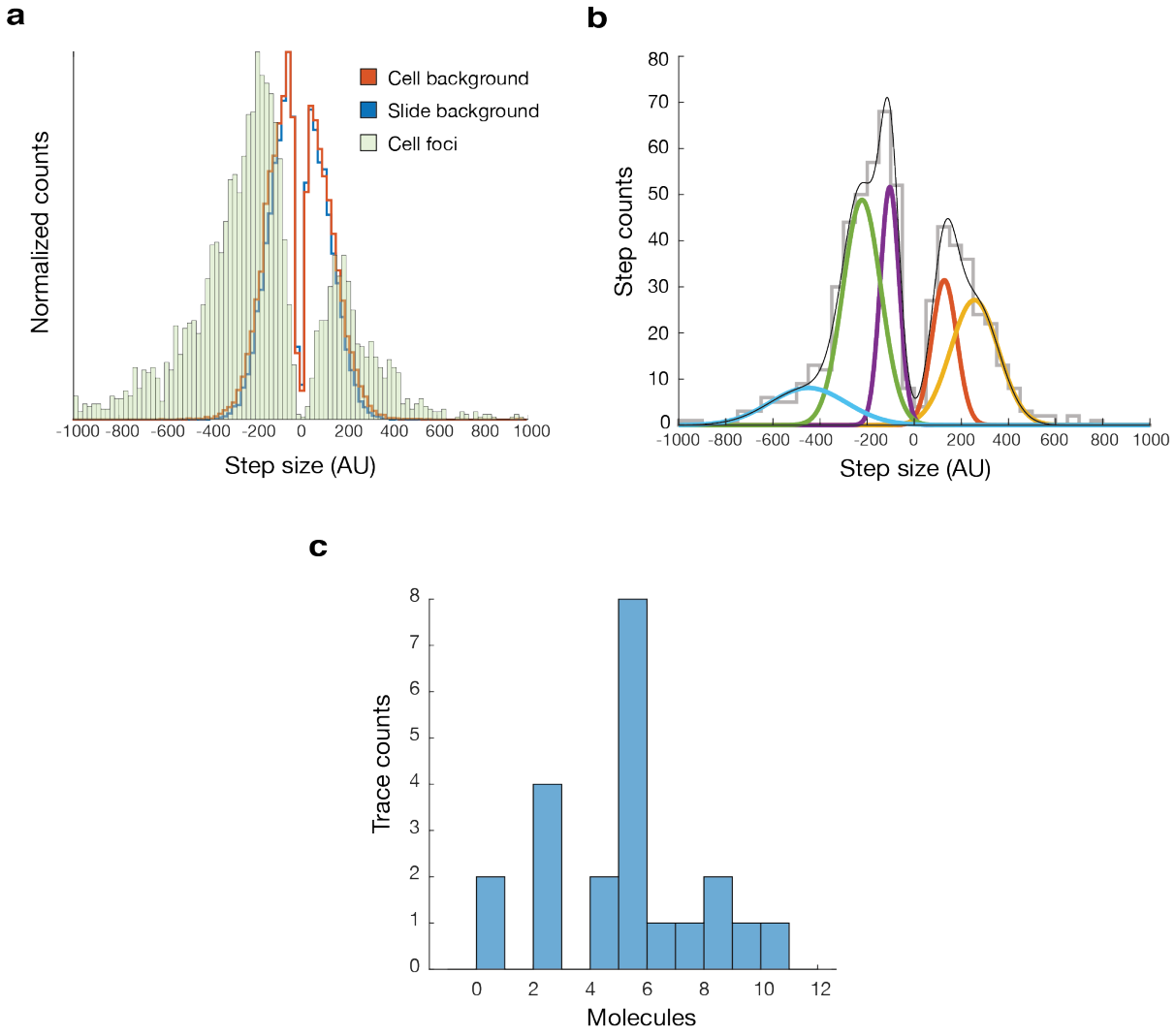

Figure 2.6. Step-size analysis methods using initial data. (a) Backgrounds for slide (blue, n = 20,464) and cells (red, n = 18,262) compared to the total population of spot steps (green, n = 3,140). Background distributions are smaller in step size than the major peaks from the spot steps. (b) Analysis of a LacO4 movie. Raw step sizes are represented by a gray staircase plot, with constituent distributions for best fit analysis plotted in color (R^2 = 0.9921). (c) Calculated molecules per LacO4 focus trace using the step size from (b) (green, -222 AU).

respect to the use of relative intensity, because there does not appear the be full occupancy, possibly due to unoccupied binding sites or to the presence of molecules in an "off" state.

Some confounds are present in this dataset. There are some foci that have significantly higher pixel counts than others, but exhibit step-wise decay indicative of molecular photobleaching with step sizes equivalent to those exhibited by loss of eGFP fluorescence. Similar foci appear in all strains expressing LacI-eGFP, regardless of the presence of LacO binding sites, but are absent in strains without the labeled protein. These observations suggest that these are aggregates of LacI-eGFP, which would account for the presence of photobleaching steps of similar (but slightly larger) size, as well as the increased molecule count. It may be necessary to tune the expression levels to eliminate these aggregates physically or to provide a means of removal through the analysis pipeline.

2.8 Conclusion and Future Directions

The methodology used in this software package shows that what appear to be real photobleaching steps can be separated from background noise in live yeast nuclei. While the data used to benchmark each portion of the pipeline were limited, the preliminary results suggest that we can identify fluorescent foci containing less than eight molecules using the automated focus-finding algorithm. We can process a substantial amount of data quickly, although some steps in the process are still labor-intensive.

Further data analysis is currently in progress using the range of strains from LacO1 – LacO16, at which point we can start to see the limitations of the application of stepwise photobleaching analysis. We would like to determine the minimum number of molecules present in a focus necessary for detection as well as accurate step-size determination. From these data,

we can begin to make computational refinements to optimize data capture and step analysis. We will also be able to observe the calculated occupancies of the different strains and relative abundance of fluorescent aggregates, giving insight into expression tuning for these fluorescent standards as well as potentially unknown targets. These occupancies will need to be corroborated through chromatin immunoprecipitation and TIRF imaging, allowing us to get snapshots of the occupancy in the low-background *in vitro* environment where photobleaching is easily quantifiable. Finally, we can begin to use this same analysis for the myriad other strains containing LacI fused with other fluorescent proteins, providing standards for a wide variety of applications, including colocalization and stoichiometry for multiple labeled participants.

In the future, I would like to optimize this pipeline in several ways. I would like to adapt a machine-learning component to computationally determine ideal parameters for the Chung-Kennedy smoothing and the edge-detection methods. Such implementation would eliminate one point of user interaction and prevent missed transitions caused by user bias, which in turn would create more accurate mean regions for step-level determination. The background selection application can also be incorporated into the automated cell selection script; automating background selection would drastically reduce the amount of time needed for collection of these data, not only by eliminating the manual selection of spots, but because the stand-alone Matlab application environment runs much less efficiently. Further, scripting background selection would allow the entire pipeline to be run on the Research Computing cluster and the time-saving benefits would be compounded.

I have created a software suite that can perform photobleaching analysis of fluorescent foci in live cells, from identifying cells, through extraction of pixel intensity information, and ending with quantification of photobleaching steps. While the pipeline requires some user input,

user-interfaced programs have been designed to reduce processing time and increase ease of use. Most importantly, these interfaces allow users real-time feedback to tune statistical parameters, which serves to eliminate the user-bias concerns associated with choosing photobleaching events "by eye." This approach applies similar statistical methods that many automated algorithms employ but can be applied to low-SNR data. Using the pipeline, we can distinguish systematic step-like noise derived from the microscope/camera system from "real" photobleaching steps; this distinction may not be possible using other automated step-detection analysis platforms. Finally, the programs have been optimized to minimize memory usage and computational time, allowing users to process large datasets that would normally take weeks in a few days.

CHAPTER 3: USING PHOSPHOROTHIOATE DNA FOR CHEAP, SEQUENCE-INDEPENDENT INTERNAL FLUORESCENT LABELING

3.1 Introduction

Protein interaction with DNA is vital to cellular function and genome stability. Conformational change in DNA structures often accompanies protein-mediated processes like gene regulation[67] and DNA damage repair[1,7]. DNA oligonucleotides modified with fluorescent dyes are useful for techniques measuring DNA bending and protein-DNA interactions, such as Förster resonance energy transmission (FRET)[6], bulk anisotropy[68], single-molecule protein-induced fluorescence enhancement (PIFE)[69], and electrophoretic analysis of nuclease activity[70].

Dye location is an important component of these methods, as FRET and PIFE are determined by inter-fluorophore and fluorophore-protein distance, respectively, and are sensitive to minute changes. Furthermore, dye attachment might perturb protein binding or cause unintended PIFE effects that might confound FRET analysis. Thus, multiple labeling sites may need to be tested for optimal results. For substrates where end-labeling is not an option, dyes must be attached internally. There are currently two commercially available methods for modification: backbone incorporation and functionalized thymine residues.

Commercial backbone labeling employs phosphoramadite chemistry, where chemical linkers mimicking a 5’ phosphate and 3’ hydroxyl allow an organic linker to be synthesized directly into the oligo. This method has the advantage of being sequence-independent and allows for interrogation of the full sequence space. Traditionally, backbone incorporation involves using cyanine dyes with the chemical moieties directly attached to opposite sides of the dye, the

drawback of which is that the dual-anchoring of the dye limits the rotational freedom of the dipole. Rigid dipole moments could yield inconsistent FRET efficiencies when using multiple dyes incorporated in this manner, as FRET efficiency is dependent upon the relative orientations of the donor and acceptor fluorophore dipoles[16]. Recently, a new phosphoramadite linker became commercially available that incorporates a single-linker into the backbone with a free amine group, which eliminates the dipole moment concerns. Alternatively, thymine residues can be included in the oligo which contain an attachment modification on the C_6 position of the nucleobase. These modifications can contain a number of attachment chemistries, thereby maximizing the compatible dye library, and the singular nature of the attachment does not limit the orientation of the dipole moment. This method, however, limits the accessible sequence space as it requires a dT, rendering it less feasible for applications where sequence specificity is paramount.

For all three currently available commercial labeling methods, creation of a library of attachment sites can become expensive quickly. Even when labeling in-house, the cost of a modified dT oligo is essentially the same as commercial incorporation of dye prior to shipment. To find a more cost-effective means of incorporation, we looked to the catalog of commercially-available DNA modifications. Phosphorothioate (PS) nucleotides are inexpensive modifications commonly used to inoculate oligos from nuclease degradation[71], where an oxygen on the 5' phosphate is replaced with a sulfur. Previous studies have shown that various attachment chemistries can be incorporated into DNA substrates containing PS nucleotides using bifunctional linkers[72,73]. One such linker, *N,N′*-bis(α-iodoacetyl)-2-2′-dithiobis(ethylamine) (BIDBE), is particularly interesting as it requires minimal synthesis and functionalizes the oligo

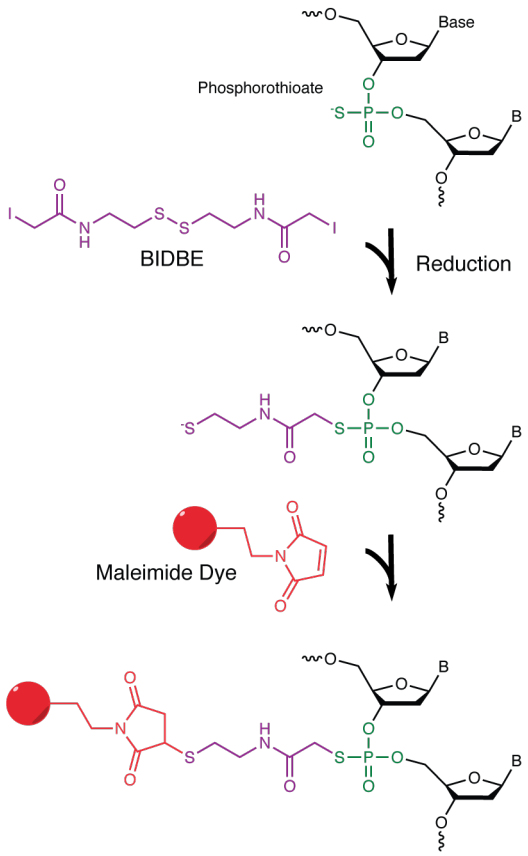

Figure 3.1. Scheme for labeling phosphorothioate DNA. The bifunctional linker BIDBE is attached to the sulfur moiety in an internal phosphorothioate linker through an SN_2 reaction. The disulfide within the linker is reduced using TCEP, producing a reactive thiol, which can then be coupled with a maleimide-functionalized fluorescent dye.

with a free thiol[74], allowing coupling with the wide spectrum of maleimide-functionalized dyes used for cysteine labeling in proteins (Fig. 3.1).

Incorporation of a single PS would create a specific labeling site that is sequence-independent and the attachment chemistry would not limit dipole orientation for FRET applications. Furthermore, the low cost of PS modifications and BIDBE starting materials would increase accessibility to site probing. Here we describe a method for incorporation of fluorescent dyes using PS-BIDBE attachment chemistry.

3.2 Results and Discussion

3.2.1 BIDBE can be synthesized with high purity

BIDBE was initially synthesized as previously described[74]; however, a high quantity of iodoacetate byproduct remained. Residual iodoacetate will compete with BIDBE for PS attachment sites, thereby hindering labeling efficiency. To overcome this issue, we added several steps that do not extend the purification time or difficulty significantly, but yield product with high purity. After the initial synthesis and subsequent centrifugation step, we washed once with 0.1 M NaOH to ensure reaction completion. Following the post-wash centrifugation, the iodoacetate was carefully recrystallized from a smaller volume of acetone, rather than dissolving in 1 ml of hot acetone as described in the literature[74]. Upon solvent evaporation, the remaining powder was noticeably more uniform; the iodoacetate appears whiter than the yellow BIDBE product. Finally, one more wash with deionized water was added, as iodoacetate is moderately soluble in water, while BIDBE is insoluble. The high purity of product was confirmed by liquid chromatography/mass spectrometry (LC-MS) (Fig. 3.2). Notably, the full mass of BIDBE was

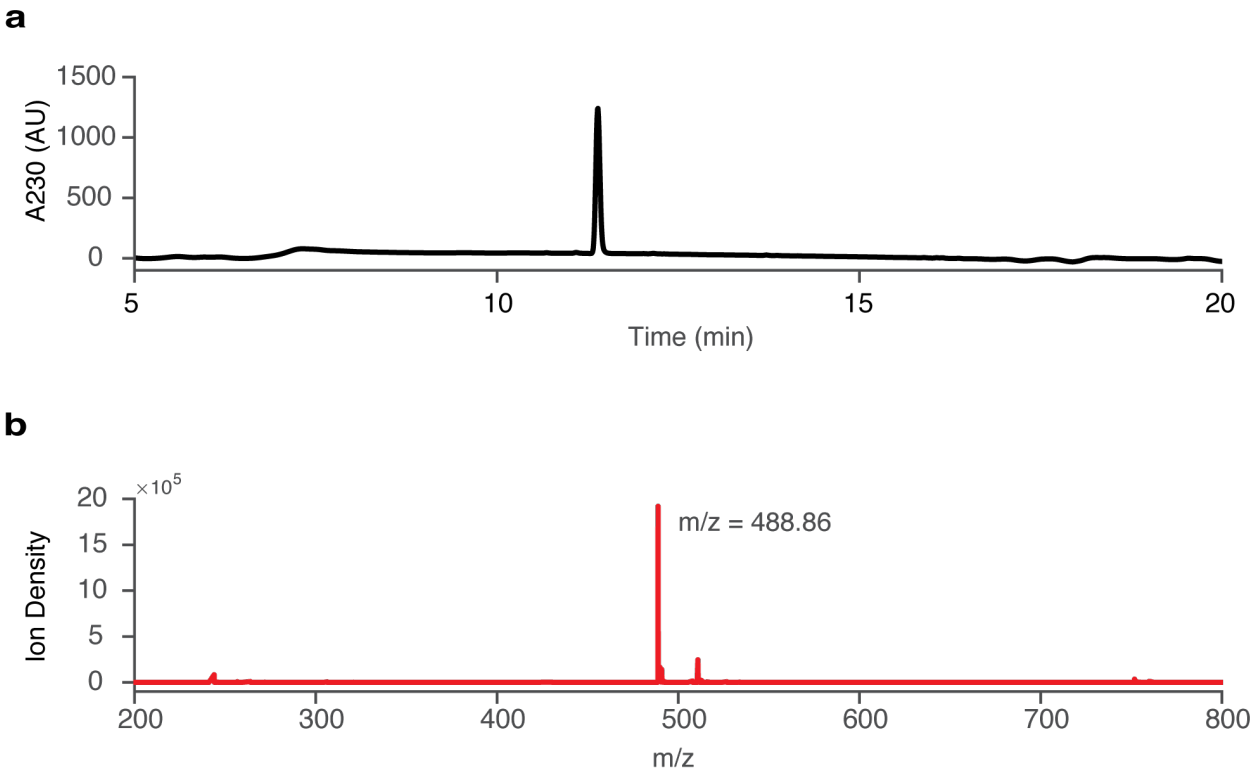

Figure 3.2. LC-MS analysis of BIDBE synthetic product. (**a**) BIDBE retention on a C-18 column with a water/acetonitrile gradient. Only a single peak is present. (**b**) Ion density of the single retained peak in positive mode. The major peak at m/z = 488.86 matches the mass of BIDBE (MW = 487.86 g/mol) with a +1 charge from an added proton. No iodoacetate (MW = 185.95 g/mol) was observable.

predominant with the modified procedure, while it was undetected in previous studies[74] (Fig. 3.2b).

3.2.2 PS oligos can be labeled with fluorescent dyes with limited effect on stability

To test the properties of PS-labeled oligos, a 24-mer containing a single PS (24mer-PS) was modified with BIDBE and reduced (24mer-PS-BIDBE), and the free thiol coupled with maleimide-functionalized Alexa Fluor 647 (24mer-PS-BIDBE-Alexa647). Samples from each step were separated using anion-exchange high-performance liquid chromatography (HPLC) to determine the extent to which the products could be separated (Fig. 3.3). 24mer-PS displays as

two diastereomer peaks (Fig. 3.3a), consistent with previous studies[75]. Upon reaction with BIDBE, a broad peak appears at an earlier elution time and coincides with a significant drop in signal from the unmodified PS oligo, suggesting BIDBE incorporates with high efficiency (Fig. 3.3b). Addition of the fluorescent dye shifts the peak down the gradient (Fig. 3.3c). It should be noted that Alexa Fluor 647 has a number of sulfate groups for increased solubility, which causes the labeled substrate to elute later; more hydrophobic dyes (e.g. ATTO dyes) display peak migration in the opposite direction. The elution differences show that purification of the dye-labeled oligo can be achieved through this method, as well as recycling any unmodified sample if desired.

The dye-conjugated PS oligo was annealed to its reverse-complement and the melting temperature (T_m) was compared to that of a homoduplex (24mer) and an oligo with Cy5 commercially incorporated in the backbone using phosphoramidite synthesis (24mer-Cy5) using UV/Vis spectrophotometry (Fig. 3.4). The measured T_m for the unmodified 24mer is 66.2 ± 0.2 °C, close to the theoretical T_m of 67 °C. While dye conjugation introduces some instability and lowers the T_m values, there is no significant difference between 24mer-PS-BIDBE-Alexa647 (60.3 ± 0.3 °C) and 24mer-Cy5 (59.9 ± 0.3 °C), suggesting the PS labeling does not incur a stability penalty over the traditional method.

Interestingly, there appears to be a biphasic melting curve for the PS-labeled oligo. This observation is likely due to the presence of the two stereoisomers. One phase of the curve appears to show a higher melting temperature than the commercially-labeled oligo at ~65°C, while the other seems to melt about 10°C lower (~55°C). When using this labeling method for applications with higher experimental temperature conditions, the fraction with the

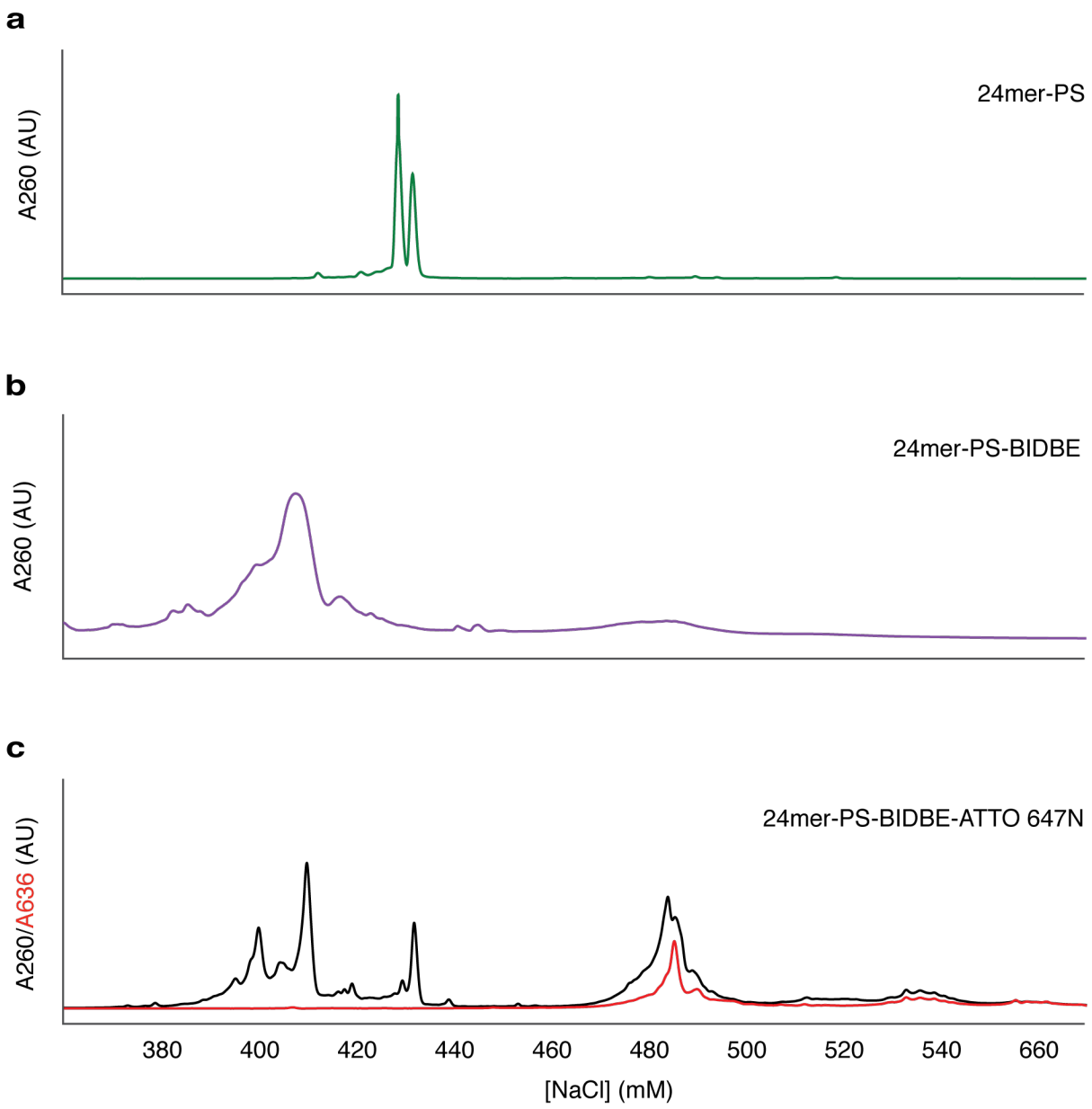

Figure 3.3. HPLC separation of a 24mer containing a single phosphorothioate throughout the labeling process. Elution time was converted to NaCl concentration to normalize retention times from different gradient starting points; elution buffer was increased at the same rate in all traces. (**a**) commercially synthesized 24mer-PS. (**b**) 24mer-PS after reaction with BIDBE linker. (**c**) 24mer-PS-BIDBE coupled with Alexa Fluor 647 maleimide dye.

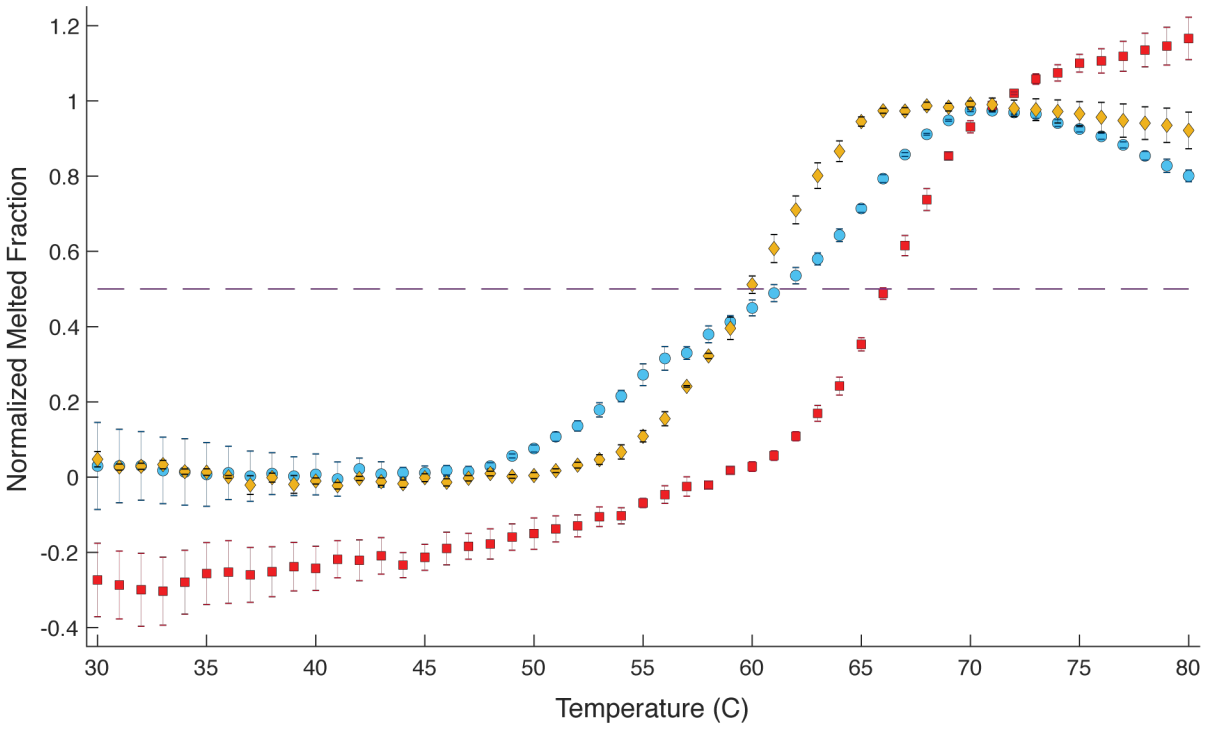

Figure 3.4. Melting temperature comparison of backbone labeling methods. Unmodified 24mer (red), 24mer-Cy5 (yellow), and 24mer-PS-BIDBE-Alexa Fluor 647 (blue) were annealed to the reverse compliment oligo. The absorbance at 260 nm was measured at 1°C intervals. Values were normalized as described in 3.4.6. 50% melting is denoted by the dashed line.

lower melting temp may alter results. However, the impact of this may be minimal when used at 37°C , where no apparent denaturing occurs, or when using an oligonucleotide longer than 24nt.

3.2.3 Labeled Holliday junctions provide a proof of concept for PS labeling

An interesting example substrate for this labeling method is the Holliday junction (HJ). Holliday junctions (HJ) are complex, multimeric DNA structures formed during homologous recombination[76] and double-strand break repair[77]; synthetic HJs are formed by annealing four compatible oligos into a cruciform structure that can adopt multiple conformations (Fig. 3.6a), the dynamics of which have been explored using single-molecule FRET. Previous collaboration between the Erie and Sekelsky labs focused on the ability of *Drosophila melanogaster* Gen (dmGen), a structure-specific endonuclease, to cleave HJs *in vitro*[78]. dmGen requires longer arms (24-25 bp) for binding, necessitating internal labeling to maintain a FRET-capable inter-dye distance[79].

To test the efficacy of bulk-fluorescence applications, two PS oligos were labeled with a FRET dye pair (Alexa Fluor 546 and Alexa Fluor 647) without HPLC purification following BIDBE attachment. The oligos were included in an HJ substrate (Fig. 3.5a) and treated with dmGen. Native polyacrylamide gel electrophoresis (PAGE) analysis (Fig. 3.5b) showed that dmGen cleaves the PS-labeled HJ with a similar rate to previous studies (Fig. 3.5c), indicating the label does not interfere with binding. All cleavage products with dye incorporation are identifiable; when attempting the same analysis using ^{32}P-labeled substrate, this result can only be achieved by performing electrophoresis under both native and denaturing conditions[78]. Interestingly, a band of smaller size than intact HJ appears in the green channel that is absent in the red (Fig. 3.5b, left). Incubation with dmGen results in the disappearance of this species, along

with the full HJ, and subsequent appearance of a product band of larger size than the products of HJ cleavage. We surmise that the original species is comprised of three of the four component oligos where the strand containing the acceptor dye is absent, likely the result of dissociation during the electroelution step of HJ purification. This substrate would have both a 5' and 3' flap available for cleavage (Fig. 3.5b, center), of which only the former is cleaved by dmGen[78]. The resulting product would be larger than any of the observable HJ cleavage products, which is consistent with our PAGE results.

For single-molecule FRET experiments (Fig. 3.6), identical HJs were made using Alexa Fluor 555 and Alexa Fluor 647 as donor and acceptor dyes, respectively. Following dye incorporation and prior to substrate annealing, component oligos were purified using anion-exchange HPLC to reduce the presence of substrate without an active donor/acceptor fluorophore. A 5' biotin present on an unlabeled strand was utilized to conjugate the substrate to a modified quartz slide bound with streptavidin. Total internal reflection imaging in physiological-salt conditions was performed, with Ca^{2+} present as a divalent cation, in the absence and presence of dmGen. The majority of traces observed in the absence of dmGen display rapid transitions between FRET states (Fig. 3.6b), while addition of dmGen tends to reduce the number of state transitions (Fig. 3.6c). In the absence of dmGen, there is a predominant distribution around 0.3 FRET, with smaller populations at higher FRET efficiencies near 0.45 and 0.65. (Fig. 3.6d). This main distribution persists upon addition of dmGen, however the higher FRET populations are not present; there appears to be a shift to a lower FRET state (0.10 – 0.15).

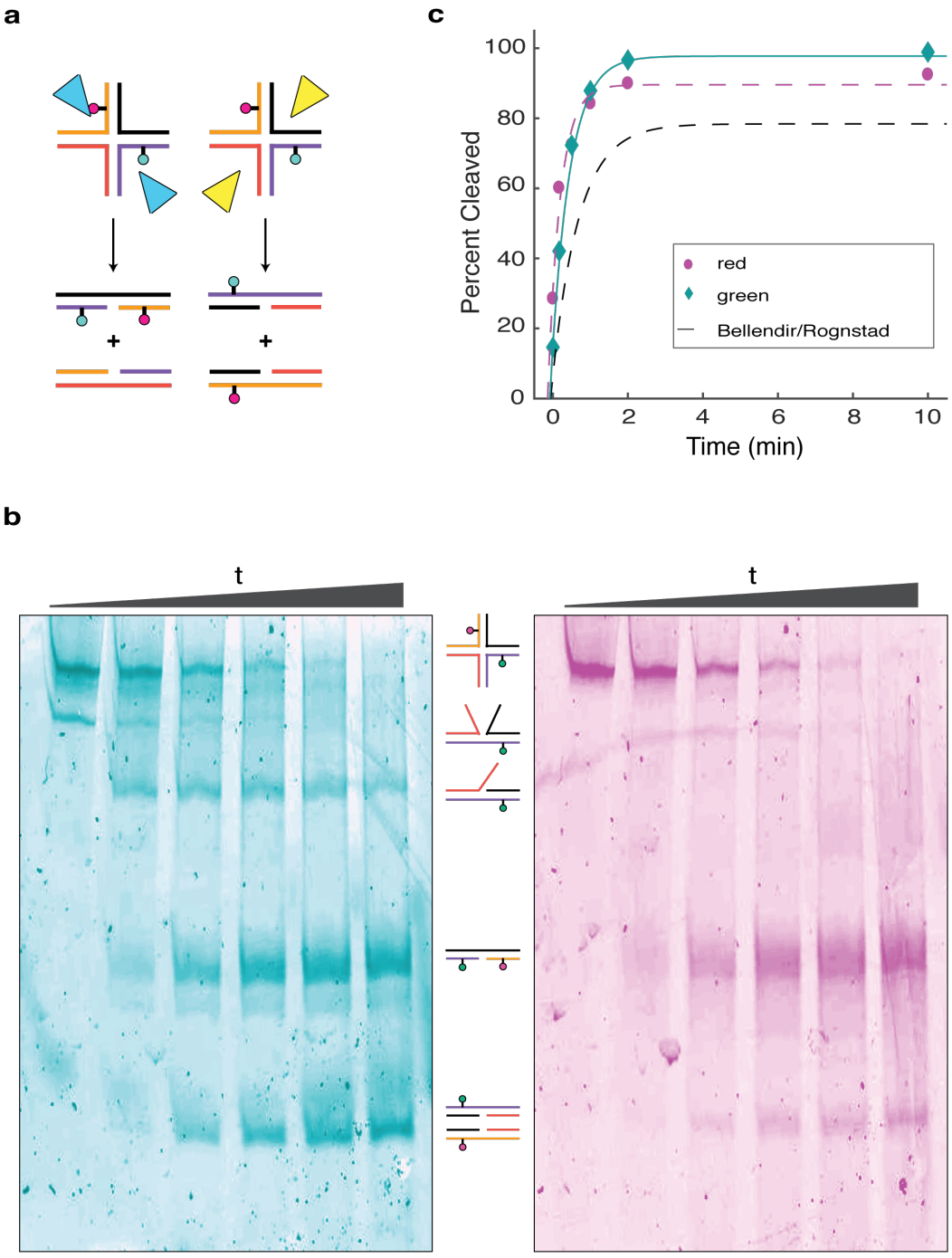

Figure 3.5. Electrophoretic analysis of cleavage and product formation of PS-labeled Holliday junctions. a, schematic of cleavage orientations and respective products. **b**, pseudocolor images of native PAGE gels filtered for donor (left, 570 nm filter) and acceptor (right, 670 nm filter) channels with illustrations of cleavage products for each identified band (center). Lanes reflect timepoints collected at 0s, 10s, 30s, 1m, 2m, and 10 min from left to right. **c**, plot of reduction in intact HJ as a percentage of total lane fluorescence (**b**) for donor (cyan) and acceptor (magenta) at each timepoint compared to cleavage kinetics determined by Bellendir et al. (*14*).

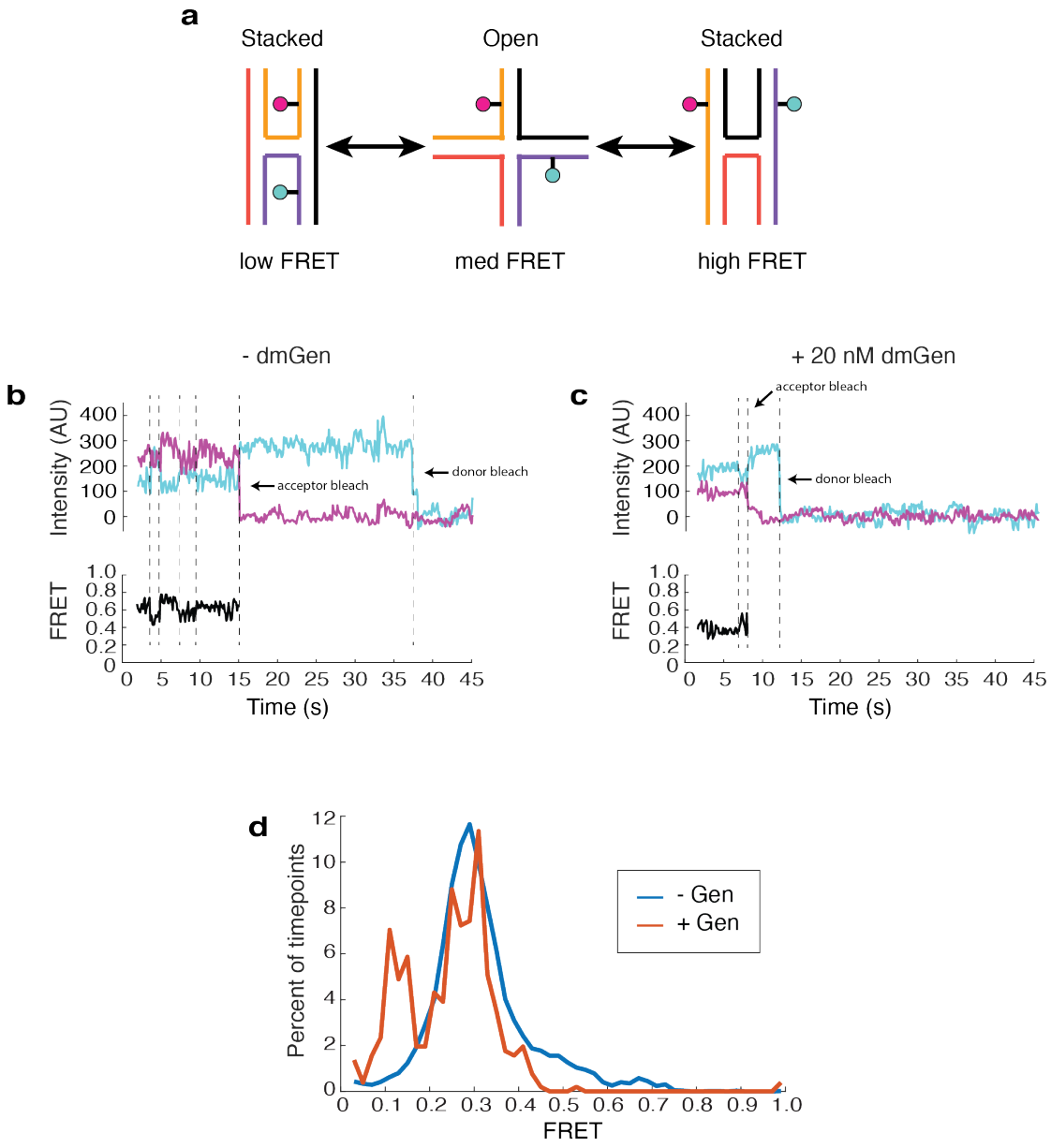

Figure 3.6. smFRET applications of Holliday junctions containing PS-labeled oligos. a, schematic of relative dye locations in various conformations of synthetic HJs. **b**, example smoothed time-intensity traces for donor (cyan) and acceptor (magenta) dyes (top) and calculated FRET trace (bottom) in the absence of dmGen. Identified state transitions are shown by dashed lines, with donor and acceptor bleaches denoted. **c,** example smoothed time-intensity (top) and FRET (bottom) traces in the presence of 20 nM dmGen. **d**, distribution of timepoint percentages by FRET efficiency value (bin size = 0.02 FRET) in the absence (blue, n = 60 molecules) and presence (red, n = 9 molecules).

3.3 Conclusion and Future Directions

We have demonstrated that DNA oligonucleotides can be fluorescently labeled using a single PS modification. A single BIDBE synthesis as described here yields approximately 10 reactions and PS-oligos can be modified in parallel to create a library of maleimide-compatible labeling sites. The process from BIDBE synthesis to dye incorporation can be done in as little as 2-3 days if no HPLC purification is required. The synthetic route is accessible, requiring no special equipment and performed in a microcentrifuge tube with inexpensive starting materials. DNA labeled in this manner exhibits stability similar to that of commercial backbone labeling, but the cost-efficiency allows for moving dye locations and interchanging dyes for different applications, while avoiding potential FRET impacts or sequence space limitations of traditional methods.

Currently, BIDBE is used at 200-fold excess to ensure complete conjugation. Preliminary tests suggest that labeling still occurs at a similar efficiency when using slightly less linker, although a titration of linker amounts would be needed to assess the optimum concentration; using less linker would further increase the cost differential between our method and commercial phosphoramidite synthesis. Furthermore, incubation time and temperature optimization would shorten the process and potentially allow for labeling of degradation-sensitive substrates. We would like to optimize the different steps of this protocol to provide a shorter, more efficient procedure. This labeling method will then be deployed to create substrates for smFRET structure-function studies of dmGen-HJ interactions.

3.4 Methods

3.4.1 Materials

Iodoacetic anhydride, cystamine di-hydrochloride, and fluorescent dyes were purchased through Sigma-Aldrich (St. Louis, MO). Oligonucleotides were ordered from Integrated DNA Technologies (Skokie, IL). The sequences used can be found in Table 3.1, with incorporation of a phosphorothioate nucleotide denoted by an asterisk. Commercial oligos were dissolved in 10 mM Tris pH 8 to 1mM concentration.

Table 1.1. Oligonucleotide sequences for PS labeling experiments

Name	Length (nt)	Sequence	Reference name (*14*)	Figure
888-Biotin	49	Biotin-GACGCTGCCGAATTCTGGCGTTAGGAGATACCGATAAGCTTCGGCTTAA	888	3.6a (red)
891-C16-PS	49	ATCGATGTCTCTAGAC*AGCACGAGCCCTAACGCCAGAATTCGGCAGCGT	891_SPB	3.6a (purple)
990-A34-PS	49	TTAAGCCGAAGCTTATCGGTATCTTAGCAATGTA*ATCGTCTATGACGTT	990_SPB	3.6a (yellow)
991_SPB	51	CAACGTCATAGACGATTACATTGCTAGCTCGTGCTGTCTAGAGACATCGAT	991_SPB	3.6a (black)
24mer	24	TTAAGCCGAAGCTTATCGGTATCT	992	3.4 (red)
24mer-PS	24	TTAAGCCGAAGCTTAT*CGGTATCT	992	3.4 (blue)
24mer-Cy5	24	TTAAGCCGAAGCTTA-Cy5-TCGGTATCT	992	3.4 (yellow)
24mer reverse compliment	24	AGATACCGATAAGCTTCGGCTTAA		

* denotes PS nucleotide

3.4.2 Synthesis of BIDBE

Iodoacetic anhydride was dissolved in dichloromethane in a 1.5 mL microcentrifuge tube, to which was added cystamine di-HCl dissolved in 0.1 M NaOH. The resultant biphasic solution was vortexed for one minute, forming a thick white precipitate consisting of iodoacetate and the bifunctional liker BIDBE. The suspension was centrifuged at 13,000 RCF for 10 min, the aqueous layer removed, and the organic layer dried *in vacuo*. The pellet was resuspended in 0.1 M NaOH to ensure completion of the reaction and the centrifugation repeated and supernatant removed. With the tube submerged in a hot water bath, boiling acetone was added dropwise and periodically vortexed until the pellet dissolved. The remaining iodoacetate was then recrystallized at room temperature and the slurry centrifuged at 13,000 RCF at 4°C for 15 min. The supernatant was removed and blown dry with air. The resulting solid was resuspended in deionized water, centrifuged at 13,000 RCF for 10 min, and the pellet was allowed to dry overnight at room temperature. The purity of BIDBE was assessed by dissolving 1 mg in acetonitrile and analyzing with an Agilent 6520 Accurate Mass QTOF LC-MS ESI positive in high-resolution mode, equipped with a Thermo Scientific Acclaim RSLC 120 C18 column (2.2 µm particle size, 120 Å pore diameter, 2.1 x 150 mm).

3.4.3 Thiol-modification of phosphorothioate DNA oligonucleotide

24mer-PS (10 nmol) was added to 36 µl 50 mM phosphate buffer pH 7. BIDBE (1 mg, 100 nmol) dissolved in dimethylformamide was added and the reaction was incubated at 50°C for 6 hr. A five-fold molar excess of TCEP was added to reduce the disulfide for 1 hr at room temperature and the reaction cleaned up using a Zymo Oligo Clean and Concentrate kit (Zymo Research), eluting in 10 mM Tris pH 7.5.

3.4.4 Fluorescent dye labeling

Maleimide-functionalized fluorescent dye was dissolved in dry DMSO and mixed at a 10:1 molar ratio with thiol-modified oligo in 50 mM Tris pH 7.2. The reaction was either incubated at room temperature for 2 hours or at 4 °C overnight. For endonuclease assays, the mixture was passed through a G50 microcentrifuge spin column to remove excess dye; for HPLC purification, no further cleanup was used.

3.4.5 HPLC purification

Chromatography was performed using a 1260 Infinity binary liquid chromatography system equipped with a Waters Gen-Pak Fax anion-exchange column (2.5 µm, 4.6 x 100 mm) equilibrated in 90% Buffer A (50 mM Tris-HCl pH 8.0) and 10% Buffer B (50 mM Tris-HCl pH 8, 1M NaCl) with a flow rate of 0.5 mL/min. To establish elution concentrations of NaCl, test samples from each stage of the process were injected and allowed to equilibrate for 5 minutes, followed by increasing Buffer B (1%/min, 10 mM NaCl/min) to 75% Buffer B (750 mM NaCl). Subsequent purifications were adjusted based on the elution times for the test samples, which depend on the length of the oligo and the hydrophobicity of the dye used, such that the initial equilibration was at least 100 mM NaCl (10% Buffer B) below elution concentration and ramped to 100 mM NaCl above elution of the final peak using the same gradient rate. The eluent was monitored at 260 nm, 555 nm, and 636 nm and elution fractions collected at 30 second intervals. Fractions containing labeled DNA were pooled and concentrated using a Zymo Oligo Clean and Concentrate spin column, eluting in 10 mM Tris pH 8.

3.4.6 DNA melting curves

Absorbance measurements were recorded using a Cary 100 UV/Vis spectrometer with a Peltier attachment (Agilent). Modified oligos were combined with equimolar reverse compliment to a final concentration of 500 nM in degassed 10 mM Tris pH 8 and 100 mM NaCl. The oligo mixtures were transferred to sealed quartz cuvettes, heated to 90°C for 5 min, cooled slowly to room temperature, and equilibrated for 30 minutes. The absorbance at 260 nm was monitored at 1°C increments from 25°C to 80°C with a heating rate of 1°C/min. The absorbance was then plotted against temperature and the regions before and after the transition region were fit linearly to correct for small changes in hyperchromicity, as well as the steepest part of the major transition. The intersection of the transition line and those of the pre- and post-transition regions were then set to 0% and 100% of the total DNA melted, respectively, to normalize for differences in total absorbance values. The melting temperature was then calculated for the temperature at which the transition line is equal to 50% melting.

3.4.7 HJ purification

Care was used throughout to protect the fluorescent dyes from light to limit photobleaching. HJ oligos (888-Biotin, 891-C16-PS, 990-A34-PS at 2 nmol each; 991_SPB at 6 nmol, 3-fold excess) were mixed in 65 µl annealing buffer (50 mM Tris pH 7.5, 50 mM NaCl, 5 mM DTT), heated to 95°C for 5 min, and cooled slowly to room temperature. The solution was mixed with an equal volume of 2X loading buffer (20 mM Tris pH 8.0, 2 mM EDTA, 10% glycerol), loaded onto an 8% native Tris-borate-EDTA (TBE) PAGE gel, and electrophoresis was performed with 200 V for 2 hr. The gel was imaged using a Typhoon Trio+ fluorescence imager (GE Healthcare), overlayed on the image, and the HJ band excised. The gel slice was

placed in dialysis tubing (45 kDa cutoff) containing 2mL TBE electrophoresis buffer. The tubing was placed in a gel box containing TBE and electroeluted with 200 V at 4°C for 2 hr, at which point the current was reversed for 15 minutes. The gel slice was removed and the eluted HJ dialyzed at 4°C against 4 L storage buffer (10 mM Tris pH 8, 50 mM NaCl), refreshing the buffer every 24 hours for 72 hours. The buffer was removed from the tubing and concentrated using a 14 kDa Microcon microcentrifuge filter (MilliporeSigma) and stored at 4°C wrapped in aluminum foil.

3.4.8 HJ endonuclease activity

Purified HJ (20 nM) was incubated with dmGen (200 nM) in 40 µl reactions (50 mM Tris pH 8, 100 µg/mL BSA, 1 mM DTT, 10% glycerol, 50 mM KCl, 5 mM $MgCl_2$). Reactions were initiated upon addition of dmGen and 5 µl aliquots were removed and quenched with 10 µl 2.5 mg/mL Proteinase K and 125 mM EDTA for 15 min at each timepoint (10 s, 30 s, 1 min, 2 min, 5 min, 10 min). The samples were diluted 1:2 with 2X loading buffer and electrophoresis was performed with an 8% native PAGE gel. Amounts of substrate cleavage and product formation were determined by the fraction of total lane fluorescence.

3.4.9 Surface preparation for TIRF imaging

Imaging was performed using quartz slides drilled to create multiple flow channels[80]. Slides and glass coverslips were bath-sonicated in anionic detergent, acetone, ethanol, and KOH (in that order), using deionized water to rinse between steps and for final storage. Aliquots of methoxy-poly(ethylene glycol)-silane (mPEG-silane, 20 mg, Laysan Bio, MW 2000) and biotin-poly(ethylene glycol)-silane (bPEG-silane, 2 mg, Laysan Bio, MW 3400) were measured in a

glove box. Slides and coverslips were blown dry under compressed air. Single aliquots of mPEG-silane and bPEG-silane were dissolved in 80 µl and 10 µl of deionized water, respectively, and 1 µl bPEG-silane was added to the mPEG-silane to create a 1:100 biotinylated/non-biotinylated ratio, vortexed, and briefly centrifuged to remove bubbles. For each slide, ~40 µl of PEG-silane mixture was applied to the channel area and covered by a coverslip, removing any air bubbles. The slide/coverslip combination was then suspended in a box partially filled with water to maintain humidity and allowed to react overnight. Upon completion, the slide and coverslip were separated, the treated surfaces rinsed thoroughly with deionized water, and blown dry with compressed air. A second treatment of mPEG-silane was then applied to fill any potential functionalization gaps, followed by rinsing and drying when sufficient reaction time had elapsed. Strips of double-sided tape were adhered to the slide between the channel holes and the coverslip attached to create flow channels.

3.4.10 Sample treatment for TIRF imaging

Functionalized flow channels were rinsed 5x with loading buffer (10 mM Tris-HCl pH 8, 50 mM NaCl, 10 mM $CaCl_2$). A streptavidin solution was diluted to 0.1 mg/ml in loading buffer, added, and allowed to react for 15 min, followed by rinsing 3x with loading buffer. Labeled and biotinylated HJ substrate was diluted to 40 pM with loading buffer and applied to the channel for 15 min. The channel was then rinsed 2x with imaging buffer (50 mM Tris-HCl pH 8, 50 mM NaCl, 10 mM $CaCl_2$ for low salt conditions; 50 mM Tris pH 8, 100 µg/mL BSA, 1 mM DTT, 10% glycerol, 100 mM KOAc, 10 mM $CaCl_2$ for endonuclease conditions). Immediately prior to imaging, an oxygen scavenging and triplet state quenching system (100 U/ml glucose oxidase,

1000 U/ml catalase, 0.05 mg/ml cyclooctatetraene, 143 mM 2-mercaptoethanol, 2% w/v glucose) prepared in the corresponding imaging buffer was added[6,7].

3.4.11 Single-molecule TIRF image collection

Donor and acceptor fluorophores were excited using 532 and 640 nm lasers, respectively, and imaged on a through-prism total internal reflection microscope with a 60X, 1.2 NA water immersion objective. Donor and acceptor channels were split using a DualView optical splitter with a 645 nm dichroic mirror and captured by an emCCD camera using 585/70 bandpass and 655 longpass filters at a 100 ms framerate for 800-900 frames. Acceptor fluorophores were excited during the first 10 frames to locate substrate molecules, followed by 790 frames of donor excitation, at which point acceptor fluorophores were excited for >10 frames to determine the extent of acceptor photobleaching.

3.4.12 FRET data analysis

TIRF image analysis was performed using custom software in Matlab[7]. Fixed fluorescent bead images were used to determine any offset between the donor and acceptor halves of the images and acceptor molecules were located using the initial 10 frames. Intensity traces were generated for each acceptor (I_A) and its corresponding donor (I_D). Raw FRET efficiency (E) was calculated by $E = I_A / (I_A + I_D)$ and smoothed using a modified Chung-Kennedy edge-preserving filter. Only those molecules exhibiting nonzero FRET and intensity consistent with a single donor/acceptor pair were considered for analysis. Traces containing at least one donor and acceptor blinking or photobleaching event were corrected for background and bleedthrough while the remainder were discarded, after which FRET efficiency was recalculated.

CHAPTER 4: MAKING AN UNDERGRADUATE LABORATORY COURSE FROM A GRADUATE RESEARCH PROJECT

The groundwork for the development was performed with Hunter Wilkins. This included initial test experiments, examination and activity design, lecture design, preparatory work, and in-course data analysis. Survey data analysis was performed by Bryant Hutson in the UNC Office of Institutional Research and Assessment. The instruction and course development were overseen by Dr. Thomas Freeman and Dr. Dorothy Erie.

4.1. Introduction

For decades, the standard of undergraduate physical and life science labs has been an expository approach, where students are guided to a known outcome by instructors in order to provide experience in the broad techniques of biochemical research[81]. While this "cookbook" approach is well-established and easily implemented across undergraduate institutions, it is inadequate as a realistic model of research in a graduate or professional environment[82,83]. These common pitfalls in traditional laboratory curricula have led to prescriptions by the American Association for the Advancement of Science (AAAS)[84] and the President's Council of Advisors on Science and Technology (PCAST)[85], among others, to update undergraduate laboratory methodology to incorporate more research-based instruction.

Traditionally, undergraduate research experience is gained through independent research overseen by a research faculty member, typically through joining the research group[86,87]. While this provides an excellent resource for undergraduates, this experience is often pursued outside of the course load[86]; furthermore, the number of available research positions limits the number of students who can participate[87]. The effect of these limitations is that most students may not have

the opportunity to experience research in a realistic setting, the impact of which is particularly pronounced within underrepresented minority (URM) student communities[88].

In recent years, course-based undergraduate research experience (CURE) courses have been increasingly incorporated to address some of these shortfalls[89]. Numerous studies have shown that such inquiry-based approaches result in greater internalization of scientific concepts; increased self-confidence in physical skills, data analysis and interpretation; and critical thinking skills involved in synthesizing scientific hypotheses and assertions[87,90–93]. Completion of research-based courses have also shown lower rates of STEM student attrition[91,94], while increased access to scientific mentorship[95] and quality research experience[96] correlates closely with increased self-identification as a scientist by URM students and, in turn, higher likelihood of persistence of students from those social backgrounds within STEM careers[97].

In an effort to broaden participation in authentic research experiences, the University of North Carolina at Chapel Hill (UNC) adopted a Quality Enhancement Plan (QEP) beginning in 2017 designed to incorporate a research-based model for undergraduate laboratory experiences. A key component of the Southern Association of Colleges and Schools Commission on Colleges (SACSCOC) reaccreditation requirements, the QEP is specified as a focused course of action that addresses a well-defined issue related to student learning. As a leading research university, UNC-Chapel Hill has identified providing meaningful research experiences to undergraduates as an institutional priority. Included in this QEP are CURE programs for the physical and life sciences, with the prescription that these CURE courses incorporate certain core characteristics (Table 1) into their curriculum outlined in the literature[98,99].

We have created an undergraduate biochemical curriculum that aims to provide a research experience comparable to a graduate-level project to be carried out over the course of

multiple successive semesters by different student cohorts. In this study, we explore how to best offer students an authentic research experience through CUREs and examine how implementing a CURE model impacts student research outcomes, including sense of research project ownership and research skill development. We describe a CURE approach implemented with a course focused on biochemistry laboratory techniques. We delineate the CURE project structure as well as the instructional strategies employed in the course. In addition to describing student research outcomes, we report the impact the CURE project had on students' perceptions and attitudes toward engaging in scientific work as well as the class itself. Our specific research questions for this study include:

1) To what degree was the CURE model successful implemented in the biochemistry course (as measured by the Laboratory Course Assessment Survey, or LCAS)?
2) What is the impact of implementing a CURE model in a biochemistry laboratory techniques course on students' research skill development in the discipline?
3) What is the impact of implement a CURE model in a biochemistry laboratory techniques course on students' perceptions and attitudes toward engaging in scientific work, including sense of project ownership, research skill development, etc.

4.2 Course Development

At UNC, the biochemistry lab is a three-credit-hour lab primarily taken by third- and fourth-year biochemistry majors. The traditional aims of the course have been to introduce students to the skills involved in the biochemical study of enzymes and scientific reporting of the findings. The class is divided into two sections of a maximum of 16 students per section. Each

section meets twice weekly for four-hour periods, as well as a single hour-long recitation period attended by both sections. In addition to a faculty instructor, there are four graduate teaching assistants (TAs), where each section has two assigned to it.

4.2.1. Rationale for choosing a research project

The initial motivation for moving away from a cookbook-style laboratory was to provide students with a shared research experience that more readily translates to a career in biochemical research. The previous target protein, alkaline phosphatase, has been rigorously studied for decades. As such, the outcomes of the experiments performed in the course are known. Conversations with students in the past provided anecdotal feedback that this bred a lack of interest in students; the feeling was that they were doing work for the sake of work.

The aim of this course is two-fold. The first aim is for students to come away with realistic expectations of what a career in science will entail. The second aim is for the research the students perform to be included in a peer-reviewed publication. Contributing to the body of knowledge allows students to feel ownership of the project and a sense of real accomplishment beyond completion of the course.

In pursuing these aims, we decided that choosing a novel target for the class at the outset was the best option, allowing the class to learn the techniques with initial direction from the instructors. This approach creates a controlled environment for students to practice the critical thinking, analysis, and experimental design skills necessary for a career in research beyond those taught in a traditional laboratory course.

When choosing a novel target to study, it is pertinent to find something that is feasible for novice scientists to approach. With this in mind, we looked for an uncharacterized homolog of a

previously studied enzyme familiar to us through our academic research. This approach means that the instructor has intimate knowledge of the process of purification and assay design that are crucial to timely progression. Students can also recapitulate the results from the characterized homolog in parallel, providing a built-in control to compare with the novel results. For our pilot study, we selected *Thermus aquaticus* (*Taq*) UvrD, a DNA helicase whose homolog is a key component in nucleotide excision repair in *Escherichia coli*. The *E. coli* homolog serves as a control for DNA unwinding and ATPase activity.

4.2.2. Course progression

The updated curriculum emphasizes the development of what we have termed the "biochemist's toolkit." We want students to understand, in both theory and practice, the array of experimental techniques at their disposal to investigate any problem that they might encounter in a research career. This toolkit includes those molecular biology techniques standard to the isolation of any enzyme – PCR, primer design, plasmid vectors, bacterial strains, separation methods – while also including more advanced experiments that may not be covered by other courses and lectures. The overarching aim of this approach is to have students work through the development of a research project as they would in a graduate environment: using the resources available (e.g. software, literature sources) to understand the limitation of the current knowledge base, design experiments using the techniques available to them, and collect and analyze data to test their hypotheses. While the scope of these experiments is naturally limited to the techniques presented to them in the course, the process is broadly applicable to identifying and solving research problems in a professional research environment.

It is not feasible to complete the vast majority of graduate-level research projects in a single semester. Furthermore, it is virtually impossible to predict *a priori* which novel targets would fall into that minority. Thus, the aim of this new curriculum is not to design a stand-alone project to be completed by a particular cohort, but rather to build each successive semester's experimental workflow upon the results obtained by previous students. The result is a dynamic syllabus that attempts to answer those questions that remain previously uncompleted or unaddressed. The adaptability of this curriculum provides the opportunity to reinforce those techniques that the students find particularly challenging and introduce more non-traditional experimental design that might otherwise be precluded by the constraints of traditional course design.

At the inception of our new curriculum, emphasis was initially placed on determining optimal conditions for purification of *Taq* UvrD and enzyme assays. Students performed initial induction tests using isopropyl-β-D-thiogalactoside (IPTG), which yielded limited soluble UvrD; this provided an opportunity to introduce the concept of autoinduction, a technique unfamiliar to the majority of the class. To reduce the protein load during column chromatography, an ammonium sulfate precipitation was employed, the optimal concentration of which was unknown for our homolog. Finally, all groups performed identical affinity chromatography. Results were monitored for each step using polyacrylamide gel electrophoresis (PAGE).

We chose to test two facets of *Taq* UvrD activity: DNA unwinding (helicase activity) and ATP hydrolysis (ATPase activity). The extent of helicase activity is determined by incubating UvrD with DNA substrate where one strand is labeled for visualization and the opposite strand contains a 3' overhang for enzyme binding. To prevent reannealing of the labeled strand, an equivalent, unlabeled "trap" oligo must be included in excess. Single- and double-stranded

substrate are subsequently separated using native PAGE, and relative amounts calculated. ATPase activity can be measured using a colorimetric Malachite Green assay, which measures the concentration of inorganic phosphate following hydrolysis.

Assay conditions for both helicase and ATPase activity needed to be defined before meaningful results could be acquired, as the literature procedures for both involved use of radioactivity. The helicase substrate was modified to include a fluorescent dye, which is considerably safer but reduces the sensitivity of the assay. The concentrations for substrate, enzyme, and trap oligo needed to therefore be determined. The conditions for the Malachite Green assay were unknown from the outset. After conditions for all of the experiments were established, the structure of subsequent semesters shifted to focus on producing enzymatic results and introduction of techniques like mutagenesis.

4.2.3. Collaboration

At the start of each semester, students complete a pre-course survey conducted by the course instructors to establish each student's experience background. Students are sorted by experience and divided into groups of four so that each group contains students with diverse levels of experience. This allows students to help teach their group about techniques they've used in other settings, which helps students reinforce their knowledge and frees teaching assistants to assist those who need more instruction.

The aims of the group structure are to facilitate cooperative progression through lab activities, promote collaborative writing and peer review, and to efficiently perform experimentation. Students are expected to divide work duties amongst themselves for

experimental preparations (e.g. making buffers, pouring gels), reducing the setup time for the more laborious experiments.

The nature of novel research necessitates troubleshooting and optimization of the various steps in the process. This is perhaps the largest departure from a traditional curriculum where the procedure is more concrete, but it serves to provide a realistic experience in graduate-level research. We can take advantage of the group structure to tackle this challenge in an efficient manner; each group can run an internal control, whether to compare a specific condition using the target enzyme or results from the characterized homolog, along with differing test conditions between groups. This method allows for many conditions to be tested simultaneously so that all members get experience with the techniques and systematic error can be distinguished from negative results. We can then compare the results of all groups to determine which conditions are optimal.

In this vein, the weekly recitation period has been converted to a lab-meeting-style discussion involving both sections. Each group performs analysis of the previous week's results and presents their conclusions, determining what the next experimental step should be. If the prior experiments were determined to have failed, the students hypothesize what possible explanations of the outcomes exist and how to experimentally troubleshoot those sources of error. This environment allows the instructors to introduce analysis techniques and software to the class as a whole, as well as incorporation of the same information as would be provided by a traditional lecture-style recitation.

For *Taq* UvrD, there were many unknowns in the procedure that would normally take a significant amount of time to establish. In purification, differing IPTG and ammonium sulfate concentrations were compared between groups, as were relative amounts of enzyme, substrate,

and trap oligo in helicase and ATPase assays. By dividing the work among the groups, we were able to test a wide variety of conditions in just a few lab periods, more than would likely be feasible for groups to test individually in an entire semester. All groups were still exposed to the conditions they didn't test through the recitation meeting, which allows all groups to interpret the gamut of results and design future experiments.

4.2.4. Computational techniques

We have included an introduction useful to computational techniques not previously included in this course, inserted during periods of downtime during points in the workflow where downtime is unavoidable (e.g. incubation periods). These lessons include crystal structure modeling in PyMol (Schrödinger, New York, NY), basic local-alignment search tools (BLAST, National Center for Biotechnology Information), and the molecular biology software suite SnapGene (GSL Biotech, San Diego, CA).

Through multiple-sequence alignment, students identify those residues that are conserved throughout homologs identified in other species. If no crystal structure exists for the target enzyme, the sequence can be submitted to protein-fold-modeling algorithms. Students can then determine those conserved residues and structural elements of functional importance that model closely to available structures, as well as major differences that may affect function. Visualization of these characteristics provide an opportunity to discuss amino acid side chain properties and the impact of changing residues to alter function, allowing students to formulate hypotheses about differences in binding or kinetics and to define change- or loss-of-function mutants that might be interesting to further pursuit. Finally, students can use molecular biology tools to design a mutagenesis experiment.

4.2.5. Presentation of experimental results

Students are required throughout the semester to present their findings in multiple formats. During the weekly recitation, groups are asked to annotate figures from recent experiments and describe their interpretation of the outcome. This requires defending those interpretations through the data and allows instructor feedback in real time. This can be presented within a particular group, between groups who tested similar conditions, or to the entire class.

We have shifted to a single, collaborative written report for each group, as opposed to multiple reports written by each individual throughout the semester. This approach more closely resembles real-world scientific writing, where multiple scientists contribute to a single body of work, a process that students likely have not participated in previously. The students are again responsible for holding each other accountable; an adjustable rubric can be established that considers the differing contributions of each member. We opt for gradual submission of components following completion of major sections of the procedure (i.e. purification, assays) so that students can have incremental feedback.

At the completion of the course, groups are required to present a poster at the UNC-wide undergraduate research symposium. For many students, this is their first experience in presenting an original research project to an external audience. The abbreviated format of a poster presentation also provides experience in distilling their results to the essential information.

4.2.6. Course materials

Novel research precludes the availability of a formal lab manual. Students are instead required to turn to literature sources, local lab protocols, and internet searches to find

Table 4.1. Overall LCAS results

Scale (Possible Range of Scores)	Collaboration (6-24)		Discovery (5-30)		Iteration (6-36)		Total (17-90)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
CHEM 530L	21.73	4.17	22.45	3.59	28.73	5.64	72.91	8.71
Original Corwin et al (2015) study								
CUREs	21.11	3.20	24.35	4.04	28.71	4.15	75.10	8.67
Traditional lab	20.87	4.02	20.77	5.82	26.53	7.00	68.15	14.76

experimental methods. This enforces skills in literature interpretation and self-sufficiency that are valuable in professional settings. Students are also provided with commercially produced video tutorials for different techniques.

In lieu of hard copies of pertinent protocols and background information, we take advantage of cloud-based services so that students have access to course-related information, literature references, and group results from each experiment. We worked in conjunction with UNC Information Technology Services to create a course site using Microsoft OneNote, as all students have access to the software the university site license for Microsoft Office 365. Student notebooks were access-limited to the student themselves and those with instructor permissions. This ensures individual responsibility for contemporaneous notetaking while allowing for periodic review by instructors. The option was given for electronic recordkeeping or for a physical notebook with the requirement that digital images of any notes taken should be uploaded within an hour of the end of a laboratory period. Each group has an associated collaboration page, where experimental data are posted. Access to all of these pages are universal, which allows students to access sources and data from anywhere, provides a space for writing and figure annotation within each group, and allows students to compare results to those of other groups.

4.3 Outcomes

In order to explore the impact of integrating these CURE elements, survey data has been collected, analyzed and examined with each iteration of the course. This allows for monitoring trends over time. One major benefit of collecting this data is that we can use the previous feedback to improve the next iteration, so it never gets stale and always improves, theoretically. For example, instead of using overall lab experience to group, assign students so that each group has someone who has experience with each of the general activities (gel, protein prep, etc). Several scales have been adapted from published instruments to measure the impact of the course on students' experiences with authentic research.

4.3.1. Laboratory Course Assessment Survey results

The LCAS is used to differentiate CUREs from other laboratory courses by measuring students' perceptions of three design features including: 1) collaboration, 2) discovery and relevance, and 3) iteration. The Collaboration subscale addresses the degree to which students are encouraged to work together and provide and respond to feedback. The subscale incorporates a four-point scale that asks respondents how often they were encouraged to engage in collaborative activities (i.e., 1=Never, 2=One or Two Times, 3=Monthly, 4=Weekly). The Iteration subscale describes the degree to which students have opportunities to revise their work to address problems or new questions as well as to improve validity of their own and others' results; this subscale incorporates a six-point Likert scale that runs from Strongly Disagree to Strongly Agree. The Discovery subscale, also using a six-point agreement scale, addresses the extent to which students have opportunities to generate new

Table 4.2. Student responses for Collaboration

Collaboration	Mean	SD
I was encouraged to discuss elements of my investigation with classmates or instructors.	3.64	0.809
I was encouraged to reflect on what I was learning.	3.55	1.036
I was encouraged to contribute my ideas and suggestions during class discussions	4.00	0.000
I was encouraged to help other students collect or analyze data.	3.36	1.206
I was encouraged to provide constructive criticism to classmates and challenge each other's interpretation.	3.36	0.924
I was encouraged to share the problems I encountered during my investigation and seek input on how to address them.	3.82	0.603

knowledge in the discipline. Responses on items across the three scales are aggregated to create overall scores for Collaboration, Discovery and Relevance, and Iteration, and a total LCAS score is comprised of the three combined scores. The possible range of scores for Collaboration is 6-24, for Discovery and Relevance is 5-30, and for Iteration is 6-36, while the overall LCAS score may range from 17 to 90.

Over time, LCAS responses (Table 4.1) have been collected and reviewed to establish how students perceive the "CUREness" of the course (Sathy, et al., under review). The overall mean LCAS score was 72.91 (SD = 8.91). The LCAS scores for this course were similar to those from the original Corwin et al. (2015) study[99], with students scoring Collaboration and Iteration slightly higher and Discovery somewhat lower. Overall, these data suggest we delivered what we intended to in terms of a departure from a traditional cookbook structure, but there are areas where improvements can be made.

The Collaboration survey results (Table 4.2) suggest students felt strongly that they were encouraged to contribute to class discussions (4.00 ± 0.000) and seek input for addressing problems they encountered in the lab (3.84 ± 0.603). These positive responses are likely due to our recitation structure, where student groups were encouraged to draw their own experimental

conclusions and report back to the class, or the low student-to-TA ratio during lab sessions. The lowest scores were in response to assisting classmates in data collection and analysis (3.36 ± 1.206) and providing criticism (3.36 ± 0.924). The latter result is unsurprising, as our current curriculum does not incorporate much peer review, but the low score for assisting others is interesting. Our group structure and the length of the experiments necessitate students to help each other collect data; furthermore, the writing assignments, figure generation, and experimental analysis are all performed as a group. Taking these course elements together, we would expect a stronger response to this question. A possible explanation is that students may have interpreted the question as assisting students from other groups, which we currently do not emphasize. Since Collaboration received the lowest response score of the three LCAS categories, future iterations should incorporate more emphasis on research as a collaborative effort, including inter-group interactions and peer review.

Discovery responses (Table 4.3) were uniformly higher than those for Collaboration, with no mean values falling below 4.00. Students rated the expectations for data explanation (5.09 ± 0.831) and argument development (4.73 ± 0.786) highest, which aligns with our focus on analysis and data interpretation. Students felt expectations were lower for delivering results of interest to the community or unknown to the instructors (4.09 ± 0.944). It is impossible to parse if this score results from the students' lack of perception that their results would be of import to the scientific community or if they assumed the instructors knew the experimental outcomes beforehand, but generally we could aim to emphasize the fact that while we as instructors have some expectation of the results, the exact outcomes are unknown to anyone and would therefore be important to expanding the knowledge base. The low score for student formulation of

Table 4.3. Student responses for Discovery

Discovery	Mean	SD
I was expected to generate novel results that were unknown to the instructor and that could be of interest to the broader scientific community or others outside of class.	4.09	1.136
I was expected to conduct an investigation to find something previously unknown to myself, other students, and the instructor.	4.45	0.934
I was expected to formulate my own research questions or hypothesis to guide an investigation.	4.09	0.944
I was expected to develop new arguments based on data.	4.73	0.786
I was expected to explain how my work has resulted in new knowledge.	5.09	0.831

research projects (4.09 ± 1.136) is to be expected at this stage in the course development. We are still troubleshooting and streamlining the experiments and course structure, including assessing what experimental volume and degree of technical difficulty the students will be able to handle within the scheduling constraints. There will at some point be a shift in design that incorporates student-led hypothesis generation and experimental design, which should improve this metric.

The category with the strongest ratings was Iteration (Table 4.4), which is the most predictable result as the students repeated most of the experiments threefold or more during troubleshooting. This point is most evident from the fact that the highest rating was given to repeating work to address experimental problems (5.27 ± 0.786). A secondary effect of multiple rounds of troubleshooting was that students were tasked with deciding which facets of the experiments to alter; the time spent in this process is reflected by the strong response to the available time to change methods (5.00 ± 1.183). The lowest score was in regard to testing new questions and hypotheses that arose over the course of the experiments (4.18 ± 1.250). This is understandable from the standpoint that the focus was on getting usable results from the prescribed assays before pursuing other avenues; we hope that as we learn more and progress

Table 4.4. Student responses for Iteration

Iteration	Mean	SD
I was expected to revise or repeat work to account for errors or fix problems.	5.27	0.786
I had time to change the methods of the investigation if it was not unfolding as predicted.	5.00	1.183
I had time to share and compare data with other students.	4.73	0.905
I had time to collect and analyze additional data to address new questions or further test hypotheses that arose during the investigation.	4.18	1.250
I had time to revise or repeat analyses based on feedback.	4.82	1.168
I had time to revise drafts of papers or presentations about my investigation based on feedback.	4.73	1.104

through the more difficult initial stages of this curriculum, we can begin to address these shortcomings.

4.3.2. Skill development

The skills development scale incorporates student ratings of their comfort level with various lab techniques. This is assessed in a pre/post manner on a five-point scale. Statistically significant increases were reported for all skills (Table 4.5), the largest growth being in enzyme activity assays (mean difference of 2.27) and working with restriction enzymes (mean difference of 1.78). Perhaps more importantly the standard deviations for each decreased. Together, these results suggest that the experimental design and iterative nature of this CURE not only accomplished the underlying goal of teaching the students an array of widely-applicable biochemical techniques, but also led to more uniform confidence in laboratory abilities among the cohort.

4.4. Limitations, Challenges, and Opportunities for Improvement

Creating a curriculum of this nature involves significant challenges. Instructor experience is of particular import; progressing smoothly through the course and providing an optimal learning experience necessitates familiarity with the methods and associated troubleshooting points. This is especially important when the main point of instructor contact is through graduate teaching assistants. We have been fortunate to have late-career graduate students to serve in a leadership role for younger teaching assistants. When this is not possible, extensive training will be required. It is imperative to have open channels of communication for planning; we scheduled TA meetings immediately following the weekly recitation so the results of both sections could be included in planning each week's experiments. The student-instructor ratio should be considered when deciding the expectations for the semester's progression and the scope of the research project as a whole.

The institutional resources available are also limiting factors when choosing a project to pursue. The structure of the UNC Chemistry departmental curriculum allows us to meet with students for three sessions each week, including two in the laboratory. Other universities may have significantly less face-to-face student with students and reinforcing concepts and techniques may be difficult. At UNC, we have access to departmental cores that allow us access to numerous analytical techniques, such as plate readers, fluorescence imagers, refrigerated cell-culture shakers, etc. This course also takes advantage of university-wide software licensing agreements to implement notebooks, collaboration, and data analysis. Not all institutions can offer similar access, so a project must be chosen such that it is feasible to perform any necessary experiments to produce publishable results.

Table 4.5. Change in self-confidence with laboratory techniques

Technique	Mean (before)	SD (before)	Mean (after)	SD (after)	Mean Difference	p-value	η^2
Enzyme activity assays	2.28	1.06	4.55	0.52	2.27	0.00	0.57
Restriction Enzyme	2.40	1.19	4.18	0.75	1.78	0.00	0.38
Bacterial Transformation	2.68	1.15	4.27	0.65	1.59	0.00	0.35
SDS PAGE	3.04	1.37	4.36	0.67	1.32	0.01	0.21
Protein Purification	3.00	1.32	4.27	0.65	1.27	0.01	0.21
Agarose Gel Electrophoresis	3.48	1.30	4.45	0.69	0.97	0.03	0.14

The dynamic nature of the syllabus means that some semesters will focus more on some techniques than others. As the course shifted from optimization of purification and assay conditions to enzymology results, students spent less time on the process of enzyme production than the previous cohort. Likewise, the initial cohort got more experience in enzyme production and troubleshooting than enzymology data analysis and interpretation. It is pertinent to include some activities that reinforce those techniques that might be stressed less at any point in the overall project, so as not to create a knowledge gap.

Finally, pursuing an unknown outcome means that the "safety nets" available in a traditional curriculum aren't available in the event of experimental failure. This has less of an impact after successive semesters, as samples of purified protein, PCR products, and cell stocks can be stored; however, for initial implementation, this can stall the project. To prevent this, we did some initial test purifications and assays with an undergraduate the summer prior to the first semester. We also embrace the unpredictability as a means of teaching students that in research, the answer is not known beforehand; repetition and troubleshooting may be frustrating, but students quickly become proficient in understanding and performing those techniques.

Class size presents a challenge for drawing broad conclusions about the success of our CURE. While maximum enrollment can accommodate 32 students across two sections in each

semester, our actual enrollment numbers are below this ceiling. Furthermore, recent departmental changes in degree requirements have shifted students toward other tracks, leading to a decline in enrollment numbers. Our present sample size should still allow for meaningful analysis of the entire cohort, but it remains to be seen how this decrease will shift student demographics.

It is also important to note that the student surveys were implemented in conjunction with the CURE curriculum, resulting in a lack of data for students preceding the change. The result is a lack of a clear control group for comparing the initial effects of CURE establishment. Further, another CURE program was established within UNC Chemistry that students in later cohorts would have taken prior matriculation into this course. It is too early to determine how prior experience with a CURE might affect student outcomes, but this could potentially influence the trends in the data. While Corwin et al.[99] serves as a general comparison for our CURE instance, the appropriateness of using survey metrics designed for a biology course to evaluate our biochemistry CURE remains to be determined. The specific activities and structure of a biochemistry course may differ significantly enough that survey wording or emphasis might cause variations in student responses.

4.5. Implications

One of the most important aspects of our CURE is the inclusion of the pre- and post-course surveys. The pre-course survey allows us to distribute laboratory experience between the lab groups, which helps create a more uniform learning environment where group members can learn from each other and TA interaction could be directed towards the most pressing concerns. Tracking the changes in student beliefs following the semester is vital to adjusting the curriculum to address shortfalls that would otherwise go unnoticed through standard course evaluations.

While the LCAS is widely used as a measure of the degree to which a class has attributes associated with CUREs – or it's "CUREness" -- the range of ratings for items on the LCAS scales may be more reflective of how the course is structured and what aspects are emphasized. For example, the high scores in data explanation and argument development most likely reflect the emphasis that our course structure placed on these aspects of the scientific process. With guidance from the instructor, students were asked to assess weekly results in their recitations and prescribe follow-up experiments to address those results, activities that require robust analytical effort. This approach meant less time was allocated to hypothesis generation. Furthermore, as the course is still early in iteration, we chose to be more guiding of the students' research goals as we are still becoming more comfortable operating within the intrinsic uncertainty of the curriculum; further iterations will include more student-led hypothesis generation and testing. The increased student ownership of the project may also improve the students' views on how they are impacting the scientific community.

Survey and anecdotal feedback from the students suggest that one of the biggest challenges was dealing with the seeming ambiguity of an open-ended scientific inquiry. This is the first experience many of the students had with approaching a question in a laboratory setting without a concrete endpoint or outcome and is starkly different compared to the binary correct or incorrect nature of standard knowledge testing. This is a challenge that persists for scientists at every level and is particularly difficult to cope with early in a career. We don't currently have a built-in mechanism to directly address why students feel the way they do about this mercurial aspect of science or discuss the metacognitive or affective aspect. In future iterations, we can do a better job of introducing the concept of succeeding through failure and providing support for students who struggle.

While multiple attempts to complete an experiment are valuable in skill development and provide a realistic expectation of a science career, they may diminish the students' sense of contribution to the scientific landscape. Our results suggest that this is a shortcoming of our current CURE implementation, although the trend could be explained by a lack of emphasis of the project on a big-picture scale or an intrinsic doubt of the veracity of an undergraduate research project in a classroom setting. The latter alternative could be addressed by introducing students to other CURE application in the literature and resulting publications.

Our results suggest that our CURE approach to the biochemistry course is delivering a worthwhile research experience to those students who might otherwise not be exposed to such an opportunity. We have created an environment of collaboration, self-teaching, and rigorous drilling of those skills and concepts necessary for a career in science. This course is still relatively young and will continue to evolve, expand, and balance to supplement those areas indicated through student surveys. Ultimately, we believe that our work can be used as a template for expanding biochemical CUREs to other institutions.

APPENDIX A: USER GUIDE FOR THE *IN VIVO* IMAGE ANALYSIS PIPELINE

A.1 Overview

This pipeline is designed to process single-channel movies in TIFF format. The general workflow for an experiment proceeds as shown in Figure A.1. CellMovieAnalysis and InVivoBatchTraces are compatible for running on the UNC Research Cluster, as they do not require user input. All scripts are compatible with Matlab 2017b or newer, while the applications require 2019a or newer. Current versions for each component are: 1.7 for CellMovieAnalysis, 0.1 for InVitroMovieAnalysis, 1.1 for InVivoBackgroundChooser, 1.6 for FilterSliders, 1.6 for InVivoBatchTraces, and 2.5 for TransitionAnalysisApp. All software files are available from the Erie Lab GitHub repository.

A.2 CellMovieAnalysis Matlab script

1. Set initialization parameters (located at top of main function)
    a. threshold (default = 3) – number of standard deviations above median used for spot thresholding.
    b. spotPixelNumber (default = 4) – number of top-intensity pixels taken from any spot found.
    c. minimumSpotSize (default = 4) – any spots found with less pixels above threshold than this number are discarded. NOTE: If the program is not in “box method” mode and this is less than spotPixelNumber, the program defaults this

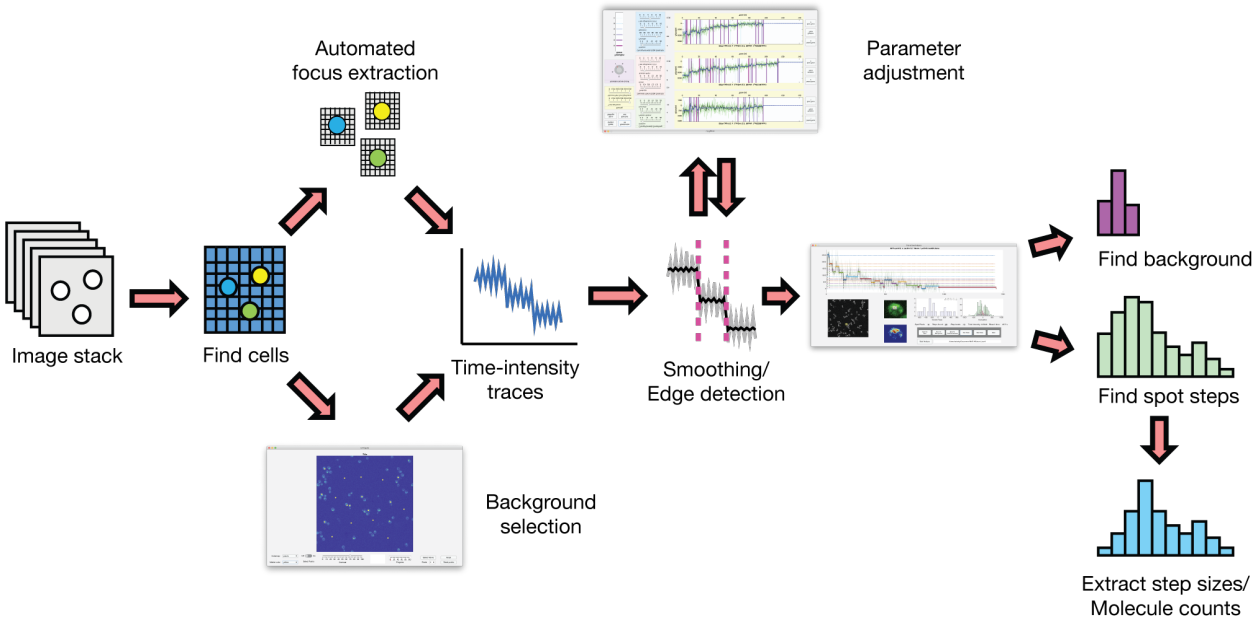

Figure A.1. Software pipeline workflow. Analysis begins by collecting a TIFF image stack. If the movie contains cells, the cells are found and those with spots are extracted; if the movie is *in vitro* data, the spots are found. The background application can be used to generate background traces. Traces are then extracted, smoothed, and edges are detected. Parameters for smoothing and edge-detection can be refined using the sliders application. Analyzed traces are then analyzed for step sizes using the step-finding application. Finally, output histograms are analyzed and background subtracted.

value to spotPixelNumber to avoid throwing errors (i.e. not having enough pixels to take).

d. spotBoxMethodOn (default = true) – when on, a 3-by-3-pixel box is drawn around the maximum-intensity pixel of the spot and the number of pixels designated by spotPixelNumber are taken. If this is turned off, all pixels in the spot are ordered by intensity and the designated number of pixels are taken.

2. Set up the directory

   a. The script queries the folder in which it resides for any TIFF files. This allows the user to place all movies into a single folder and the script will cycle through them all.

b. All other TIFF images should be removed from the folder to avoid throwing errors.

3. Call the script
    a. This can be done in an interactive Matlab workspace or via a bash script for job submission on the Research Computing cluster. Information on job submission can be found on the UNC Information Technology Services website.
4. Output files
    a. The script creates a text log ("BatchLog_" followed by the date) beginning with the timestamp of initialization, followed by a list of those files analyzed with basic information (file size, movie size, number of cells found, and the number that contained foci), and finally a timestamp for completion.
    b. A subdirectory is created for each movie analyzed. This subdirectory contains:
        i. Time-intensity traces (.traces files) for each cell with a spot – these are binary files with the same structure as FRET analysis
            1. Movie length (32-bit integer)
            2. Number of traces (16-bit integer)
            3. Trace data (16-bit integer) – these data are in the following order, repeating for each frame:
                a. Frame number
                b. "Acceptor" intensity – this is fake acceptor data created by inverting the intensity relative to the maximum intensity for the entire raw ("donor") trace. This nomenclature is maintained to reduce the amount of change to the FRET

batch trace analysis script. Fake acceptor data allows for better transition detection later.

c. "Donor" intensity – background-corrected intensity from the movie frame. The median background intensity from the pixels around the spot is calculated for the frame and subtracted from each spot pixel's intensity. The corrected intensities are then summed.

d. If there are multiple foci in a single cell, the fake acceptor and donor values repeat in this manner for each additional focus

i. Image maps in PNG format (for viewing)

1. Region map – the cell numbering map, in "parula" colormap. This allows the user to see which cells were identified and those that were removed.

2. Threshold map – for all cells with spots, the intensity values are converted to multiples of that cell's standard deviation above its median. This is scaled from 0 to the set threshold value (see initialization parameters, above) and presented in "jet" colormap. This allows the user to view the context of the spot pixels in terms of how close surrounding pixels were to the threshold.

3. Spot map – all pixels chosen for background superimposed on the composite image (contrast adjusted). The spot pixels appear empty so the composite intensity can be seen easily. The colormap for the

background pixels is not indicative of value other than to show contrast between different and overlapping background regions.

ii. Image maps in binary format (.erie files) – matrix values, no colormaps. These are used downstream by the TransitionAnalysis app

1. Region map – see above, unsigned 16-bit integers
2. Threshold map – see above, single-precision values
3. Spots – binary mask of pixels selected as spots, unsigned 16-bit integers
4. Backgrounds – binary mask of pixels selected as background, unsigned 16-bit integers
5. Composite – composite image, single-precision values

A.3 InVitroMovieAnalysis Matlab script

2. Set initialization parameters (top of main function)
    a. threshold (default = 5) – number of standard deviations above median used for spot thresholding.
    b. spotPixelNumber (default = 4) – number of top-intensity pixels taken from any spot found.
3. Call script in interactive Matlab session
4. Select directory
    a. The script will prompt the user to select a directory
    b. Any TIFF files in the selected directory will be processed, so limit files to *in vitro* movies, lest the outputs be unpredictable or errors be thrown

5. Output files
    a. The script creates a text log (“BatchLog_” followed by the date) beginning with the timestamp of initialization, followed by a list of those files analyzed with basic information (file size, movie size, number of cells found, and the number that contained foci), and finally a timestamp for completion.
    b. A subdirectory is created for each movie analyzed. This subdirectory contains:
        i. Time-intensity traces (.traces files) for all spots found – these are binary files with the same structure as FRET analysis
            1. Movie length (32-bit integer)
            2. Number of traces (16-bit integer)
            3. Trace data (16-bit integer) – these data are in the following order, repeating for each frame:
                a. Frame number
                b. “Acceptor” intensity – this is fake acceptor data created by inverting the intensity relative to the maximum intensity for the entire raw (“donor”) trace. This nomenclature is maintained to reduce the amount of change to the FRET batch trace analysis script. Fake acceptor data allows for better transition detection later.
                c. “Donor” intensity – background-corrected intensity from the movie frame. The median background intensity from the pixels around the spot is calculated for the frame and

subtracted from each spot pixel's intensity. The corrected intensities are then summed.

A.4 InVivoBackgroundChooser Matlab application

1. Open the application. This can be done outside of or within a Matlab session. Opening from the directory panel in Matlab will open App Designer, at which point the user can select "Run" at the top (indicated by a green "play" icon).
2. The UI window will open (Figure A.2).
3. Press the "Select Movie" button to open the file selection prompt (limited to TIFF image stacks). After selecting a movie, it may take some time to load the composite image.
    a. The loaded composite image is processed in the same manner as the CellMovieAnalysis script (see section A.2)
4. To select points for background traces, switch the "Select Points" toggle to "on." While Select Points is active, clicking on the composite image will place solid circle markers at the chosen coordinates.
    a. NOTE: To use the window functions available by hovering the mouse over the image (zoom, pan, etc.), toggle Select Points off while manipulating the image to avoid recording the clicks. While zoomed in, point-selection can be resumed, as the image coordinates scale.
5. Optional: The contrast of the composite image can be adjusted using the slider underneath the image. Use the dropdown menus on the far-left to change the colormap of

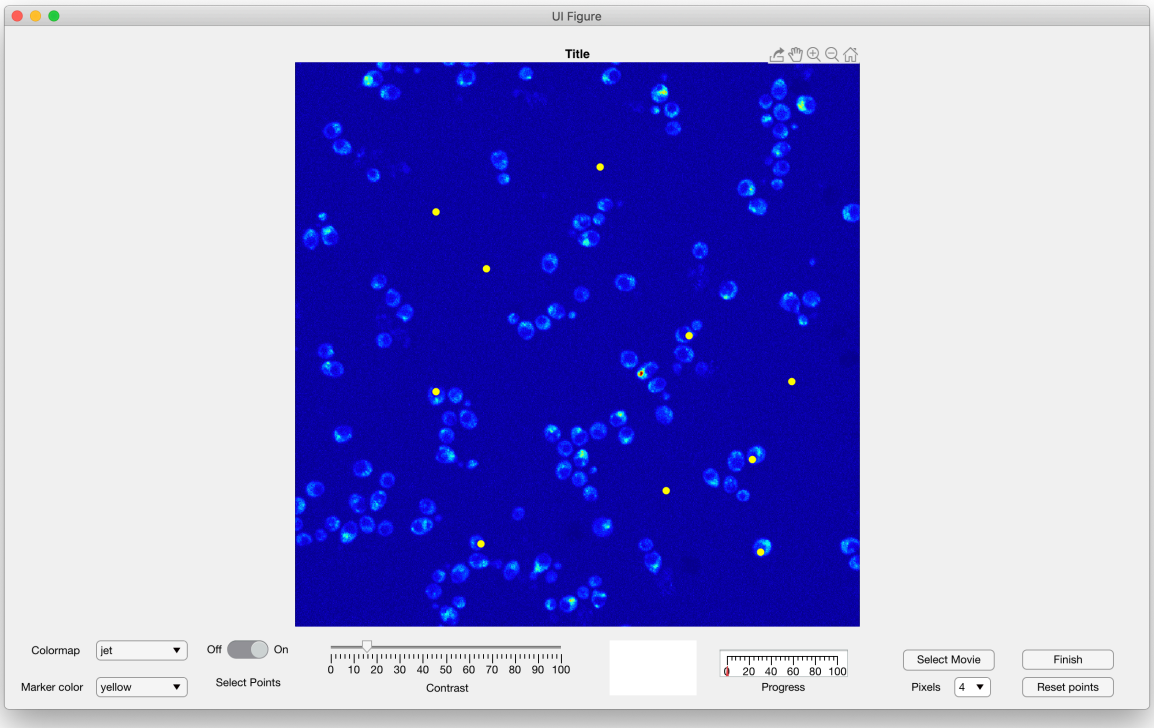

Figure A.2. InVivoBackgroundChooser application interface.

the image and the color of the click markers. The "Reset Points" button will remove all selected points. At the time of writing this, there is not yet an option to remove individual points.

6. Once desired points are selected, adjust the number of pixels taken for analysis using the "Pixels" dropdown menu.
    a. This value is equivalent to the spotPixelNumber variable from the CellMovieAnalysis script.
7. Press the "Finish" button to begin analysis.
    a. Progress can be monitored through the text box and progress gauge underneath the image. The text box displays the current step and the progress gauge shows the percent completion of this step.

b. The analysis is performed using the "box method" (see CellMovieAnalysis, section A.2) around each pixel that was clicked.
c. Background is taken as a two-pixel border around the 3-by-3 box.
d. If the selected pixel is near the edge of a cell, the algorithm will determine which region (e.g. cell or slide) that comprises the majority of the combined background/spot area and only included pixels from that region.
e. Time-intensity traces (.traces files) are output as previously described, with "_Background" included in the title.
f. No map files are output

A.5 FilterSliders Matlab application

1. Open the application. This can be done outside of or within a Matlab session. Opening from the directory panel in Matlab will open App Designer, at which point the user can select "Run" at the top (indicated by a green "play" icon).
2. The UI window will open (Figure A.3).
3. Select time-intensity traces to display. This can be accomplished in several ways:
    a. Pressing the "Randomize All" button in the upper-left corner of the interface will prompt the user to select a directory containing ".traces" files and select three random time-intensity traces from the dataset. These will not include duplicates, but can include multiple traces from the same file.
    b. Pressing the "Random Trace" button next to a display window will prompt the user to select a directory as described above, but this will only select a single

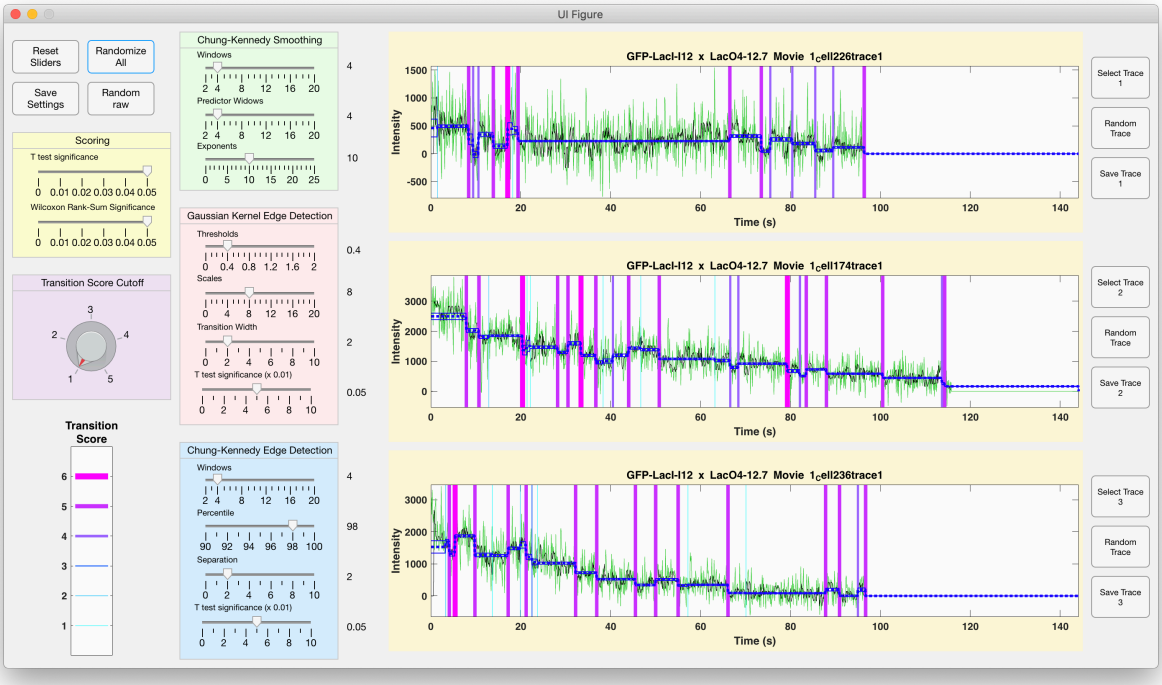

Figure A.3. FilterSliders application interface.

random trace to be display in the respective display window. This option is useful for viewing traces from different experiments held in separate directories.

c. Pressing the “Select Trace” button next to a display window allows the user to manually select a “.traces” file. If the file contains multiple time-intensity traces, a pop-up window will display the number of traces and ask which trace to load.

d. NOTE: The software will work even if not all windows are displaying a trace.

4. Raw trace data is displayed in green, while the smoothed data is black. The regional means appear as a dashed blue line, while the bounds of the 95% confidence interval are thinner, solid blue lines above and below the mean. Transitions are displayed as vertical lines of increasing thickness for increasing score, the key for which is found in the lower-left corner of the user interface.

5. Adjust the smoothing and edge-detection parameters using the sliders. Sliders will automatically adjust to the valid value nearest that selected and the value is displayed in a text box immediately to the right of the slider. The trace(s) will automatically update upon any change in parameter values.
    a. NOTE: The p-value for the edge-detection t-tests differ from the "Scoring" p-values. Changing the value of the edge-detection p-values will reduce the number of edges found prior to scoring. The scoring p-values do not change the edges found by the detection methods, but instead change the threshold for allotting the score points for those statistical tests (the point for detection by each method is still accrued).
    b. NOTE: Changing the "Transition Score Cutoff" removes those transitions scored below the threshold. After applying the cutoff, transitions may remain with scores lower than this due to the fact that removal of transitions will change the results for calculating regional means, confidence intervals, and t-tests, all of which are used for scoring. This score removal is only performed once, otherwise it might continue iterating until all most or all transitions are removed.
6. Optional: Results of smoothing and edge-detection on each display trace can be saved using the "Save Trace" button next to the respective display window.
7. To save the current parameter setting, press the "Save Settings" button in the upper-left of the user interface.
    a. The settings are saved as a text file – titled "BatchTraceConfig.txt" – that can be read into the InVivoBatchTraces script (see below).

A.6 InVivoBatchTraces Matlab script

1. NOTE: This script is mostly the same as the FRET version, with a few exceptions. It is currently set to use only donor data, although it should work with donor and acceptor (since we create fake acceptor data during trace generation). The same scoring system used in the FilterSliders application has been implemented here as well. The output file has been modified to remove output of acceptor data and to include the scoring metrics.
2. Set up directory
    a. The file handling has been changed to query the current directory and all subdirectories for ".traces" files. This allows the user to move whole folders of traces instead of individual files, but it is important to note that any output files will be overridden if the data is reanalyzed.
    b. This file system was designed to be used with the Research Computing cluster. The output from the CellMovieAnalysis script will result in the directory containing subdirectories for each movie with the respective trace files. Placing this script in the same directory as the CellMovieAnalysis script will allow all resulting traces to be analyzed.
3. Set up initialization parameters
    a. The filter parameters are no longer tied to a pop-up window. The script will search the current directory for a configuration file titled "BatchTraceConfig.txt" that is output by the FilterSliders application. This file can be edited manually if necessary.

b. If no configuration file is located, the script will run with the default values:
    i. Chung-Kennedy Smoothing
        1. Window size: 4
        2. Predictor window size: 4
        3. Exponents: 10
    ii. Gaussian Kernel Edge-Detection
        1. Threshold: 0.4
        2. Scales: 8
        3. Transition width: 2
        4. t-Test p-value: 0.05
    iii. Chung-Kennedy Edge-Detection
        1. Window size: 4
        2. Percentile: 98
        3. Window separation: 2
        4. t-Test p-value: 0.05
    iv. Scoring
        1. Wilcoxon Rank-Sum p-value: 0.05
        2. t-Test p-value: 0.05
        3. Score Cutoff: 1

4. Call the script
    a. The script can be called in an interactive Matlab session or via a bash script on the Research Computing cluster. Information on job submission can be found on the UNC Information Technology Services website.

5. Output files
    a. The output files are text files with the ".trans" extension, which allows for file searching by the TransitionAnalysis application.

A.7 TransitionAnalysisApp – Matlab application

1. NOTE: There are several variables with hard-coded values. These are the first 5 variables listed in the *editable* "Properties" section of the code (i.e. not the section that is grayed out). To access the code, open the application in the directory window of a Matlab session to open the App Designer window and select the "Code View" tab in the upper-right corner of the middle panel.
    a. timestep – framerate of the movie (default = 0.1, equivalent to 100 ms)
    b. bleachThreshold – intensity below which we consider complete photobleaching (default = 400 AU)
    c. imageHeight – y-dimension of the image (default = 1024). NOTE: Changing to non-square images will require more editing.
    d. colNum – anonymous function for finding column number from linear indexing (uses 1024 for the image height)
    e. rowNum – anonymous function for finding row number from linear indexing (uses 1024 for the image height)
2. Open the application. This can be done outside of or within a Matlab session. Opening from the directory panel in Matlab will open App Designer, at which point the user can select "Run" at the top (indicated by a green "play" icon).
3. The UI window will open (Figure A.4).

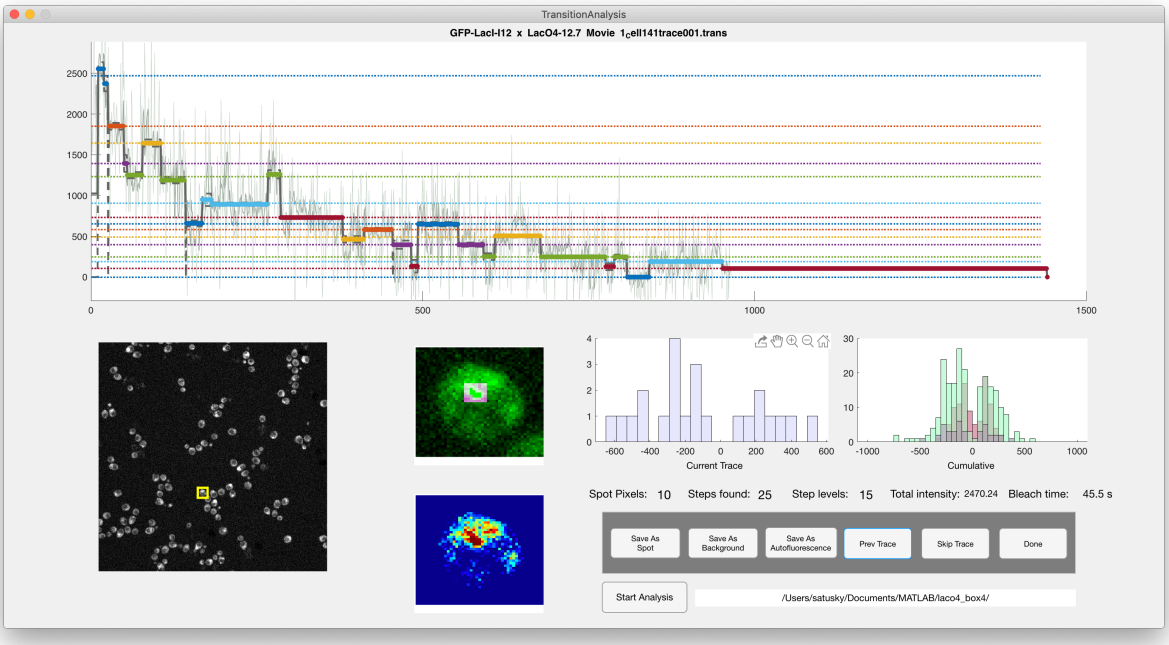

Figure A.4. TransitionAnalysis application interface.

4. Press the “Start Analysis” button at the center-bottom of the interface to open the directory-selection prompt.
5. A pop-up window will open so the user can name the output file. If a file with the same name already exists, the user must confirm the override or rename the output file.
6. Data Analysis
    a. Trace Display Window
        i. Trace data is displayed in the prominent window. Raw data is displayed in light gray and smoothed data is in darker gray.
        ii. Step levels are arbitrarily colored for contrast purposes and displayed as a dashed line at the y-position of the weighted mean. While the colors are

not representative of the value of the levels, the regions contributing to each level are highlighted in the matching color.

iii. The SEM for each region is represented by a dashed black line above and below the region.

b. Cell maps

i. These maps are not displayed for background or *in vitro* traces, or if there are no ".erie" map files in the selected directory.

ii. The contrast-adjusted composite image is displayed in the bottom left. The current cell is highlighted by a yellow box.

iii. A cropped map of the cell intensities is shown in green, with a magenta box highlighting the pixels chosen for background around the current spot.

iv. The cropped threshold map for the current cell is displayed in the "jet" colormap for added context.

c. Histograms

i. The step sizes for the current trace are displayed in the left histogram window. The x-axis is set to automatically scale, so the scale will change depending on the step profile of the trace.

ii. The cumulative step sizes for all traces analyzed are displayed in the right histogram window. This histogram has a fixed x-axis scale from -1000 to 1000. Up to three histograms will be displayed, depending on how the user has classified the traces: spot (green), background (red), and autofluorescence (purple).

d. Trace Stats – displayed as text below the histograms

i. Spot Pixels – number of pixels in the spot above threshold
ii. Steps Found – number of steps identified
iii. Step Levels – number of step levels identified
iv. Total Intensity – intensity of the highest step level
v. Bleach Time – time at which the intensity drops below the threshold considered complete bleaching

e. User Interface
    i. Binning trace by type – press the "Save As" button for the corresponding data type to bin as spot, background, or autofluorescence.
    ii. To return to the previous trace, press the "Prev Trace" button. If the previous trace data was binned, the data will be removed from the cumulative histogram. If the trace was skipped, no data is removed.
    iii. To skip a trace without saving the data, press the "Skip Trace" button.
    iv. The "Save All as Single Histogram" button provides an alternative means of saving the data. When pressed, all traces in the directory are binned automatically as spots. This method is ideal for *in vitro* experiments that contain a high volume of traces.
    v. To remove all analyzed data and select a new file at any time during analysis, press the "Start Analysis" button. The user will be prompted to confirm this choice.

7. Save the analyzed data by pressing the "Done" button. A prompt will appear asking the user if they would like blinks to be removed. Selecting either option results in display of the histograms and the option to proceed or revert.

a. Blink Removal

    i. Blinks are deemed to have occurred when a down-step is immediately followed by a return to the same level.

    ii. If the user selects blink removal, the down- and up-step from each blink is removed and all histograms are displayed with the change from blink removal overlaid.

    iii. The user then has the option to continue with blink removal or to save the dataset without removing known blinks.

8. Output files

    a. Step levels and step heights are written to a two-column, tab-delimited text file with the suffix "_Steps" for easy import into statistical software.

    b. A more descriptive tab-delimited text file contains the details of the dataset on a trace-by-trace basis.

        i. TraceName – filename of the trace

        ii. SpotType – 0 (spot), 1 (background), or 2 (autofluorescence)

        iii. NumPixels – number of pixels in the spot above threshold

        iv. MaxIntensity – intensity of the highest step level

        v. BleachFrame – frame at which the intensity drops below the threshold considered complete bleaching

        vi. NumLevels – number of step levels identified

        vii. NumSteps – number of steps identified

        viii. StepLevels – the values that follow are the weighted mean intensities for the step levels in descending order

ix. StepHeights – the values that follow are the intensity sizes of steps in chronological order

REFERENCES

1. Kunkel, T. A. & Erie, D. A. DNA MISMATCH REPAIR. Annu. Rev. Biochem. 74, 681–710 (2005).

2. Iyer, R. R., Pluciennik, A., Burdett, V. & Modrich, P. L. DNA Mismatch Repair: Functions and Mechanisms. Chem. Rev. 106, 302–323 (2006).

3. Fedier, A. & Fink, D. Mutations in DNA mismatch repair genes: implications for DNA damage signaling and drug sensitivity (review). Int. J. Oncol. 24, 1039–1047 (2004).

4. McKinnon, P. J. & Caldecott, K. W. DNA Strand Break Repair and Human Genetic Disease. Annu. Rev. Genomics Hum. Genet. 8, 37–55 (2007).

5. Loeb, L. A., Loeb, K. R. & Anderson, J. P. Multiple mutations and cancer. Proc. Natl. Acad. Sci. U. S. A. 100, 776–781 (2003).

6. Sass, L. E., Lanyi, C., Weninger, K. & Erie, D. A. Single-Molecule FRET TACKLE Reveals Highly Dynamic Mismatched DNA−MutS Complexes. Biochemistry 49, 3174–3190 (2010).

7. LeBlanc, S. J. et al. Coordinated protein and DNA conformational changes govern mismatch repair initiation by MutS. Nucleic Acids Res. (2018) doi:10.1093/nar/gky865.

8. Monico, C., Capitanio, M., Belcastro, G., Vanzi, F. & Pavone, F. S. Optical Methods to Study Protein-DNA Interactions in Vitro and in Living Cells at the Single-Molecule Level. Int. J. Mol. Sci. 14, 3961–3992 (2013).

9. Axelrod, D., Thompson, N. L. & Burghardt, T. P. Total internal inflection fluorescent microscopy. J. Microsc. 129, 19–28 (1983).

10. Axelrod, D. [1] Total internal reflection fluorescence microscopy in cell biology. in Methods in Enzymology vol. 361 1–33 (Academic Press, 2003).

11. Combs, C. A. Fluorescence Microscopy: A Concise Guide to Current Imaging Methods. Curr. Protoc. Neurosci. 50, (2010).

12. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. Nat. Methods N. Y. 5, 507–16 (2008).

13. Qiu, R. et al. Large conformational changes in MutS during DNA scanning, mismatch recognition and repair signaling. EMBO J. 38, (2019).

14. Erie, D. A. & Weninger, K. R. Single molecule Studies of DNA Mismatch Repair. DNA Repair 20, 71–81 (2014).

15. Lakowicz, J. R. Principles of fluorescence spectroscopy. (Springer, 2006).

16. Ranjit, S., Gurunathan, K. & Levitus, M. Photophysics of Backbone Fluorescent DNA Modifications: Reducing Uncertainties in FRET. J. Phys. Chem. B 113, 7861–7866 (2009).

17. Snapp, E. Design and Use of Fluorescent Fusion Proteins in Cell Biology. Curr. Protoc. Cell Biol. Editor. Board Juan Bonifacino Al CHAPTER, Unit-21.4 (2005).

18. Specht, E. A., Braselmann, E. & Palmer, A. E. A Critical and Comparative Review of Fluorescent Tools for Live-Cell Imaging. Annu. Rev. Physiol. 79, 93–117 (2017).

19. Stephens, D. J. & Allan, V. J. Light Microscopy Techniques for Live Cell Imaging. Science 300, 82–86 (2003).

20. Chudakov, D. M., Matz, M. V., Lukyanov, S. & Lukyanov, K. A. Fluorescent Proteins and Their Applications in Imaging Living Cells and Tissues. Physiol. Rev. 90, 1103–1163 (2010).

21. Keppler, A. et al. A general method for the covalent labeling of fusion proteins with small molecules in vivo. Nat. Biotechnol. N. Y. 21, 86–9 (2003).

22. Keppler, A., Pick, H., Arrivoli, C., Vogel, H. & Johnsson, K. Labeling of fusion proteins with synthetic fluorophores in live cells. Proc. Natl. Acad. Sci. U. S. A. 101, 9955–9959 (2004).

23. Soh, N. Selective Chemical Labeling of Proteins with Small Fluorescent Molecules Based on Metal-Chelation Methodology. Sensors 8, 1004–1024 (2008).

24. Pantoja, R., Rodriguez, E. A., Dibas, M. I., Dougherty, D. A. & Lester, H. A. Single-Molecule Imaging of a Fluorescent Unnatural Amino Acid Incorporated Into Nicotinic Receptors. Biophys. J. 96, 226–237 (2009).

25. Zhang, M., Li, M., Zhang, W., Han, Y. & Zhang, Y.-H. Simple and efficient delivery of cell-impermeable organic fluorescent probes into live cells for live-cell superresolution imaging. Light Sci. Appl. 8, 1–11 (2019).

26. Giepmans, B. N. G., Adams, S. R., Ellisman, M. H. & Tsien, R. Y. The Fluorescent Toolbox for Assessing Protein Location and Function. Science 312, 217–224 (2006).

27. Dougherty, D. A. Unnatural amino acids as probes of protein structure and function. Curr. Opin. Chem. Biol. 4, 645–652 (2000).

28. Narancic, T., Almahboub, S. A. & O'Connor, K. E. Unnatural amino acids: production and biotechnological potential. World J. Microbiol. Biotechnol. 35, 67 (2019).

29. Joglekar, A. P., Bouck, D. C., Molk, J. N., Bloom, K. S. & Salmon, E. D. Molecular architecture of a kinetochore-microtubule attachment site. Nat. Cell Biol. Lond. 8, 581–5 (2006).

30. Coffman, V. C., Wu, P., Parthun, M. R. & Wu, J.-Q. CENP-A exceeds microtubule attachment sites in centromere clusters of both budding and fission yeast. J. Cell Biol. 195, 563–572 (2011).

31. Liesche, C. et al. Automated Analysis of Single-Molecule Photobleaching Data by Statistical Modeling of Spot Populations. Biophys. J. 109, 2352–2362 (2015).

32. Weninger, K., Bowen, M. E., Choi, U. B., Chu, S. & Brunger, A. T. Accessory Proteins Stabilize the Acceptor Complex for Synaptobrevin, the 1:1 Syntaxin/SNAP-25 Complex. Structure 16, 308–320 (2008).

33. Ulbrich, M. H. & Isacoff, E. Y. Subunit counting in membrane-bound proteins. Nat. Methods 4, 319–321 (2007).

34. Long, F., Zeng, S. & Huang, Z.-L. Localization-based super-resolution microscopy with an sCMOS camera Part II: Experimental methodology for comparing sCMOS with EMCCD cameras. Opt. Express 20, 17741–17759 (2012).

35. Chung, S. H. & Kennedy, R. A. Forward-backward non-linear filtering technique for extracting small biological signals from noise. J. Neurosci. Methods 40, 71–86 (1991).

36. Tsekouras, K., Custer, T. C., Jashnsaz, H., Walter, N. G. & Pressé, S. A novel method to accurately locate and count large numbers of steps by photobleaching. Mol. Biol. Cell 27, 3601–3615 (2016).

37. Chen, Y., Deffenbaugh, N. C., Anderson, C. T. & Hancock, W. O. Molecular counting by photobleaching in protein complexes with many subunits: best practices and application to the cellulose synthesis complex. Mol. Biol. Cell 25, 3630–3642 (2014).

38. McKinney, S. A., Joo, C. & Ha, T. Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling. Biophys. J. 91, 1941–1951 (2006).

39. Messina, T. C., Kim, H., Giurleo, J. T. & Talaga, D. S. Hidden Markov model analysis of multichromophore photobleaching. J. Phys. Chem. B 110, 16366–16376 (2006).

40. Carter, B. C., Vershinin, M. & Gross, S. P. A Comparison of Step-Detection Methods: How Well Can You Do? Biophys. J. 94, 306–319 (2008).

41. Leake, M. C. et al. Stoichiometry and turnover in single, functioning membrane protein complexes. Nature 443, 355–358 (2006).

42. Zhang, H. & Guo, P. Single molecule photobleaching (SMPB) technology for counting of RNA, DNA, protein and other molecules in nanoparticles and biological complexes by TIRF instrumentation. Methods 67, 169–176 (2014).

43. Dobrucki, J. W., Feret, D. & Noatynska, A. Scattering of Exciting Light by Live Cells in Fluorescence Confocal Imaging: Phototoxic Effects and Relevance for FRAP Studies. Biophys. J. 93, 1778–1786 (2007).

44. Gru$\upbeta$mayer, K. S., Yserentant, K. & D.-P. Herten. Photons in - numbers out: perspectives in quantitative fluorescence microscopy for in situ protein counting. Methods Appl. Fluoresc. 7, 012003 (2019).

45. Kalashnikov, M. et al. Assessing light scattering of intracellular organelles in single intact living cells. Opt. Express 17, 19674–19681 (2009).

46. Maslanka, R., Kwolek-Mirek, M. & Zadrag-Tecza, R. Autofluorescence of yeast Saccharomyces cerevisiae cells caused by glucose metabolism products and its methodological implications. J. Microbiol. Methods 146, 55–60 (2018).

47. Verdaasdonk, J. S., Lawrimore, J. & Bloom, K. Determining absolute protein numbers by quantitative fluorescence microscopy. in Methods in Cell Biology vol. 123 347–365 (Elsevier, 2014).

48. Coffman, V. C. & Wu, J.-Q. Counting protein molecules using quantitative fluorescence microscopy. Trends Biochem. Sci. 37, 499–506 (2012).

49. Haase, J. et al. A 3D Map of the Yeast Kinetochore Reveals the Presence of Core and Accessory Centromere-Specific Histone. Curr. Biol. 23, 1939–1944 (2013).

50. Lawrimore, J., Bloom, K. S. & Salmon, E. D. Point centromeres contain more than a single centromere-specific Cse4 (CENP-A) nucleosome. J. Cell Biol. 195, 573–582 (2011).

51. McGuire, H., Aurousseau, M. R. P., Bowie, D. & Blunck, R. Automating Single Subunit Counting of Membrane Proteins in Mammalian Cells. J. Biol. Chem. 287, 35912–35921 (2012).

52. Das, S. K., Darshi, M., Cheley, S., Wallace, M. I. & Bayley, H. Membrane Protein Stoichiometry Determined from the Step-Wise Photobleaching of Dye-Labelled Subunits. ChemBioChem 8, 994–999 (2007).

53. Coffman, V. C. & Wu, J.-Q. Every laboratory with a fluorescence microscope should consider counting molecules. Mol. Biol. Cell 25, 1545–1548 (2014).

54. Bradford, K. C. Structure-Function Studies of the Initiation Response of Human Mismatch Repair Proteins to DNA Containing a Mismatch. (The University of North Carolina at Chapel Hill, 2015).

55. Du, M., Kodner, S. & Bai, L. Enhancement of LacI binding in vivo. Nucleic Acids Res. 47, 9609–9618 (2019).

56. Arias-Castro, E. & Donoho, D. L. Does median filtering truly preserve edges better than linear filtering? Ann. Stat. 37, 1172–1206 (2009).

57. Reyes-Lamothe, R., Sherratt, D. J. & Leake, M. C. Stoichiometry and Architecture of Active DNA Replication Machinery in Escherichia coli. Science 328, 498–501 (2010).

58. Watkins, L. P. & Yang, H. Detection of Intensity Change Points in Time-Resolved Single-Molecule Measurements. J. Phys. Chem. B 109, 617–628 (2005).

59. Eddy, S. R. What is a hidden Markov model? Nat. Biotechnol. N. Y. 22, 1315–6 (2004).

60. Kalafut, B. & Visscher, K. An objective, model-independent method for detection of non-uniform steps in noisy signals. Comput. Phys. Commun. 179, 716–723 (2008).

61. Carter, N. J. & Cross, R. A. Mechanics of the kinesin step. Nat. Lond. 435, 308–12 (2005).

62. Kerssemakers, J. W. J. et al. Assembly dynamics of microtubules at molecular resolution. Nature 442, 709–712 (2006).

63. Canny, J. A Computational Approach to Edge Detection. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-8, 679–698 (1986).

64. Arant, R. J. & Ulbrich, M. H. Deciphering the Subunit Composition of Multimeric Proteins by Counting Photobleaching Steps. ChemPhysChem 15, 600–605 (2014).

65. Gauer, J. W. Single-molecule fluorescence studies of DNA bending during prokaryotic mismatch repair initiation. (The University of North Carolina at Chapel Hill, 2016).

66. McCann, J. J., Choi, U. B., Zheng, L., Weninger, K. & Bowen, M. E. Optimizing methods to recover absolute FRET efficiency from immobilized single molecules. Biophys. J. 99, 961–970 (2010).

67. Dickerson, R. DNA bending: the prevalence of kinkiness and the virtues of normality. Nucleic Acids Res. 26, 1906–1926 (1998).

68. Heyduk, T. & Lee, J. C. Application of fluorescence energy transfer and polarization to monitor Escherichia coli cAMP receptor protein and lac promoter interaction. Proc. Natl. Acad. Sci. 87, 1744–1748 (1990).

69. Hwang, H. & Myong, S. Protein induced fluorescence enhancement (PIFE) for probing protein–nucleic acid interactions. Chem Soc Rev 43, 1221–1229 (2014).

70. Genschel, J. et al. Interaction of proliferating cell nuclear antigen with PMS2 is required for MutLα activation and function in mismatch repair. Proc. Natl. Acad. Sci. 114, 4930–4935 (2017).

71. Eckstein, F. Nucleoside Phosphorothioates. Annu. Rev. Biochem. 54, 367–402 (1985).

72. Lee, J. H. et al. Site-Specific Control of Distances between Gold Nanoparticles Using Phosphorothioate Anchors on DNA and a Short Bifunctional Molecular Fastener. Angew. Chem. Int. Ed. 46, 9006–9010 (2007).

73. Lee, J. H., Wong, N. Y., Tan, L. H., Wang, Z. & Lu, Y. Controlled Alignment of Multiple Proteins and Nanoparticles with Nanometer Resolution via Backbone-Modified Phosphorothioate DNA and Bifunctional Linkers. J. Am. Chem. Soc. 132, 8906–8908 (2010).

74. Ludueña, R. F., Roach, M. C., Trcka, P. P. & Weintraub, S. N,N-Bis(a-iodoacetyl)-2,2'-dithiobis(ethylamine), a Reversible Crosslinking Reagent for Protein Sulfhydryl Groups. Anal. Biochem. 117, 76–80 (1981).

75. Frederiksen, J. K. & Piccirilli, J. A. Chapter 14 - Separation of RNA Phosphorothioate Oligonucleotides by HPLC. in Methods in Enzymology vol. 468 289–309 (Academic Press, 2009).

76. Cromie, G. A. et al. Single Holliday Junctions Are Intermediates of Meiotic Recombination. Cell 127, 1167–1178 (2006).

77. West, S. C. Molecular views of recombination proteins and their control. Nat. Rev. Mol. Cell Biol. 4, 435–445 (2003).

78. Bellendir, S. P. et al. Substrate preference of Gen endonucleases highlights the importance of branched structures as DNA damage repair intermediates. Nucleic Acids Res. 45, 5333–5348 (2017).

79. Zhou, R. et al. Junction resolving enzymes use multivalency to keep the Holliday junction dynamic. Nat. Chem. Biol. 15, 269–275 (2019).

80. Weninger, K., Bowen, M. E., Chu, S. & Brunger, A. T. Single-molecule studies of SNARE complex assembly reveal parallel and antiparallel configurations. Proc. Natl. Acad. Sci. U. S. A. 100, 14800–14805 (2003).

81. Bell, J. K. et al. CUREs in biochemistry-where we are and where we should go: CUREs in Biochemistry. Biochem. Mol. Biol. Educ. 45, 7–12 (2017).

82. Domin, D. S. A Review of Laboratory Instruction Styles. J. Chem. Educ. 76, 543 (1999).

83. Buck, L. B., Bretz, S. L. & Towns, M. H. Characterizing the Level of Inquiry in the Undergraduate Laboratory. J. Coll. Sci. Teach. Wash. 38, 52–58 (2008).

84. Vision and change in undergraduate biology education: a call to action : final report of a national conference. (American Association for the Advancement of Science, 2011).

85. Gates, S. J. & Mirkin, C. Engage to Excel. Science 335, 1545–1545 (2012).

86. Weaver, G. C., Russell, C. B. & Wink, D. J. Inquiry-based and research-based laboratory pedagogies in undergraduate science. Nat. Chem. Biol. 4, 577–580 (2008).

87. Russell, S. H., Hancock, M. P. & McCullough, J. Benefits of Undergraduate Research Experiences. Science 316, 548–549 (2007).

88. Estrada, M., Hernandez, P. R. & Schultz, P. W. A Longitudinal Study of How Quality Mentorship and Research Experience Integrate Underrepresented Minorities into STEM Careers. CBE—Life Sci. Educ. 17, ar9 (2018).

89. Auchincloss, L. C. et al. Assessment of Course-Based Undergraduate Research Experiences: A Meeting Report. CBE—Life Sci. Educ. 13, 29–40 (2014).

90. Gregerman, S. R., Lerner, J. S., Hippel, W. von, Jonides, J. & Nagda, B. A. Undergraduate Student-Faculty Research Partnerships Affect Student Retention. Rev. High. Educ. 22, 55–72 (1998).

91. Seymour, E., Hunter, A.-B., Laursen, S. L. & DeAntoni, T. Establishing the benefits of research experiences for undergraduates in the sciences: First findings from a three-year study. Sci. Educ. 88, 493–534 (2004).

92. Russell, C. B. & Weaver, G. C. A comparative study of traditional, inquiry-based, and research-based laboratory curricula: impacts on understanding of the nature of science. Chem Educ Res Pr. 12, 57–67 (2011).

93. Brownell, S. E. & Kloser, M. J. Toward a conceptual framework for measuring the effectiveness of course-based undergraduate research experiences in undergraduate biology. Stud. High. Educ. 40, 525–544 (2015).

94. Lopatto, D. Survey of Undergraduate Research Experiences (SURE): First Findings. Cell Biol. Educ. 3, 270–277 (2004).

95. Cole, Darnell. Do Interracial Interactions Matter? An Examination of Student-Faculty Contact and Intellectual Self-Concept. J. High. Educ. 78, 249–281 (2007).

96. Hurtado, S., Cabrera, N. L., Lin, M. H., Arellano, L. & Espinosa, L. L. Diversifying Science: Underrepresented Student Experiences in Structured Research Programs. Res. High. Educ. 50, 189–214 (2009).

97. Chang, M. J., Eagan, M. K., Lin, M. H. & Hurtado, S. Considering the Impact of Racial Stigmas and Science Identity: Persistence Among Biomedical and Behavioral Science Aspirants. J. High. Educ. 82, 564–596 (2011).

98. Kloser, M. J., Brownell, S. E., Chiariello, N. R. & Fukami, T. Integrating Teaching and Research in Undergraduate Biology Laboratory Education. PLoS Biol. 9, e1001174 (2011).

99. Corwin, L. A., Graham, M. J. & Dolan, E. L. Modeling Course-Based Undergraduate Research Experiences: An Agenda for Future Research and Evaluation. CBE—Life Sci. Educ. 14, es1 (2015).