THE DEVELOPMENT AND IMPLEMENTATION OF MICROSCOPY STRATGIES FOR INVESTIGATING PROTEIN DIFFUSION AND CHROMATIN BINDING

Michael August Tycon

A dissertation submitted to the faculty of 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
2013

Approved by:

Christopher Fecko

Nancy Allbritton

Dorothy Erie

Linda Spremulli

Mark Wightman

ABSTRACT

Michael August Tycon:
The Development and Implementation of Microscopy Strategies for
Investigating Protein Diffusion and Chromatin Binding
(Supervised by Dr. Christopher J. Fecko)

Nearly all cellular processes, notably transcription, translation, and genomic repair, are enacted by multiprotein complexes that coalesce into functional assemblies in response to constantly fluctuating cellular demands. A complex interplay of endogenous and exogenous cellular cues regulates the assembly and activity of these complexes by both active and passive mechanisms, with a current fundamental dilemma in the field of molecular biology being the elucidation of the mechanisms governing the assembly of these supramolecular complexes. Such complexes arise through two processes, the nucleation of macromolecular assemblies and target binding site recognition. Collectively, this phenomenon is anthropomorphized as"protein recruitment", yet this term conceals the underlying physical interactions that govern the spatiotemporal formation of such assemblies, turning protein activity into a series of "black boxes" with prescribed functions. In response to this overarching question, microscopy technologies were tailored to investigate the mechanisms of these two inextricable facets of protein recruitment. Thus, during my tenure in the Fecko Laboratory, I have been concerned with the big picture while simultaneously looking at the very small.

Methods were developed enabling the observation of model systems of complex recruitment dynamics and have been used to illustrate paradigms of biological function. An initial effort was focused on designing optical systems for observation of DNA repair protein diffusion. The ability to generate user-defined DNA photolesions in real time, a highly characterized binding site of many classes of DNA repair proteins, creates opportunities for optical imaging experiments in which protein behavior before and after a biological perturbation can be observed. To this end a two-photon DNA damage method was developed, which enabled the production of UV-type DNA photolesions by blue light and is highly compatible with conventional laser-scanning optical microscopy configurations. This visible light damage method was compared to alternative damage induction processes, and the advantages of the two-photon method enumerated.

Continuing towards an integrated system for observing protein diffusion, a popular single-molecule imaging DNA immobilization and visualization technique was characterized. In this work, the extent of optically-induced DNA binding site artifacts was established with a unique pairing of a widefield microscopy based single-molecule and gel electrophoresis based ensemble biochemical DNA damage assays. The results indicated that many commonly used DNA visualization practices, from imaging parameters through fluorescent intercalaters, lead to extensive photodamage and can perturb native DNA-protein interactions.

Later work shifted away from single molecule investigations and towards studying the diffusion dynamics of large macromolecular complexes *in vivo*. A unique two-photon FRAP microscopy and image processing technique was developed and used to characterize the

diffusion of RNA Polymerase II subunits in live cell nuclei. The findings substantiate a hybrid model of macromolecular assembly in which a broad distribution of macromolecular species allow for mechanistic flexibility in the assembly of transcription complexes. This provides evidence for further speculation on mechanisms controlling gene expression.

.

To my parents, friends, and teachers
To Uncle Marty, the first person to teach me that a doctor is not necessarily an MD
To Miss Belle for giving me a reason to stick around
To my departed Firebird which brought me from Canada to North Carolina
&
To my dear Ford Truck which will take me away.

ACKNOWLEDGMENTS

"If you try and take a cat apart to see how it works, the first thing you have on your hands is a non-working cat"

-Douglas Adams

The modern world of highly interdisciplinary scientific research cannot occur alone in an intellectual vacuum, much to the chagrin of those that prefer to work holed up in a dark room. Rather, it requires intellectual and practical collaboration at every step, in a mutualistic relationship in which all should benefit from the synergy of talents. In the same vein, this truth applies to the education that molds such researchers, in which learning is a partnership between students and teachers, both friends and professors. To this end, there are many people I would like to thank whom either contributed to my intellectual development, directly assisted me in my research, or helped keep me sane throughout my time in the chemistry department of the University of North Carolina.

Foremost, I would like to thank my advisor, Dr. Christopher J. Fecko. We collectively gambled on each other and I like to think we both learned a great deal along the way. Whether through the sheer amount of time we have spent together or the efficiency of his teaching over the past five years, he has helped to shape my understanding of the scientific method and made me into a better (more skeptical) scientist. No one that spends any amount of time with Chris cannot come to appreciate his critical reasoning and the extent of his academic insight- I hope some of this has rubbed off on me.

Under the direction of our fearless leader, my lab group has often been a rag-tag collection of misfits, and I would like to individually acknowledge them- Lori Dorward (now Nichols), Ian McNeil, Matthew Daddysman, and my undergraduate Catherine Dial, for their assistance over the past years. Mr. Daddysman deserves special recognition for his patience as I struggled through MATLAB help files (which I swear are written in ancient Greek). He is a truly encyclopedic resource- from ornithology through NASCAR trivia, and always willing to lend an ear. Additionally, Miss Dial deserves commendation for her labors, dealing with an office full of guys and being too naive to understand that working on Sunday evening was not a reasonable request to make of a summer intern.

No mention of my graduate school career would be complete without a nod to the usual incoming class of analytical students in 2008. We were and still are a neurotic, OCD-laden, and resourceful bunch. They said we were too slow for graduate school but too dumb to quit, and that has been largely accurate. In particular, my best-wishes go out to Natalie Bjorge and Joe Gateri. My first three years at UNC would have been a dull and sober place were it not for the antics of Miss Bjorge. As for Mr. Gateri- with how many people can you have an entire conversation in movie quotations?

Foremost among the class of UNC Chemistry 2008, my affections and many thanks to the lovely Miss Anna M. Belle. She is a wonderfully neurotic and quirky girl,quick to carry other people in times of distress while slow to ask for help herself. You have been a constant source of support throughout the mess that has been the past five years. Thanks for the early morning goodnight calls and remember to sit back and relax sometimes. My tailgate is always down for a drink.

The past five years would be much less memorable were it not for Pasha Takmakov, Paul Walsh, Richard Kiethley, and Scott Nichols. Thank you all for the intellectual and social contributions to my well-being. Additionally, I would like to thank Holly Wolcott and Punya Navaratnarajah for assistance in troubleshooting difficult experimental techniques at great expense to their own time.

I did not get to graduate school through my handwork alone and would like to thank the many people who helped before I arrived at UNC. For me, my pursuit of scientific knowledge has always been accompanied by friends. My McGill crew closed the library down on many an occasion, and Amir Amiri and Jordan Wilson were always there besides me on the late nights. Further, my thanks go out to both Professor Eric Salin and Professor David Burns. Collectively, you both believe in a meritocracy, recognized my abilities, and gave me my first chance at research.

Anytime I recollect on those that have helped shape my appreciation for science and mold my work ethic, my thoughts immediately conjure up my high school biology and chemistry teachers- Mrs. Mary Jane Roethlin and Mrs. Morturano. I am likely still trying to impress you both and never have been able to pick a favorite subject.

I would like to thank my sister, Laura, for needling me during my graduate school career and goading me to the finish line. Finally, my gratitude and love to my parents, Joseph and Mary Tycon. You both always seemed to effortlessly cultivate a spirit of creativity and intellectualism in our house. You both endured my countless questions and my unreasonable demands. While you may have stopped proof-reading my papers years ago, I'll always be happy to discuss them with you both.

Table of Contents

LIST OF TABLES ..... **xiii**

LIST OF FIGURES ..... **xiv**

LIST OF ABBREVIATIONS ..... **xvii**

CHAPTER 1

INTRODUCTION

What is Protein Recruitment and how do We Study it? **1**

Differing Systems to Study Protein Recruitment **1**

Strategies to Investigate Protein-Target Binding Site Recognition **3**

1. Protypical DNA Damage Repair Pathway 3

2. Optically Manipulating DNA 4

3. Designing a Platform for the In Vitro Study of Repair Pathways 6

4. Sensitized Methods of Photochemical DNA Damage Induction 7

Strategies for Investigating the Assembly of Macromolecular Complexes **10**

1. Models of Macromolecular Protein Assembly Dynamics in Cell Nuclei 10

2. Mechanism Evaluation: Choosing the Right Time and Place 12

Research Aims and Scope **14**

References **15**

CHAPTER 2

GENERATION OF DNA PHOTOLESIONS BY TWO-PHOTON ABSORPTION OF A FREQUENCY-DOUBLED TI:SAPPHIRE LASER

Overview: **19**

Introduction **20**

Materials and Methods **22**

1. Materials 22

2. DNA sample preparation 23

3. QPCR assay of DNA damage 24

4. UV irradiation 25
5. Femtosecond laser irradiation 25
Results and Discussion 28
1. Development of a QPCR assay of DNA damage 29
a. Conditions for quantitative PCR 30
b. Statistical treatment of randomly distributed DNA photolesions 33
2. UV-induced DNA photodamage 35
3. Two-photon absorption-induced DNA photodamage 38
Conclusion 47
References 48

CHAPTER 3

QUANTIFICATION OF DYE-MEDIATED PHOTODAMAGE DURING SINGLE-MOLECULE DNA IMAGING

Overview: 51
Introduction 52
Materials & Methods 54
1.Observing double-strand photocleavage using flow-stretched DNA 54
a. Surface functionalization, microfluidic chamber fabrication, and DNA substrate preparation 54
b. DNA staining and injection for SMI 56
c. Single-molecule imaging 56
d. Radical scavenger buffer preparation 57
2.Single-molecule image processing 57
3.Ensemble DNA damage assay 58
a. Bulk DNA sample preparation 58
b. Bulk sample irradiation 59
c. Gel electrophoresis 59
4.Ensemble damage assay: Ascorbic acid mediated DNA degradation 59
5.Gel quantification 60
Results 60
1. Double-strand photocleavage of individual DNA molecules 60
2. Ensemble study of single and double-strand photocleavage 65
3. Kinetic modeling of DNA strand cleavage 68

a. Modeling for DNA cleavage 68
b. Fitting ensemble data 74
c. Fitting double-strand photocleavage of flow-stretched DNA 77
d. Effect of scavengers 79
e. Extrapolation between SMI conditions and ensemble studies 80
4. Degradation of DNA by ascorbic acid 80
Discussion 83
Conclusion 87
References 89

CHAPTER 4

RNA POLYMERASE II SUBUNITS EXHIBIT A BROAD DISTRIBUTION OF MACROMOLECULAR ASSEMBLY STATES IN THE INTERCHROMATIN SPACE OF CELL NUCLEI

Overview: 91
Introduction 92
Materials and Methods 98
1.Fly Strains 98
2.Salivary Gland Extract Preparation 98
3.Two-photon microscopy configuration and FRAP Procedures 99
Results 100
1.Automated "shotgun ptFRAP" data collection 100
2.Different recovery dynamics observed for RNAPII subunits 101
3.Confirming the distribution of heterogeneous RNAPII subunit complexation states 108
4.Distribution modeling: decomposing apparently anomalous recovery curves into components exhibiting Brownian diffusion 110
Discussion 117
1.A new perspective for in vivo diffusion: apparent anomalous diffusion 117
2.RNAPII distributions indicate an intermediate assembly mechanism 122
Conclusion 126
References 128

CHAPTER 5

DETERMINING THE UNDERLYING DISTRIBUTIONS OF MULTIPLE SIMULTANEOUS DIFFUSING SPECIES FROM FRAP SIMULATIONS

Overview: **131**

Introduction **131**

Computations **133**

1.The Distribution Model: 133

2.Distributions of Diffusing Species: 134

3.Incomplete FRAP Recovery Simulations 135

Results and Discussion **138**

1.Accuracy of Predicting a Binary Mixture 144

2.Accuracy of Predicting a Biologically Relevant Distribution 144

3.Accuracy of Predicting a Binary Mixture with an Artificial Immobile Fraction 145

4.Accuracy of Predicting a Gamma Distribution with an Artificial Immobile Fraction 146

Conclusions **150**

References **151**

APPENDIX A: QUANTIFICATION OF GEL ELECTROPHORESIS DATA USING FOUR GAUSSIAN PEAKS TO OVERCOME BACKGROUND HETEROGENEITIES **152**

APPENDIX B: AUTOMATED QUANTIFICATION OF DNA MOLECULE STRAND CLEAVAGE **172**

APPENDIX C: AUTOMATED "SHOTGUN PTFRAP" IMAGING PROCESSING PROGRAMS **185**

APPENDIX D: SUPPORTING INFORMATION FOR THE CHAPTER 4 ..... **189**

1.High expression levels of fusion proteins are not responsible for the observed anomalous diffusion 189

2.Determining the resolution of the Point FRAP method 192

3.Establishing the robustness of the Distribution model on experimental data 195

4.Slow Diffusion Components under the FRAP resolution method are not required for an accurate fit 198

5.FRAP fitting results for each dataset 200

6.FRAP diffusion fitting results for individuals datasets and ensemble averages **201**

LIST OF TABLES

Table 3.1- Characteristic parameters describing the single strand breakage rates by imaging condition 78

Table D.1-FRAP diffusion fitting results for individuals datasets and ensemble averages 201

LIST OF FIGURES

Figure 2.1-Schematic diagram of the irradiation apparatus. 27

Figure 2.2-Detection of DNA photolesions using quantitative PCR. 32

Figure 2.3- Poisson statistics are required to determine the number of DNA lesions from the quantitative PCR assay 34

Figure 2.4- UV dose dependent lesion formation. 37

Figure 2.5- Power-dependent damage produced by irradiation of DNA samples with focused femtosecond pulses at 425 nm, 450 nm and 475 nm. 42

Figure 3.1- SMI strand cleavage assay and damage quantification for flow-stretched, YOYO-stained lambda DNA at a dye to nucleotide ratio of 1:4. 63

Figure 3.2- Ensemble breakage assay and damage quantification. 67

Figure 3.3- Stochastic DNA damage model and fitting of the ensemble data. 70

Figure 3.4- Comparison of the single-strand breakage rates (n) obtained by fitting results of the ensemble (A) or the SMI (B) damage assays to the stochastic DNA damage model. 76

Figure 3.5- Extrapolation of the intensity-dependent SMI single-strand breakage rates to the laser intensity used for the ensemble measurements. 82

Figure 3.6- Ascorbic acid mediated DNA Damage. 84

Figure 4.1- Image Collection and Automated Processing Methodology “Shotgun ptFRAP” 95

Figure 4.2- Comparison of in vivo subunit recovery dynamics. 102

Figure 4.3- Summary of the best-fit apparent anomalous modeling parameters. 106

Figure 4.4- Comparison of in vitro subunit recovery dynamics. 109

Figure 4.5- Brownian diffusion coefficient distributions. 112

Figure 5.1: Extracting a binary mixture from a simulated FRAP curve at different SNR. 136

Figure 5.2: Extracting a gamma distribution from a simulated FRAP curve at different SNR 139

Figure 5.3: Inclusion of an artificial immobile fraction impairs fitting by the distribution model on datasets with 50 dB SNR. 140

Figure 5.4: Inclusion of an artificial immobile fraction impairs fitting by the distribution model on datasets with 35 dB SNR. 141

Figure 5.5: Results of extracting the underlying distribution from a gamma function input with the inclusion of an artificial immobile fraction at 50 dB SNR 142

Figure 5.6: Results of extracting the underlying distribution from a gamma function input with the inclusion of an artificial immobile fraction at 35 dB SNR. 143

Figure 5.7: Effect of including an artificial immobile fraction on distribution fitting to a binary mixture without noise 148

Figure 5.8: Effect of including an artificial immobile fraction on distribution fitting to a gamma distribution without noise 149

Figure A.1- Output of gel_analysis2, indicating the region of interest for each lane. 160

Figure A.2- Representative output of a single lane analysis. 164

Figure A.3- Output of the Gaussian fits to each DNA band for every lane. 170

Figure A.4- Quantification of the three plasmid forms from the initial gel image. 171

Figure B.1- Compiling datafiles into an image stack 175

Figure B.2- Initial frame of a time-lapse movie recording the cleavage of elongated DNA molecules. 176

Figure B.3- False-color output used to guide user selection of intact DNA molecules. 177

Figure B.4- DNAid1 output enabling user selection. 178

Figure B.5- Final output and resulting mask. 179

Figure B.6- Output of the DNAidstack2 program. 183

Figure D.1- High expression levels of fusion proteins are not responsible for the observed anomalous diffusion 191

Figure D.2- Determining the Resolution of the Point FRAP Method 194

Figure D.3- Establishing the Robustness of the Distribution Model on Experimental Data..... 197
Figure D.4- Fit quality excluding diffusion components under FRAP resolution..... 199

LIST OF ABBREVIATIONS

AA- Ascorbic Acid

AFM- Atomic Force Micrscopy

APTES- (3-Aminopropyl)triethoxysilane

b- Critical distance in basepairs

B(t)- Nicked or singly broken DNA molecules

BBO- β Barium Borate

BME- β-mercaptoethanol

bp- Basepairs

BSA- Bovine Serum Albumin

CALI- Chromophore Assisted Laser Inactivation

CI- Confidence Interval

CPD- Cyclopyrimidine Dimer

D_{eff}- Effective Diffusion Coefficent

DNA- Deoxyribonucleic Acid

dNTP- Deoxynucleotide

DSB- Double Strand Break

EDTA- Ethylenediaminetetraacetic acid

EMCCD- Electron Multiplied Charge Coupled Device

eV- Electron Volt

FCS- Fluorescence Correlation Spectroscopy

FLIP- Fluorescence Loss in Photobleaching

FRAP- Fluorescence Recovery after Photobleaching

FWHM- Full Width at Half Maximum

GFP- Green Fluorescent Protein

GM- Goeppert Mayer

gp- Temporal laser pulse shape

HPLC- High Performance Liquid Chromatography

Hz- Hertz

KI- Potassium Iodide

KIO_3- Potassium Iodate

$Mg(OAc)_2$- Magnesium acetate

mM- millimeter

NA- Numerical Aperture

NA_2- number of photons absorbed per nucleotide per second

NaCl- Sodium Chloride

$NaHCO_3$- Sodium Bicarbonate

NER- Nucleotide Excision Repair

OD- Optical Density

P(n)- Probability that a DNA strand has n lesions

PCR- Polymerase Chain Reaction

PEG- Polyethylene Glycol

PSF-Point Spread Function

ptFRAP- Point Fluorescence Recovery after Photobleaching

QPCR- Quantitative Polymerase Chain Reaction

QY-Φ_D, Quantum Yield of dimerization

RFP- Red Fluorescent Protein

RNAP- Ribonucleic Acid Polymerase

ROS- Radical Oxygen Species

RSD- Relative Standard Deviation

SD- Standard Deviation

SMI- Single Molecule Imaging

SSB- Single Strand Break

TE- Tris EDTA

TIR- Total Internal Reflection

TIRFM- Total Internal Reflection Fluorescence Microcopy

TPA- Two-Photon Absorption

Tris- tris(hydroxymethyl)aminomethane

U(t)- Undamaged DNA molecules

UV- Ultraviolet

v/v- volume per volume

XFP- Fluorescent Protein

Y-Linearized DNA molecules

μ- Average number of lesions on each DNA strand

σ- Two-photon cross section

τp-Pulse duration

ω- Beam diameter

CHAPTER 1

INTRODUCTION
WHAT IS PROTEIN RECRUITMENT AND HOW DO WE STUDY IT?

"You can observe a lot just by watching"

-Yogi Berra

Differing Systems to Study Protein Recruitment

The cellular interior is a crowded environment, containing a high density of dissolved biological solids and bearing little resemblance to typical *in vitro* reconstitutions[4].Through this viscous and obstacle laden matrix, proteins must migrate the cytoplasmic and nuclear environs, interact with binding partners, and recognize target binding sites. Protein recruitment is the broad term used to describe this process in which multiple binding partners assemble in the cellular environment to conduct a particular metabolic function. While the specifics such as interaction order, location of nucleation, and sub-assembly intermediates will have inevitable differences depending on the specific metabolic function under consideration[5], two elements are constant- assembly and target site recognition of macromolecular complexes. Details of each of these processes are marked by uncertainty; even the interplay of these processes is often not well understood.The questions behind protein assembly concern the timing, duration, and location of the interaction events that lead to the formation of an active complex. Distinct but complimentary, target site recognition chiefly concerns the molecular mechanisms by

which an active complex, either partially or fully assembled, locate a unique binding site, often a miniscule genomic element in comparison to the entire nuclear material[6]. For protein complexes involved in genome metabolism, it has recently been shown that a sharp delineation between these processes is not possible (Chapter 4).

Underlying all aspects of protein recruitment are the transport mechanisms, active or passive, by which proteins traverse the cellular interior[7, 8]. It is through interrogating these transport mechanisms and identifying their signaturesthat we can hope to gain insights into the mechanistic details of recruitment. Given the dynamic nature of protein transport and simultaneous requirements of capturing spatial and temporal details of the processes, optical microscopy has emerged at the forefront of tools uniquely suited for such investigations. In addition to passive imaging techniques that enable high resolution visual observations, powerful perturbation methods and spectroscopies such as Fluorescence Recovery after Photobleaching (FRAP), Fluorescence Loss in Photobleaching (FLIP), and Fluorescence Correlation Spectroscopy (FCS), have evolved allowing *in vivo* measurements of transport dynamics[9, 10]. Further, recent instrumentation advances have opened up the field of single molecule imaging (SMI); giving experimenters the ability to track and manipulate individual biomolecules in both artificially enhancedbiological and synthetic*in vitro* systems[11, 12].

The following research will initially focus on a unique pairing of optical and physical DNA manipulation techniques, joined together in creating a flexible*in vitro*SMIplatform with the possibility of interrogating the mechanisms of DNA-protein binding site recognition in a DNA repair context. Novel techniques to damage DNA in a user controlled and quantitative manner

are discussed, along with important implications for evaluating the results of many optical imaging experiments. Later, variants onhigh time resolution FRAP methods will be discussed and applied to the investigation of the spatiotemporal formation of large protein complexes in the context of DNA transcription. Given the possible mechanistic universality of the underlying chemical and physical interactions of protein recruitment, two highly conserved pathways will be considered. Initially, the most ubiquitous DNA repair pathway, Nucleotide Excision Repair (NER)[13], is used as a model system to drive the development of the optical platform to study protein recruitment *in vitro*. Next, arguably one of the most crucial genome metabolic processes, transcription by RNA Polymerase II[14], will be considered as a paradigm of *in vivo* supramolecular assembly.

Strategies to Investigate Protein-Target Binding Site Recognition

1. Protypical DNA Damage Repair Pathway

The chemical stability of DNA and simplistic elegance of its replication often obscures the myriad ways in which damage can be incurred, through the action of endogenous cellular factors (typically radical oxygen species) or exogenous mutagenic agents, particularly ultra-violet (UV) or ionizing radiation[15].These agents can cause structural changes as significant as strand breaks or dimer formation between adjacent bases. These various forms of damage, collectively termed mutations, lead to loss of genomic fidelity and resulting disease states.

In response to these general chemical and structural insults, complex biochemical pathways evolved to address these damaging effects. Three major classes of DNA repair have been thus far identified, each uniquely suited to correct a particular type of damage. All three

classes are highly conserved in both prokaryotes and eukaryotes[13], underscoring the common mechanistic universality. The least specific repair pathway, nucleotide excision repair (NER), is responsible for correcting damage that results in structural alterations to DNA[13], operating through excision of oligonucleotides flanking the damage site.

Common to all three pathways of DNA repair is the concerted action of multiprotein complexes which must be sequentially recruited to the site of damage amidst the vast majority of highly dynamic chromatin[16, 17].While several models of NER action have been proposed, most feature 3-dimensional, diffusion mediated nuclear transport to enable rapid surveillance of the nuclear volume coupled with occasional 1-dimensional sliding diffusion along the DNA backbone. NER is best understood in the model system *Escherichia coli*, where the Uvr A, B, and C endonuclease system demonstrate concerted action to identify and remove damage sites. Thus DNA repair pathways offer an excellent opportunity to observe site-specific protein recruitment. Once coupled with strategies to induce DNA damage in real-time that initiate the recruitment process, the entire process can be tracked.

2. Optically Manipulating DNA

The most commonly considered DNA damage, or lesion,targeted by NER repair systems arepyrimidine dimers (termed cylcobutane pyrimidine dimers, or CPDs) formed upon exposure to UV radiation. Such lesions occur in the presence of approximately 260 nm light due to the large DNA extinction coefficient at this wavelength and are the result of relaxation of $\pi \rightarrow \pi^*$ transitions of neighboring thymine bases[18].As a critical first step in studying the localization of repair proteins to the damage sites they bind, it is necessary to develop a method to enable the

real-time generation of CPDs with high spatial resolution. These lesions function as user controlled binding sites, triggering the switch from scanning to binding of damaged DNA. While photolesions are usually formed by exposure to 260 nm emission from UV light sources, this results in a random spatial distribution of lesions throughout the sample[19].Advances using polycarbonate masks with 3-5 μm holes to restrict UV exposure have reduced the 2D regions of lesion formation to smaller than a cell nucleus[20, 21]. However, such spatial control is still very poor in comparison to the resolution offered by modern microscopic techniques and worse still in comparison to the biological length scales needed to discern differences in diffusion modality. Further, no spatial control is possible in the third dimension. Since the poor transmission of light below 350 nm restricts the pairing of UV light sources with a microscopy-based apparatus, two-photonirradiation has been harnessed as a means to deliver UV energy with conventional optics.

Two photon absorption (TPA) induced DNA damage has the advantage of generating photolesions in a three-dimensionally pre-defined region of space using visible light and is therefore compatible with standard microscopy optics. Nonresonant multiphoton absorption is the process in which two or more photons interact with a molecule simultaneously (within 10^{-18} s) to generate an excited state equivalent in energy to the summation of the absorbed photons[22]. Thus, instead of requiring UVC photons to initiate photophysical DNA damage, the same photoreactions can be triggered by the multiphoton absorption of visible light[23-27]. Since the probability of TPA depends quadratically on the intensity of the incident light, a large photon flux is required for simultaneous absorption, usually limited to the focal waist of an objective lens [28-31]. This property is exploited to achieve sub-micron depth discrimination in

two-photon microscopy and photodamage production. Depth discrimination is then paired with equatorial control provided by the raster scanning of laser-scanning microscopy, allowing for the precise irradiation of microscale spatial volumes.

3. Designing a Platform for the In Vitro Study of Repair Pathways

In contrast to previous decades in which traditional biochemical techniques were employed to study bulk systems [13, 32], researchers now prefer SMI methods that offer the spatiotemporal resolution required to decipher protein dynamics on a biologically relevant timescale and to observe biological variability in nanoscopic systems. The implementation of single molecule detection is primarily based on the application of optical fluorescence microscopy due to the high contrast acquired by the use of bright fluorophores against dark backgrounds, even in biologically relevant aqueous environments. All such implementations require reducing the sample size under investigation to a sub-100 fl volumes[33]. Currently, the principle techniques to restrict the sample volume are total-internal reflection fluorescence microscopy (TIRFM) and laser scanning methods such as confocal or multiphoton microscopy. In the former, the sample volume investigated is limited by the effective field of illumination created by the very shallow evanescent field that results from reflection off an interface causing total-internal reflection (TIR) [34].

To achieve sample immobilization and provide a restricted imaging volume, schemes for the immobilization of DNA molecules tethered to a glass substrate and elongated by hydrodynamic flow have been independently developed by several groups [11, 35, 36]. Thisrestricts DNA molecules near the surface of a microscope-slide based flow cell while supporting the

molecules above a biologically inert surface[11, 37]. This provides a method to couple multiphotonphotolesion formation with a TIRF based imaging apparatus.

The direct imaging of fluorescently labeled Uvr protein components engaged in a search complex[35, 38], pre and post lesion induction, provides the most direct means to ascertain the mechanism by which target search occurs.To this end,pairingSMI methods with a novel, TPA real-time induction of protein recruitmentwould further elucidate the intricacies of NER in the highly characterized biological systems system.

4. Sensitized Methods of Photochemical DNA Damage Induction

Modern high-resolution optical microscopy is premised upon the use of the fluorescent marker species for the identification and tracking of intracellular or purified biological components. In the case of biological tissue imaging, fluorophores can be endogenously expressed XFP variants or exogenously incorporated molecules, either actively or passively uptaken from the environment. Markers have been engineered that are specific for cellular substructures, targeting incorporation into lipophilic domains for membrane studies or that exhibit high binding affinities to DNA to mark nuclear locations or track genomic processes. Further extending the utility of microscopy to probe highly dynamic biological processes, high-quantum efficiency fluorophores coupled with advancements in optical image collection have resulted in the bourgeoning field of single molecule microscopy for both *in vitro* and *in vivo*applications. Given the high signal-to-noise requirements of such experiments, these studies have led to the use of increasingly high optical intensities compared to conventional widefield imaging.

Paramount among the assumptions made in the use of fluorescent reporter molecules is that they do not perturb the system under observation. Unfortunately, this assumption is not always valid. The optical excitation of light-emitting molecules (fluorophores) often results in photodamage arising from chemical reactions of the fluorophore in its lowest energy electronic excited state, leading to photochemical damage. The most probable pathway for energy relaxation from this excited state is photon emission, but there exist other possible excitation-relaxation pathways that can produce reactive intermediates. These pathways can lead to fluorophore photobleaching, a permanent chemical rearrangement of the fluorophore where fluorescence is no longer the primary relaxation pathway. Most fluorophores undergo 10^5 – 10^6 excitation cycles before photobleaching; entry into this non-emissive state may indicate the production of reactive species [39, 40]. The production of these damaging species may be cryptically occurring even without a visible loss of fluorescence from the sample. In either case, photochemical damage is typically cumulative as it relies upon the net number of excitation events only and not the rate at which the excitation events occur.

Excited fluorophores can occasionally interact with their solvent environment creating short-lived, damaging radical species capable of destabilizing or destroying neighboring biomolecules. The process begins when molecular fluorophores are promoted to a singlet excited state by visible light. One mode for the energetic relaxation of these species is to emit a photon; however, the high cycling rate induced by high light intensities used in confocal or MPM increases the population of triplet state species (the triplet state quantum yield can be as high as 5% for some molecular fluorophores). Molecular oxygen, which exists in a triplet ground state configuration, can readily interact with this excited state fluorophore. Energy

transfer between molecular oxygen and the excited fluorophore results in the formation of singlet oxygenand electron transfer between the two speciescreates a super-oxide and a fluorophore radical. All of these species, termed radical oxygen species (ROS) are highly reactive and are generated by the favorable downhill energetics of electron transfer to ground state oxygen, coupled with the rapid diffusion of molecular oxygen and therefore frequent interactions [40]. These highly unstable species are rapidly quenched in aqueous environments leading to the formation of hydroxyl radicals. The short-lived hydroxyl radical is the prime damage mediating species, resulting in radical induced damage to proximal biomolecules [41].

ROS are frequently generated when imaging nucleic acids stained with intercalating dyes, in both *in vitro* and *in vivo* applications. This can lead to widespread genomic damage, the effect of which must be carefully considered when using DNA stains [42]. The formation of damaging hydroxyl radicals proximal to the site of fluorophore incorporation results in species that can attack DNA to produce various forms of oxidative radical photodamage [43], notably single strand breaks [44, 45]. Individual damage events typically cleave only one strand of the DNA sugar-phosphate backbone [46, 47]; the accumulation of many single-strand breaks leads to double-strand cleavage [18]. Since many proteins involved in DNA replication and repair bind to single-stranded DNA [6, 15, 48], the presence of single strand breaks induced by photo-excitation of intercalating dyes could strongly bias protein-DNA interactions. Additionally, wide-spread genomic damage can induce apoptotic pathways resulting in cell death. This is likely to induce artifacts in experiments probing native biological function.

Although the generation of damage mediating radicals is detrimental for most experiments, it can offer a degree of spatiotemporal user control in instances when initiating cellular damage is desirable [42, 45]. The common DNA intercalating dyes used for imaging application, such as Hoechst and DAPI (*in vivo* use) or YOYO-1, TOTO-1, Picogreen, and related dye monomers (*in vitro* staining), are all capable of selectively targeting DNA for fragmentation[49]. The incorporation of these intercalating dyes enables DNA fragmentation to be initiated at particular wavelengths in a dose-dependent manner. This is useful for studies of DNA damage and repair mechanisms, where localized photochemical damage can be used to elucidate repair pathways. It has been shown that careful selection of the type of dye and DNA binding mode can be applied to tune the DNA backbone cleavage, biasing damage towards double strand cleavage or single strand breaks [50].

Strategies for Investigating the Assembly of Macromolecular Complexes

1. Models of Macromolecular Protein Assembly Dynamics in Cell Nuclei

The second facet of protein recruitment that we have targeted for investigation concerns the spatiotemporal formation of the macromolecular complexes responsible for most cellular processes, in particular genome metabolism. The post-processing of nascent RNA transcripts by the spliceosome and transcription of DNA by RNA Polymerase II (RNAP II) represent the epitome of supramolecular complexes essential for genome metabolism[5], in which function is well resolved but assembly is poorly understood[1, 8, 51]. Elucidating the mechanisms of multi-protein complex formation is imperative not only since these interactions underlie the initiation of cellular metabolic processes, but address the fundamental concerns of genomic functionality[52].

Currently, two competing models of macromolecular assembly, categorized as either top-down[53] or bottom-up[54], are jockeying for acceptance, with a large body of literature supporting both propositions. In top-down assembly, the constituents of the final complex are hypothesized to bind one another prior to DNA interactions and form a stable macromolecular machine termed a "factory"[53, 55]. Such factories likely persist for a long duration in the cellular environment, stabilized by the numerous binding interactions of the many subunits, and represent the most efficient initiation of a metabolic function. This approach is supported by the well-documented observation of large, multi-mega Dalton RNAP complexes that have been identified by optical and electron microscopy, as well as mass spectrometry[56-59]. These factories have been found to persist *in vivo*and *in vitro* for long durations, even when transcription halts[60]. It remains unclear how the factory initially assembles, either in a concerted, step-wise manner, or through uncorrelated, stochastic interactions.

In contrast, bottom-up assembly hypothesizes *de novo* formation of the full complex each time a metabolic process is initiated, with subunits binding to the target site as the crucial first step of assembly. Such an approach would lead to highly inefficient initiation of metabolic processes[14], but is well supported by the large body of work documenting the dynamic and transient binding interactions of many nuclear proteins[54]. Detailed FRAP studies of RNAP I and RNAP II indicate that individual subunits and associated transcription factors do not remain stably incorporated into active complexes, but rather exchange with a nucleoplasmic pool of unengaged proteins. Further, in some systems, transcription initiation has been documented to be initiated with low efficiency, but following RNAP II complex formation, to proceed with high efficiency, which supports *de novo* assembly[3, 14]. Again, it is unclear whether the assembly

proceeds through a step-wise process, or through stochastic binding interactions of the component subunits at their genomic site of action. In the latter variant, assembly would be particularly inefficient since most stochastic interactions would likely be out of sequence and lead to an aborted intermediate. While FRET evidence has accumulated that indicates spliceosome subunits do form partially assembled intermediates, their role in the final assembly is not yet understood[5].

2. Mechanism Evaluation: Choosing the Right Time and Place

The difficulty in resolving the mechanisms of protein assembly stem in part from the large body of evidence in support of both opposing models. The top-down assembly model relies heavily on structural observations in which molecular factories can be visualized; however, these observations are handicapped by a lack of simultaneous spatial and temporal resolution. Optical methods have often been able to reproducibly observe punctate nuclear structures corresponding to active protein assemblies[1, 256, 61], yet the spatial resolution is lacking to determine the true size of the complexes [57](though the recent development of live cell super-resolution optical microscopy may provide such insights) . While complimentary observations have been made through electron microscopy[57], such techniques lack the temporal resolution to confirm the long time duration over which these protein complexes must remain intact to qualify as factories. Additionally, bulk biochemical studies that have demonstrated the activity and stability of purified complexes[62, 63] can perturb function due to the non-native solvent environment. The bottom-up assembly model is predicated largely on optical microscopy work that has conclusively confirmed the dynamic exchange of most complexed nuclear proteins. However, such findings do not rule out the formation of stable

factories following macromolecular assembly. Further, biochemical studies that have suggested step-wise or stochastic assembly mechanisms again suffer from a lack of cellular context, in which molecular crowding or subunit confinement could drastically alter protein interactions[64].

In general, all studies of macromolecular assembly have been complicated by the confounding presence of chromatin, which provides varying degrees of molecular confinement and presents nucleation sites for complex formation[3, 3, 14, 65, 66]. In fact, both assembly models posit cellular molecular crowding and the resulting reduced diffusional mobility as favorable evidence. Bottom-up assembly is viewed as benefiting from the reduced diffusional mobility of complex subunits, which would lengthen interaction times and promote more frequent collisions, thereby promoting macromolecular assembly from stochastic collision events. In contrast, proponents of the top-down assembly mechanism cite the crowded nuclear environment as favorable for maintaining the stability of an assembled complex, yet the inability of a large factory to effectively diffuse throughout the nuclear volume is often overlooked. Only by observing protein behavior with high spatiotemporal resolution in a model system where the effects of chromatin can be eliminated, can the initial stages of macromolecular assembly be discerned.

In practice, optical microscopies coupled with cell types with known architectures can be exploited to achieve these requirements. Our research group has made extensive use of high resolution FRAP microscopy, along with polytene cell lines, to capture protein *diffusion in vivo* and distinguish between the influences of chromatin and molecular crowding.

Importantly, recent work completed by our group has indicated that a hybrid mechanism likely mediates complex formation. We have found that large macromolecular assemblies exhibit remarkable stability both *in vivo* and *in vitro*, yet likely form through the stochastic assembly of partially assembled intermediates with or without the assistance of chromatin nucleation sites.

Research Aims and Scope

Through my graduate research, methodologies have been developed for gaining insights into the multifaceted phenomenon of protein recruitment. As detailed in Chapter 2, my initial projects focused on the development of single molecule imaging techniques, and confirm two-photon DNA photodamage with visible light. Damage cross sections were determined for biologically relevant DNA samples at different visible wavelengths. This work was later extended for *in vivo*use by my lab mate, providing a powerful tool to initiate DNA damage and enzymatic repair in a user controlled setting. Subsequently, in a project stemming from an effort to couple the two-photon damage assay with a DNA manipulation platform, the rate of DNA photodamage mediated by commonly used DNA intercalating dyes was quantified. As described in Chapter 3, these results were confirmedapplicable for a wide range of imaging conditions enabling fellow researchers to evaluate how their optical imaging configurations perturb biological samples. Finally, Chapter 4 covers my transition to *in vivo* systems and investigation of the macromolecular assembly mechanisms of RNAP II. This fruitful work resulted in a new understanding of multiprotein nucleation processes and allows us to speculate as to a modular control mechanism over gene expression.

REFERENCES

(1) Hemmerich, P. *Zellbiologie* **2005***, 31*, 18.

(2) Matera, A. G.; Izaguire-Sierra, M.; Praveen, K.; Rajendra, T. K. *Dev. Cell* **2009***, 17*, 639-647.

(3) Dundr, M.; Hoffmann-Rohrer, U.; Hu, Q.; Grummt, I.; Rothblum, L. I.; Phair, R. D.; Misteli, T. *Science* **2002***, 298*, 1623-1626.

(4) Rippe, K. *Curr. Opin. Genet. Dev.* **2007***, 17*, 373-380.

(5) Rino, J.; Carmo-Fonseca, M. *Trends Cell Biol.* **2009***, 19*, 375-384.

(6) Houten, B. V.; Croteau, D. L.; Vecchia, M. J. D.; Wang, H.; Kisker, C. *Mut. Res.* **2005***, 577*, 92-117.

(7) van Mameren, J.; Peterman, E. J. G.; Wuite, G. J. L. *Nucleic Acids Res.***2008***, 36*, 4381-4389.

(8) Hager, G.; Elbi, C.; Becker, M. Curr. Opin. Genet. Dev.**2002***, 12*, 137.

(9) Mueller, F.; Mazza, D.; Stasevich, T. J.; McNally, J. G. *Curr. Opin. Cell Biol.* **2010***, 22*, 403-411.

(10) Krichevsky, O.; Bonnet, G. *Rep Prog Phys* **2002***, 65*, 251-297.

(11) Graneli, A.; Yeykal, C. C.; Prasad, T. K.; Greene, E. C. *Langmuir* **2006***, 22*, 292-299.

(12) Xie, X. S.; Choi, P. J.; Li, G.; Lee, N. K.; Lia, G. *Annu. Rev. Biophys.***2008***, 37*, 417-444.

(13) Friedberg, E. C.; Walker, G. C.; Siede, W. *DNA repair and mutagenesis;* ASM Press: Washington D.C., 1995; .

(14) Darzacq, X.; Shav-Tal, Y.; de Turris, V.; Brody, Y.; Shenoy, S. M.; Phair, R. D.; Singer, R. H. *Nat Struct Mol Biol* **2007***, 14*, 796-806.

(15) Friedberg, E. C. *Nature* **2003***, 421*, 436-440.

(16) Mone, M. J.; Bernas, T.; Dinant, C.; Goedvree, F. A.; Manders, E. M. M.; Volker, M.; Houtsmuller, A. B.; Hoeijmakers, J. H. J.; Vermeulen, W.; Driel, R. v. *Proc. Natl. Acad. Sci. U. S. A.* **2004***, 101*, 15933-15937.

(17) Gorman, J.; Chowdhury, A.; Surtees, J. A.; Shimada, J.; Reichman, D. R.; Alani, E.; Greene, E. C. *Mol. Cell* **2007***, 28*, 359-370.

(18) Patrick, M. H.; Rahn, R. O. *Photochemistry and Photobiology of Nucleic Acids;* Academic Press: New York, 1976; Vol. II, pp 35-96.

(19) Patrick, M. H. In *Physical and chemical properties of DNA;* Wang, S. Y., Ed.; *Photochemistry and Photobiology of Nucleic Acids*; Academic Press: New York, 1976; Vol. 2, .

(20) Katsumi, S.; Kobayashi, N.; Imoto, K.; Nakagawa, A.; Yamashina, Y.; Muramatsu, T.; Shirai, T.; Miyagawa, S.; Sugiura, S.; Hanaoka, F.; Matsunaga, T.; Nikaido, O.; Mori, T. *J. Invest. Dermatol.* **2001***, 117*, 1156-1161.

(21) Mone, M. J.; Volker, M.; Nikaido, O.; Mullenders, L. H. F.; Zeeland, A. A. v.; Verschure, P. J.; Manders, E. M. M.; Driel, R. v. *EMBO Rep.* **2001***, 2*, 1013-1017.

(22) Lukas, C.; Melander, F.; Stucki, M.; Falck, J.; Bekker-Jensen, S.; Goldberg, M.; Lerenthal, Y.; Jackson, S. P.; Bartek, J.; Lukas, J. *EMBO Journal* **2004***, 23*, 2674-2683.

(23) Trautlein, D.; Deibler, M.; Leitenstorfer, A.; Ferrando-May, E. *Nucleic Acids Res.* **2010***, 38*, e14.

(24) Daddysman, M.; Fecko, C. *Biophys. J.* **2011***, 101*, 2294-2303.

(25) Tycon, M. A.; Chakraborty, A.; Fecko, C. J. *J. Photochem. Photobiol. B, Biol.* **2011***, 102*, 161-168.

(26) Gut, I. G.; Hefetz, Y.; Kochevar, I. E.; Hillenkamp, F. *J Phys. Chem***1993***, 97*, 5171-5176.

(27) Hefetz, Y.; Dunn, D. A.; Deutsch, T. F.; Buckley, L.; Hillenkamp, F.; Kochevar, I. E. *J. AM. CHEM. SOC.,***1990***, 112*, 8528-8532.

(28) Zipfel, W. R.; Williams, R. M.; Webb, W. W. *Nat. Biotechnol.* **2003***, 21*, 1369-1377.

(29) Bekker-Jensen, S.; Lukas, C.; Melander, F.; Bartek, J.; Lukas, J. *J. Cell Biol.***2005***, 170*, 201-211.

(30) Bekker-Jensen, S.; Lukas, C.; Kitagawa, R.; Melander, F.; Kastan, M. B.; Bartek, J.; Lukas, J. *J. Cell Biol.***2006***, 173*, 195-206.

(31) Denk, W.; Strickler, J. H.; Webb, W. W. *Science* **1990***, 248*, 73-76.

(32) Friedberg, E. C. *DNA Repair* **2002***, I*, 855-867.

(33) Lakowicz, J. R., Ed.; In *Principles of Fluorescence Spectroscopy;* Springer: New York, 2006; .

(34) Axelrod, D. In *Total internal reflection fluorescence microscocpy;* Torok, P., Kao, F., Eds.; Optical Imaging and Microscopy; Heidelberg: Berlin, 2007; Vol. 87, pp 195-236.

(35) Kad, N.M., Wang, H., Kennedy, G.G., Warshaw, D.M., Van Houten,B. *Mol. Cell* **2010***, 37*, 702-713.

(36) - Tanner, N. A.; - Loparo, J. J.; - van Oijen, A. M. - *J Vis Exp* , - e1529.

(37) Blainey, P. C.; Oijen, A. M. v.; Banerjee, A.; Verdine, G. L.; Xie, X. S. *PROC. NATL. ACAD. SCI. U.S.A.***2006***, 103*, 5752-5757.

(38) Wang, H.; Tessmer, I.; Croteau, D. L.; Erie, D. A.; Houten, B. V. *Nano Letters* **2008***, 8*, 1631-1637.

(39) Halliwell, B.; Aruoma, O. I. *FEBS Lett.* **1991***, 281*, 9-19.

(40) Schweitzer, C.; Schmidt, R. *Chem. Rev.* **2003***, 103*, 1685-1757.

(41) SIES, H. *Eur. J. Biochem.***1993***, 215*, 213-219.

(42) Limoli, C. L.; Ward, J. F. *Radiat. Res* **1993***, 134*, 160-169.

(43) Saran, M.; Bors, W. *Radiat Environ Biophys* **1990***, 29*, 249-262.

(44) Teoule, R. *Int. J. Radiat. Biol.* **1987***, 51*, 573-589.

(45) Ward, J. F. *Int. J. Radiat. Biol.* **1990***, 57*, 1141-1150.

(46) Guo, H.; Tullius, T. D. *PROC. NATL. ACAD. SCI. U.S.A.,.* **2003***, 100*, 3743-3747.

(47) Siddiqi, M. A.; Bothe, E. *Radiat. Res , 112*, 449-463.

(48) Caldecott, K. W. *Nature Rev. Genet.***2008***, 9*, 619-631.

(49) Akerman, B., Tuite, E. *Nucleic Acids Res.* **1996***, 24*, 1080.

(50) Tycon, M. A.; Dial, C. F.; Faison, K.; Melvin, W.; Fecko, C. J. *Anal. Biochem.* **2012***, 426*, 13-21.

(51) Cardoso, M. C.; Leonhardt, H. *J. Cell. Biochem.* **1998***, 70*, 222-230.

(52) Misteli, T. *Cell* **2007***, 128*, 787.

(53) Cook, P. R. *J. Mol. Biol.* **2010***, 395*, 1-10.

(54) Phair, R. D.; Misteli, T. *Nature* **2000***, 404*, 604.

(55) Cook, P. R. *Science* **1999***, 284*, 1790-1795.

(56) Melnik, S.; Deng, B.; Papantonis, A.; Baboo, S.; Carr, I. M.; Cook, P. *Nature Methods* **2012***, 8*, 963.

(57) Eskiw, C.; Fraser, P. *J. Cell Sci.* **2011***, 124*, 3676.

(58) Chakalova, L.; Debrand, E.; Mitchell, J. A.; Osborne, C. S.; Fraser, P. *Nature Rev. Genet,***2005***, 6*, 669-677.

(59) Osborne, C.; Chakalova, L.; Brown, K.; Carter, D.; Horton, A.; Debrand, E.; Goyenechea, B.; Mitchell, J.; Lopes, S.; Reik, W.; Fraser, P. *Nat. Genet.* **2004***, 36*, 1065-1071.

(60) Mitchell, J. A.; Fraser, P. *Genes Dev.***2008***, 22*, 20.

(61) Wilson, C. J.; Chao, D. M.; Imbalzano, A. N.; Schnitzler, G. R.; Kingston, R. E.; Young, R. A. *Cell* **1996***, 84*, 235-244.

(62) Schneider, D. A.; Nomura, M. *PROC. NATL. ACAD. SCI. U.S.A.,* **2004***, 101*, 15112-15117.

(63) Grummt, I. *Genes Dev.,***2003***, 17*, 1691.

(64) Hancock, R. *J. Struct. Biol.* **2004***, 146*, 281-290.

(65) Misteli, T.; Gunjan, A.; Hock, R.; Bustin, M.; Brown, D. T. *Nature* **2000***, 408*, 877-881.

(66) Kimura, H.; Sugaya, K.; Cook, P. R. *J. Cell Biol.,* **2002***, 159*, 777-782.

CHAPTER 2

GENERATION OF DNA PHOTOLESIONS BY TWO-PHOTON ABSORPTION OF A FREQUENCY-DOUBLED TI:SAPPHIRE LASER

"The microscope with its accessories is by far the least understood, the most inefficiently operated, and the most abused of all laboratory instruments"

-Charles Shillaber

Overview:

The formation of spatially localized regions of DNA damage by multiphoton absorption of light is an attractive tool for investigating DNA repair. Although this method has been applied in cells, little information is available about the formation of lesions by multiphoton absorption in the absence of exogenous or endogenous sensitizing agents. Therefore, we have investigated DNA damage induced *in vitro* by direct two-photon absorption of frequency-doubled femtosecond pulses from a Ti:sapphire laser. We first developed a quantitative polymerase chain reaction assay to measure DNA damage, and determined that the quantum yield of lesions formed by one-photon absorption of 254 nm light is 7.86×10^{-4}. We then measured the yield of lesions resulting from exposure to the visible femtosecond laser pulses, which exhibited a quadratic intensity dependence. The two-photon absorption cross section of DNA has a value (per nucleotide) of 2.6 GM at 425 nm, 2.4 GM at 450 nm, and 1.9 GM at 475 nm. A comparison of these *in vitro* results to several *in vivo* studies of multiphoton

photodamage indicates that the onset of DNA damage occurs at lower intensities *in vivo*; we suggest possible explanations for this discrepancy.

Introduction

Irradiation by ultraviolet (UV) light is one of the most extensively used methods for exploring the biological consequences of DNA damage and repair. Nucleic acids exhibit an absorption maximum near 260 nm, but efficiently absorb light with wavelengths between 200-300 nm [1, 2]. The most common method of photolesion formation is by exposure to 254 nm radiation from low pressure mercury lamps. Although simple to implement, this method creates photolesions with a random spatial distribution; it is often desirable to generate photolesions in a well-defined location to study protein dynamics in response to DNA damage. Due to the extremely poor UVB and UVC transmission of common glasses and mirrors, light in this range is difficult to manipulate via a microscopy-based apparatus. This prohibits the easy pairing of short-wave UV lasers with conventional microscopy optics [3]. More recently, other methods to generate localized photolesions have been applied to observe the response of fluorescently-tagged repair proteins in live cells [4]. One technique introduces UVC light through 3-5 μm pores in a polycarbonate filter [5-8]. When applied to cultured mammalian cells, the DNA damage is localized to a region that is smaller than the nucleus, but still immense in comparison to molecular length scales. Another method to introduce DNA damage involves laser-based irradiation of pre-sensitized cells in the 300-405 nm range [9-16]. While generating photolesions that are localized to smaller 2D regions, it suffers from the potentially serious drawback that the sensitizing agent could perturb the natural response of the biological system to damage.

Ultimately, while these methods can localize the extent of DNA damage in two dimensions, they do not offer confinement in the third dimension.

As an alternative, we explored the use of multiphoton absorption of DNA as a means to produce photolesions with conventional optics. Nonresonantmultiphoton absorption is the process in which two or more photons interact with a molecule simultaneously to generate an excited state equivalent in energy to the summation of the absorbed photons [17]. Since simultaneous absorption requires a large photon flux, the probability of two-photon absorption depends quadratically on the intensity of the incident light. This property is exploited to achieve depth discrimination in two-photon microscopy since absorption can only occur at the focal point of an objective lens as it is the region of highest intensity [18, 19]. Similarly, two-photon absorption-induced DNA damage has the advantage of generating photolesions in a three-dimensionally pre-defined region of space, which is superior to the spatially random and widespread regions of damage induced by widefield UV illumination. Additionally, it does not require the introduction of an exogenous sensitizer that could perturb normal cellular functions.

In our work, blue femtosecond pulses of light produced by frequency-doubling the output of a Ti:sapphire laser are focused on homogenous solutions of DNA *in vitro*. Although blue light is relatively harmless to most biomolecules, absorption of multiple blue photons in the focused region can excite transitions similar to those caused by exposure to UV light, thus generating localized DNA photolesions. While multiphoton irradiation has previously been used to generate DNA photodamage *in vivo*[20-22], the potential role of sensitizing agents (both

naturally occurring and intentionally added) as mediators of energy transfer have not been fully considered. More information is needed about the amount of direct multiphoton absorption of DNA, so that this phenomenon can be applied in conjunction with ultrasensitive microscopy-based methods to study DNA repair protein dynamics [23, 24].

It is challenging to assay DNA photolesions produced by two-photon absorption because of the inherently microscopic conditions in which they are produced. DNA damage assays premised on techniques as varied as gel electrophoresis [25], HPLC [26], and radiolabeling [27] require significantly more sample than is that contained in the ~femtoliter focal volume of an objective lens. To compensate, we have adopted the approach of irradiating 10 μL droplets by repeatedly raster scanning a focused laser beam through the sample in different axial planes using a laser scanning system. We have also developed a highly sensitive quantitative polymerase chain reaction (QPCR) to detect DNA damage. By combining these techniques, we have observed two-photon absorption-induced DNA damage, and determined the relevant absorption cross sections at 425, 450, and 475 nm. A comparison of our results to previously published *in vivo* studies indicates that the generation of photodamage by two-photon absorption *in vitro* requires higher intensities than expected based on the *in vivo* experiments.

Materials and Methods

1. Materials

High-purity grade chemicals were purchased from Fisher Scientific or Sigma-Aldrich. pBR322, EcoRI, Nb.BsmI, bovine serum albumin (BSA) and dNTPs were obtained from New England Biolabs. The rTth DNA polymerase PCR system and accompanying reagents were

purchased as the GeneAmp XL PCR kit from Applied Biosystems, and custom primers were synthesized by Integrated DNA Technologies. A QIAquick PCR Purfication kit was obtained from Qiagen and the Quant-iT^{TM}PicoGreendsDNAreagent was obtained from Invitrogen. PCR was performed in an EppendorfMastercycler, absorption measurements were made on a NanoDrop 1000 Spectrophotometer and the PicoGreen fluorescence assay was read on a BMG PheraStar plate reader. UV-induced DNA damage was generated with a SpectrolineCrosslinker containing 254 nm tubes (the crosslinker was operated with only half of the maximum number of bulbs to reduce the photon flux). The laser setup is described in detail below.

2. DNA sample preparation

The PCR amplification efficiency of supercoiled DNA is poor [28], so the samples used to develop the QPCR assay and for subsequent irradiation studies were prepared from linearized pBR322 DNA. Additionally, concerns that commercial products may contain trace amounts of photosensitizers motivated us to use DNA samples generated in-house by PCR.

We linearized supercoiled pBR322 with EcoRI (5 units/mg plasmid, recommended by New England BioLabs), confirmed the product by 1% agarose gel electrophoresis, and then amplified it using the GeneAmp XL PCR kit. The initial PCR reaction mixture was composed of sterile water, 5 pg/µL linearized pBR322, 1X rTth buffer, 200 µMdNTPs, 1.2 µMMg(OAc)$_2$, 0.1 mg/mL BSA, and 0.4 µM of each primer. The rTth polymerase was diluted in 1X rTth buffer and 1 unit was added to each amplification reaction. The primers sequences, which amplify a 4.3 kb fragment of pBR322, are[28]:

pBR102F (5'-CAGGCACCGTGTATGAAATCTA-3')

pBR399R (5'- TGGATCTCAACAGCGGTAAGA-3')

The dNTPs and primer solutions were stored as aliquots to avoid excessive freeze thaw cycles. The DNA was amplified using a three-step temperature program: initial denaturation at 94^{o}C for 1 min, then 28 cycles of denaturation at 94^{o}C (15s), annealing at 62^{o}C (30s), and elongation at 66^{o}C (240s) in the EppendorfMastercycler. The PCR product from several tubes was consolidated and purified with a PCR cleanup kit. The DNA concentration was measured by absorption at 260 nm (typical concentrations after PRC ~180 ng/uL) and stored as single-use aliquots at -80° C for use as a DNA template in subsequent experiments.

As a control for the QPCR assay, a portion of the PCR-generated linearized pBR322 was enzymatically nicked with Nb.BSmI, which cleaves only one strand of the double-stranded DNA substrate. The enzyme was heat inactivated and removed using the PCR cleanup kit. The DNA concentration was measured by absorption at 260 nm and stored as single-use aliquots at -80° C for subsequent experiments.

3. QPCR assay of DNA damage

The QPCR assay was used to amplify DNA templates that have been diluted with Millipore water to a working concentration of 0.05ng/µL. The initial PCR reaction mixture is identical to the aforementioned mixture used to generate template. Each QPCR assay run includes four mandatory controls: an undamaged pBR322, a serial dilution at half the concentration of the undamaged pBR322 to ensure the assay is functioning properly, the 0.5 lesion/strand nicked pBR322 used to monitor sample amplification, and a blank sample prepared without template. The QPCR assay was run for 14 cycles at the three-step

temperature program described above. This number of PCR cycles used was determined empirically with the goal of maintaining a two-fold increase in amplification between control samples. Samples were run in duplicate or triplicate.

Following the PCR amplification, the PCR products were quantified using the PicoGreen DNA quantification assay. The samples were prepared in a 96 well plate to be processed by the PheraStar plate reader, with filters corresponding to the 488/520 nm excitation/emission spectrum of PicoGreen. In addition to the PCR products, a set of pBR322 standards made by serial dilution was run to calculate the final concentration of the amplified products and to calculate the true starting concentrations of the template stocks. The dilution series always included a blank sample (water) to correct the fluorescence measurements. The PCR products were diluted with TE buffer (10mM Tris-EDTA, pH 8.3, adjusted with dilute NaOH and HCl) and mixed with diluted PicoGreen solution as per the manufacturer's instructions.

4. UV irradiation

In order to determine the quantum yield of photolesions in response to 254nm UV light, linearized pBR322 (0.05 ng/µL) was irradiated in a UV oven with an emission peak at 254 nm. Sample aliquots (20µL) and a KI/KIO_3 chemical actinometer (20µL) [29, 30] were simultaneously irradiated on a glass slide for varying exposure times, generating damage at a range of UV dosages [29, 30]. The number of incident photons was determined by the actinometer.

5. Femtosecond laser irradiation

In order to investigate photolesion formation that results from two-photon absorption, homogenous DNA samples were irradiated by focused 400-500 nm ultrashort pulses using the apparatus diagrammed in Figure 2.1. Our setup used tunable near-infrared, ~140 fs pulses produced at 80 MHz by a Coherent Chameleon Ultra II Ti:sapphire oscillator. An electro-optic modulator and polarizer placed directly after the laser controlled the intensity used for each experiment. We generated the second harmonic frequency of the pulses by focusing the beam into a 2 mm path length β-barium borate crystal cut for type-I phase matching. The focal length of the lens, and thus doubling efficiency, was somewhat limited by the requirement that the visible beam be relatively symmetric and free of astigmatism. The residual near-infrared light was rejected with a contrast ratio of at least 100:1 by reflecting the beam off of two dichroic mirrors. The visible femtosecond pulses were introduced into a home-built laser-scanning microscope based on an Olympus IX81 inverted microscope. Mirrors mounted on computer-controlled galvanometers determined the angle with which the laser beam enters the objective lens. To irradiate a large field of view but maintain sufficiently high peak intensities, a 10X, 0.30NA objective lens was used to focus the beam within the sample. The back aperture of the lens was slightly underfilled to maximize transmission while maintaining a tight focus. The beam was raster scanned in a sinusoidal pattern through each axial plane of the sample. The focal plane was adjusted by translating the objective lens in the axial dimension using the motorized nosepiece of the microscope.

To determine the intensity dependence of photolesion formation, identical samples were held in separate wells of a 384-well microplate and irradiated by subsequent scans at various powers. Samples were separated by a sufficient number of wells to avoid

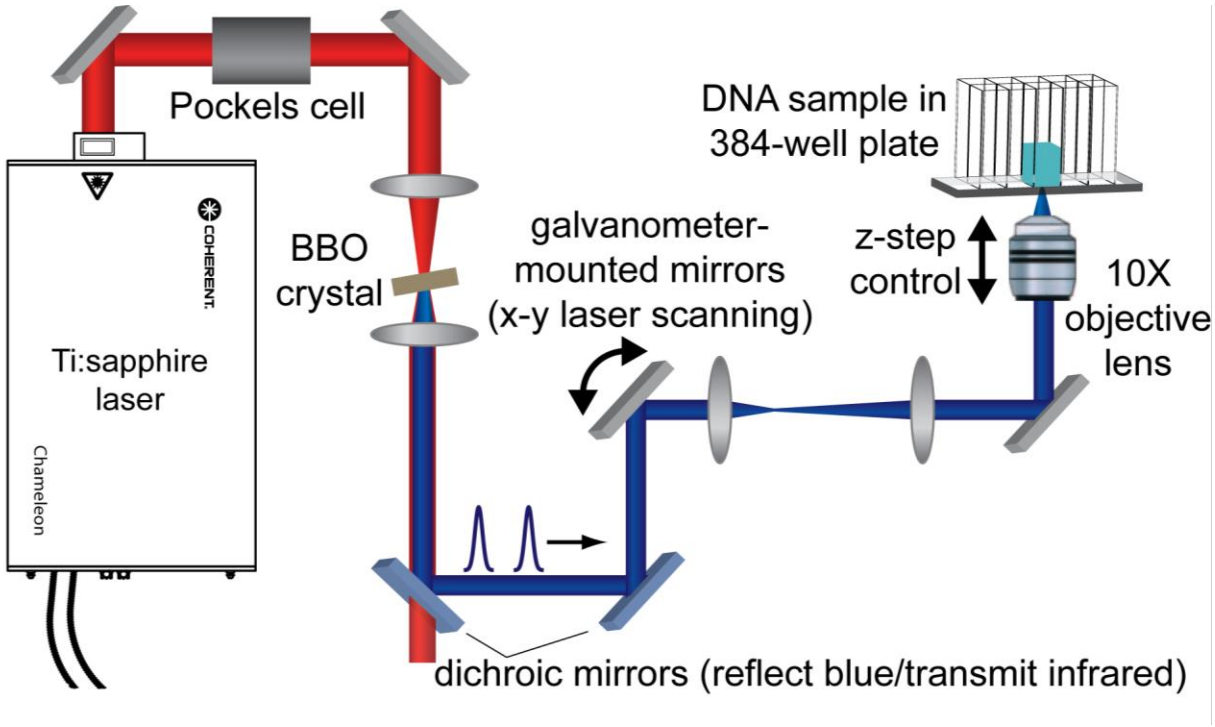

Figure 2.1-Schematic diagram of the irradiation apparatus. Femtosecond pulses from a tunable Ti:sapphire laser are attenuated by a Pockels cell/polarizer and subsequently focused into a β-barium borate (BBO) crystal that doubles their frequency. After removing the residual fundamental light with dichroic mirrors, the second-harmonic beam is focused into the sample using a 10X 0.3NA objective lens. It is raster scanned through each axial plane of the sample using galvanometer-mounted mirrors.

crosstalk. A metal guide and micrometer were used to position different wells reproducibly with respect to the objective lens. The sample in each well was irradiated by raster scanning a focused beam through the sample using the galvanometer-mounted mirrors, and then repeating for each axial plane. The number of axial scans performed was determined by the height of the liquid column in the well, and it was confirmed that over-scanning the height of the well did not cause additional photodamage. The axial plane spacing was approximately equal to the full-width at half-maximum of the calculated point spread function, and the number of laser pulses incident on each point in the sample was estimated from the calculated point spread function and average velocity. The irradiation intensity was adjusted for each sample by varying the incident laser power using the aforementioned electro-optic modulator without adjusting the beam focusing, and the incident laser power measured at the objection prior to irradiation. The point spread function was measured at 450 nm by imaging 100 nm fluorescent beads immobilized on a glass surface. For each wavelength, 10 µL of linearized pBR322 samples (0.05 ng/µL) was irradiated at a series of incident powers. Each sample was removed from the microwell plates and the amount of DNA damage was evaluated using the PCR-based assay described above. Care was taken to ensure that the control samples were treated identically to irradiated samples, with the exception of exposure to the laser.

Results and Discussion

This work proceeded in three phases: the development of a QPCR assay to quantify the formation of DNA photolesions, measurement of the damage induced by exposure to a UV light source, and quantification of the damage induced by multiphoton absorption.

1. Development of a QPCR assay of DNA damage

The assay used to quantify the formation of photolesions after irradiation was based on a method developed by Van Houtenand co-workers[31-33]. The method is premised on a reduction in DNA polymerase transcription efficiency by strand breaks or by bulky forms of damage, such as thymine dimers, which block the progress of polymerases not containing exonuclease activity[34]. In the first round of PCR amplification, a single lesion removes a damaged strand from future replication, as the truncated transcription product will not be able to anneal with the primers required to initiate the next round. Thus damaged sample populations are not amplified as quickly due to the reduction in the number of strands available for transcription, and will manifest damage relative to an undamaged control sample (Fig. 2.2-a). The sensitivity of this assay is related to the length of the PCR target, since a single lesion in a long template causes a larger reduction in the quantity of DNA produced than a single lesion in a shorter template. The use of long DNA templates and the ability of PCR to amplify sub-nanogram quantities of starting template makes this assay ideal for measuring low damage rates of microscale samples[33].

The number of photolesions or damage sites is determined by measuring the ratio of the amplification of the damaged DNA samples to an undamaged control sample, as described below. The degree of amplification (total DNA synthesized) after the samples are subjected to the PCR reaction is determined by fluorescence measurements made on a multiwell plate reader after addition of the DNA binding fluorophore PicoGreen. A tenet of this assay to produce quantitative results is that a change in the sample input concentration produces a linear change in the amplification. Therefore, implementation of the assay required

optimization of a quantitative PCR protocol for the DNA template under investigation followed by validation that amplification linearity was reliable.

a. Conditions for quantitative PCR

PCR reactions proceed through three phases: early exponential growth, reduced efficiency "leveling off" as reagents become limiting, and finally saturation or plateau[33]. In order to maintain the linear relationship between sample input and output concentrations, the QPCR reaction must be kept in the exponential phase, where a semilog analysis yields a linear plot. For a quantitative assay, the cycle number chosen is a compromise between saturation and signal to noise limits. The cycle used should be low enough for undamaged samples to remain in the exponential phase while amplifying, but high enough to yield a large degree of amplification of the control relative to damaged samples to achieve a good signal to noise ratio when measuring the amount of PCR product.

We determined the optimal quantitative amplification conditions by generating PCR growth curves of a linearized pBR322 DNA template at 0.05 and 0.025 ng/μL concentrations and selecting the cycle number that best corresponded to a two-fold amplification, in our case cycle 14 (Fig. 2.2-b). It should be noted that the actual template concentration used in each experiment was measured using the PicoGreen assay, and kept dilute enough to avoid high amplification nearing PCR saturation.

After establishing the number of PCR cycles required to achieve a linear dependence, the dynamic range of the assay was determined by amplifying a serial dilution of pBR322 and observing the range over which a linear response was maintained (Fig. 2.2-c). Samples were

analyzed in triplicate and a relative standard deviation (RSD) of less than 5% was typical. A linear dynamic range of approximately 50:1 was established. The upper end of this range is set by the need to remain in the exponential PCR region and the lower end is limited by variability in the background fluorescence in the PicoGreen assay (which is the source of the small y-intercept in the fit of Fig. 2.2-c). This range set the boundaries for output amplifications still considered reliable, and contributed to the determination that a pBR322 input concentration for PCR of 0.05 ng/μL provided the desired sensitivity.

After optimizing the assay, two controls were included in all subsequent PCR amplifications to ensure quantitative results each time samples were analyzed. The first control measured the PCR amplification efficiency of the undamaged template by including a "half-template" sample made by two-fold serial dilution. This sample was expected to show a 1:2 amplification compared to the undamaged template, and deviations from this value indicate a problem with assay. The second control was a PCR-generated linearized pBR322 template that has been nicked on one strand by the enzyme Nb.BSmI, and used at the same concentration as the undamaged template. This sample also acted as a reference for the amplification by mimicking damage to one strand of each duplex (this "damage" is the result of a deterministic cleavage as opposed to the statistically random process of photodamage described below). A significant deviation from the expected 50% amplification value for either control sample indicates a problem with the undamaged template or PCR conditions; in this event, the assay results were discarded.

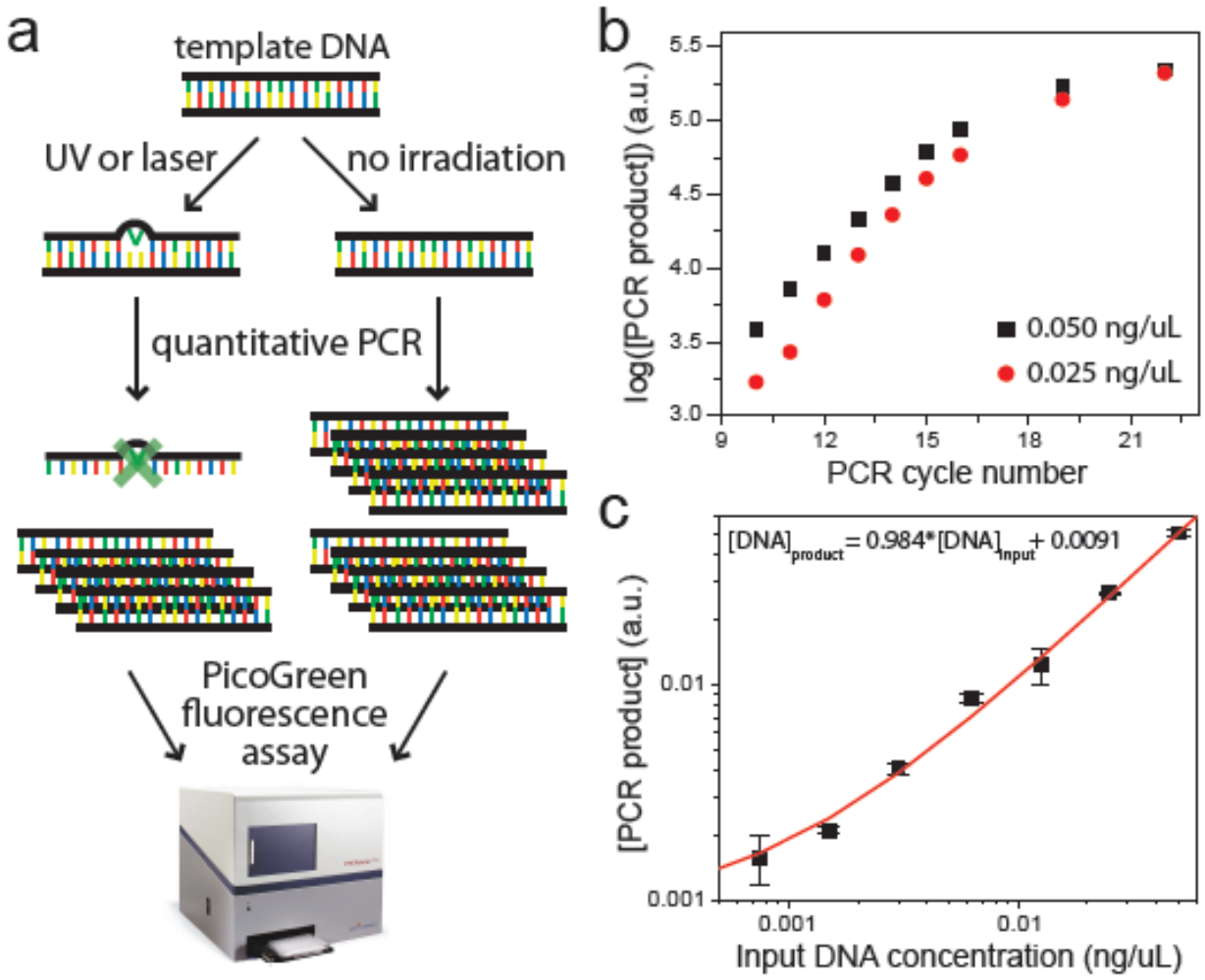

Figure 2.2-Detection of DNA photolesions using quantitative PCR.(a) The presence of a bulky photolesion (x) in the irradiated sample causes a reduction in PCR amplification. **(b)** Depiction of the exponential, linear, and plateau phase of the PCR reaction. The exponential phase cycle resulting in a two-fold increase in amplification was chosen, cycle 14 for our samples. **(c)** A dilution series of the input amount of DNA was used to determine the range over which amplification remained linear. A minor fluorescence background in the PicoGreen assay causes small deviations from linearity at the lowest concentrations, limiting the dynamic range of the assay to ~50:1.

b. Statistical treatment of randomly distributed DNA photolesions

Since the number of damage sites on a single strand is not detected directly, the amount of damage must be treated statically. This results from an inability to distinguish the reduction in amplification from multiple lesions on the same strand from the reduction in amplification from a single lesion (Fig. 2.3). The formation of photolesions is a random process governed by the Poisson distribution [35], which is applicable to situations involving occurrences that happen at a well-defined average rate but that are independent of previous events. The probability *P* that a specific strand has exactly *n* lesions if the average number of lesions per strand is μ is given by:

$$\boldsymbol{P(n)} = \frac{\mu^{\boldsymbol{n}} \boldsymbol{e}^{-\mu}}{\boldsymbol{n}!} \quad (1)$$

The average number of lesions formed per strand can be determined from the probability of detecting a strand that is devoid of lesions, known as the zero class probability (*n = 0*).

The QPCR assay only amplifies undamaged strands, so its output is directly proportional to the zero class probability [32]. Therefore, the average number of lesions formed on each strand is calculated from the measurable ratio of the amount of DNA produced in the PCR reaction of the irradiated DNA to the amount produced from unirradiated DNA:

$$\mu = -\ln\left(\frac{\text{DNA produced from irradiated template}}{\text{DNA produced from unirradiated template}}\right) \quad (2)$$

This ratio is determined from the fluorescent intensities of the final PCR reaction mixtures in a PicoGreen assay.

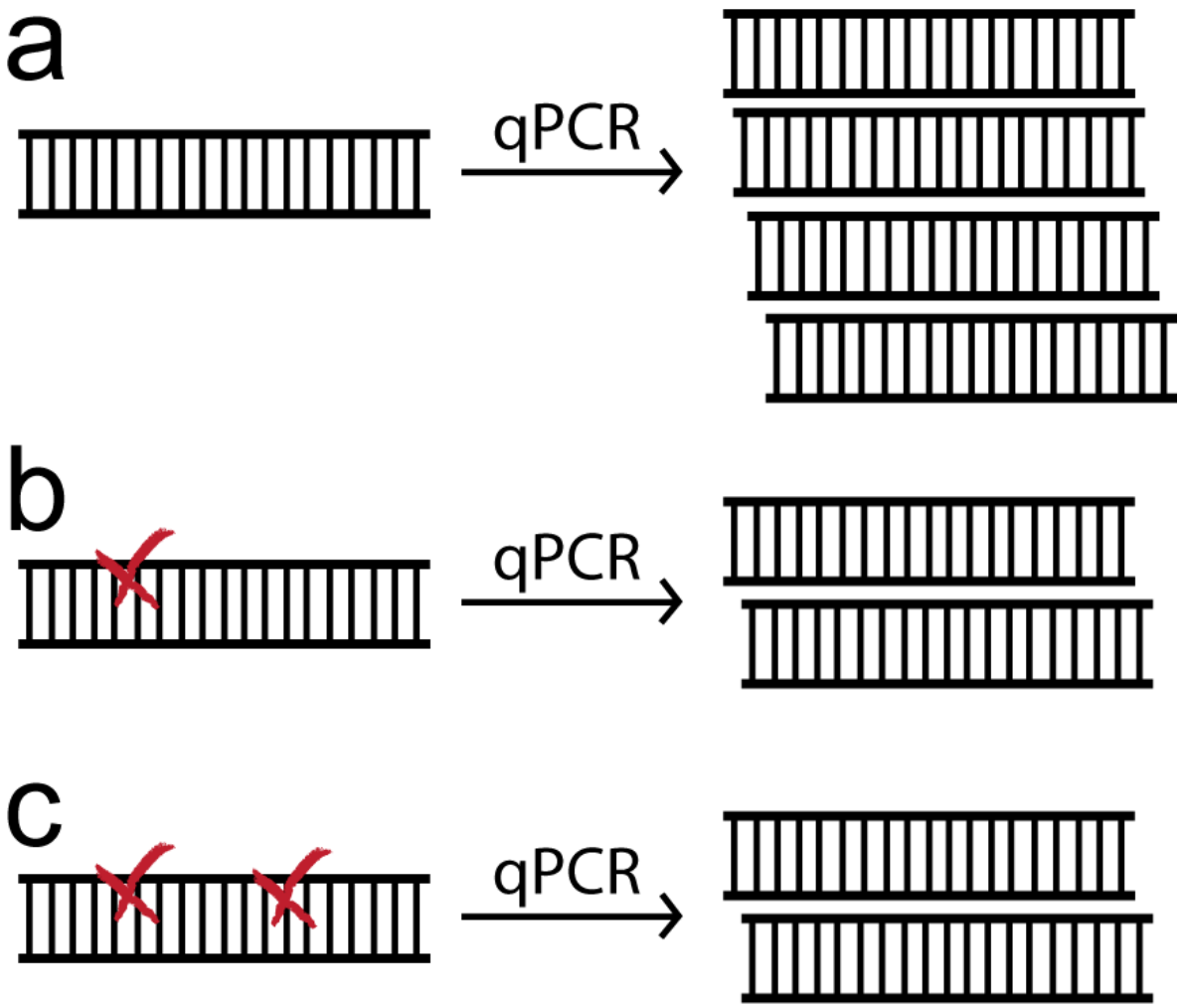

Figure 2.3- Poisson statistics are required to determine the number of DNA lesions from the quantitative PCR assay(a) Undamaged duplex DNA gives rise to four daughter strands are one round of PCR. **(b)** A single photolesions (x) prevents one strand from being replicated, resulting in a two-fold reduction in amplification. **(c)** Additional photolesions on the same strand do not cause a further reduction in amplification. Thus the process must be treated statistically using the Poisson distribution. The measurable variable is the amplification ratio of irradiated DNA to unirradiated DNA, which is equivalent to the probability of no lesions occurring.

2. UV-induced DNA photodamage

We evaluated the ability of the QPCR assay to detect photolesions by investigating the damage resulting from DNA exposure to well defined dosages of 254 nm UV radiation. The exposures were conducted in a Spectroline UV oven and the amount of radiation was measured by means of a chemical actinometer. This method enabled the accurate determination of the number of photons incident on the sample by means of measuring the formation of UV sensitive product with a spectrophotometer. The amount of lesions produced by the UV light was quantified using the QPCR assay.

The rate of lesions formation dependence exhibited a well defined linear response up to a threshold exposure nearing 3 x 10^{-12}einsteins followed by a plateau of ~4 lesions/strand (Fig. 2.4). Typical error estimates on the assay lesion measurements were around 10% RSD for these exposures. The linear region is consistent with a one photon excitation process in whichabsorption is directly proportional to incident intensity. The plateau region has two explanations. It could represent the equilibrium point of lesion formation where further exposure photo-excites the reverse reaction. This type of behavior has been witnessed before in irradiations of *E. coli* plasmids where it was attributed to photosteady state [27]. Alternatively, it could be due to the dynamic range of the PCR assay, corresponding to extensive damage that resulting in amplification values to close to the minimum detection limit. Based on a dynamic range of 50:1, it would not be surprising for the assay to exhibit saturation behavior around a value of –ln(1/50) ~ 3.9 lesions/strand.

The initial linear region of the damage curve was used to determine the quantum yield of lesion formation. A linear regression was performed to obtain the slope (Fig. 2.4), which was used in conjunction with the pBR322 concentration to obtain the QY of lesion formation. We have defined Φ_D here as moles photolesions/moles photons absorbed, which does not account for thymine proximity or abundance. This represents a value indicative of the damage rates that could be realized with genomic DNA. The calculation of the UV dosage absorbed accounted for the differential absorption of 254 nm light by DNA as compared to the absorbing actinometer species [29] by the ratio of their photon absorptions using Beer's Law. These values were estimated from the optical density of the actinometer as reported by Rahn et al. (OD=200 at 254nm) and the optical density of double stranded nucleic acid (OD=1 at 260 nm for 50 ng/µL). The pathlength of irradiation was the radius (0.21216 cm) of the 20 µL hemisphere to which the sample was assumed to conform. These corrections yielded a value for the Φ_D of 7.86 (± 0.73) x 10^{-4}.

Our experimentally determined quantum yield is considerably smaller than often cited Φ_D of 0.02 [36] (determined for *E. coli* samples) but similar in magnitude to the more recent value of 1.8 x 10^{-3} determined for pBR322 by Gut *et. al*[36].This discrepancy between our reported value and previous investigations could be the result of the selectivity of our assay, since abasic sites that contribute to the lesions detected in enzyme based assays are not detected by QPCR [37], or due to the nature of the nucleic acids under investigation. Many previous investigations have employed free nucleotides in solution [36] or homo-oligmers of thymine as model systems for dimer formation [1]. These systems have the potential to overestimate the Φ_D by placing neighboring thymines in configurations that may optimize

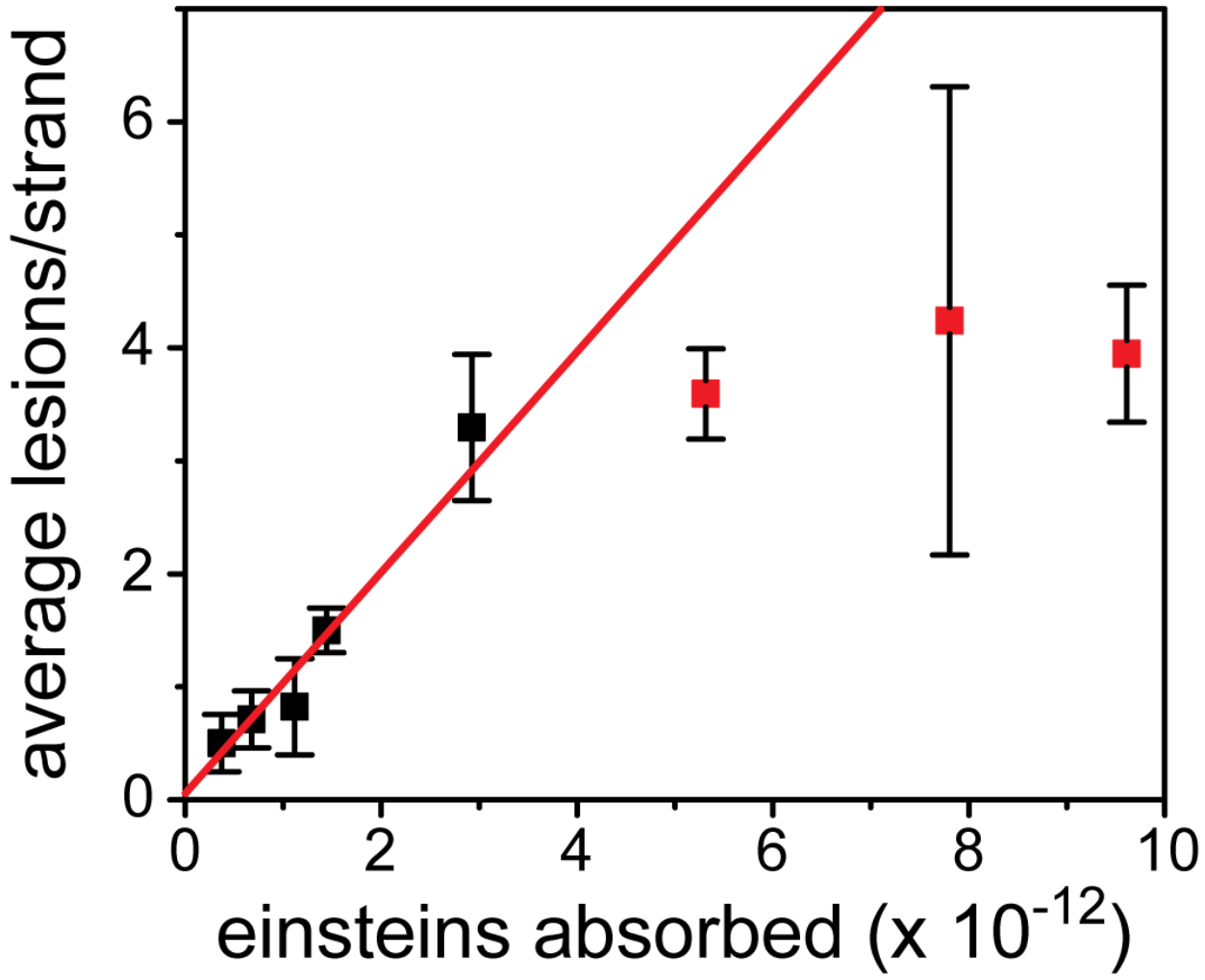

Figure 2.4- UV dose dependent lesion formation. Samples exposed to well-defined amounts of 254 nm light were analyzed with the quantitative PCR reaction. Damage produced at low dosages fit to a linear function to determine the quantum yield of lesion formation (black squared); saturation behavior was observed at higher dosages (red squares).

dimerization [25]. More native configurations of genomic DNA could reduce the possibility of dimer formation through a low abundance of thymine bases compared to chain length, spatial separation of adjacent pyrimidines along the chain, and reduction in the occurrence of favorable thymine-thymine configurations due to secondary structure formation [38, 39].

As a further comparison, if the photon absorption is scaled by the number of adjacent thymine bases (as is the case with the often cited value of 0.2) , which accounts for the number of absorption events that can lead to dimer formation, then Φ_D is revised to 0.0133, in close agreement with previous determinations. These two approaches imply different models of energy transfer along the DNA chain length following an absorption event. Assuming each nucleotide has a similar absorption cross section, the latter model (in which the quantum yield is scaled for the number of adjacent thymines) corresponds to a mechanism in which all absorption events generate an exciton that propagates along the DNA chain until neighboring thymines are encountered (thymines exhibit the highest probability for dimerization of the four nucleotides). Thus energy transfer is efficient and can potentially occur over long distances. The former calculation, in which the cross section is unscaled by neighboring thymine abundance, implies that energy transfer is very limited, with dimerization only occurring if the absorption event occurs in close proximity to neighboring thymine bases. Comparisons to work conducted with poly-thymine, which indicated a QY much higher than the value we determined, tend to suggest the latter model in which excitons can travel large distances over DNA before arriving at a thymine-thymine energy trap [1].

3. Two-photon absorption-induced DNA photodamage

Based on the two-photon absorption cross section data of fluorescent molecules, it is possible to achieve two-photon absorption in the wavelength region corresponding to twice the one-photon absorption spectrum for many species, but the two-photon absorption maximum can be blue-shifted relative to twice the one-photon absorption maximum [19]. This line of reasoning predicts that DNA could exhibit two-photon absorption of visible light between 400-600 nm, with a maximum at or below 520 nm. The near-infrared pulses produced by a femtosecond Ti:sapphire laser can be frequency-doubled to produce visible pulses in this wavelength range with peak intensities sufficiently high to achieve two-photon absorption in molecules with a reasonable cross section. Therefore, we decided to irradiate DNA with focused femtosecond visible pulses and employ the QPCR assay to quantify the extent of two-photon absorption-induced damage. To produce a realistically detectable amount of damaged DNA, we employed our multiphoton microscope setup to scan the focused beam through a small volume of DNA solution. We chose a 10X 0.3NA objective lens to maximize the scan area, with the realization that the beam waist is larger than is typically used for high-resolution imaging. This required us to irradiate samples with higher average laser powers than are used in multiphoton microscopy.

We used irradiation wavelengths of 425 nm, 450 nm, and 475 nm. These wavelengths were chosen to maximize power available after frequency-doubling our Ti:sapphire laser output, since we determined that the incident power needed to be greater than 100 mW in order to obtain detectable amounts of damage. Within experimental error, the rate of lesion formation had a quadratic dependence on the incident power (Fig. 2.5), a defining characteristic of two-photon absorption. The RSD on the control samples of the PCR process

was found to be less than 8%, with the absolute error ± 0.62 Lesions/Strand. The most parsimonious model to describe the power dependent damage curves was determined using the F-test. This allowed the effect of higher order regression models to be distinguished, which confirmed the data was best modeled by a quadratic expression.

After confirming that two-photon absorption can lead to damage, we calculated the two-photon absorption cross section of DNA from our data. Our method for calculating this quantity was derived from a basic definition of the two-photon cross section, in which the number of photons absorbed per nucleotide per second, NA_2, is proportional to the product of the two-photon cross section of lesion formation and the square of the intensity. Our experimental observables are related to the time-averaged quantities:

$$\langle NA_2 \rangle = \sigma_2 \langle I^2 \rangle \tag{3}$$

The value of NA_2 can be calculated from the observed number of lesions per strand by:

$$\langle NA_2 \rangle = \frac{Lesions/_{Strand}}{\Phi_D * T * n} \tag{4}$$

Where Φ_D is the quantum yield of lesion formation (assumed to be the same for both a one or two photon process), *T* is the interaction time of the absorber and the incident light, and *n* is the number of nucleotides per strand of pBR322. The value of *T* was estimated by dividing the diameter of the beam (2*ω), where ω is the beam radius, by the scan rate of the raster laser beam, s.

Given the important dependence of the peak laser power and two-photon cross sections on the beam profile, we measured the point spread function of the beam at 450 nm by

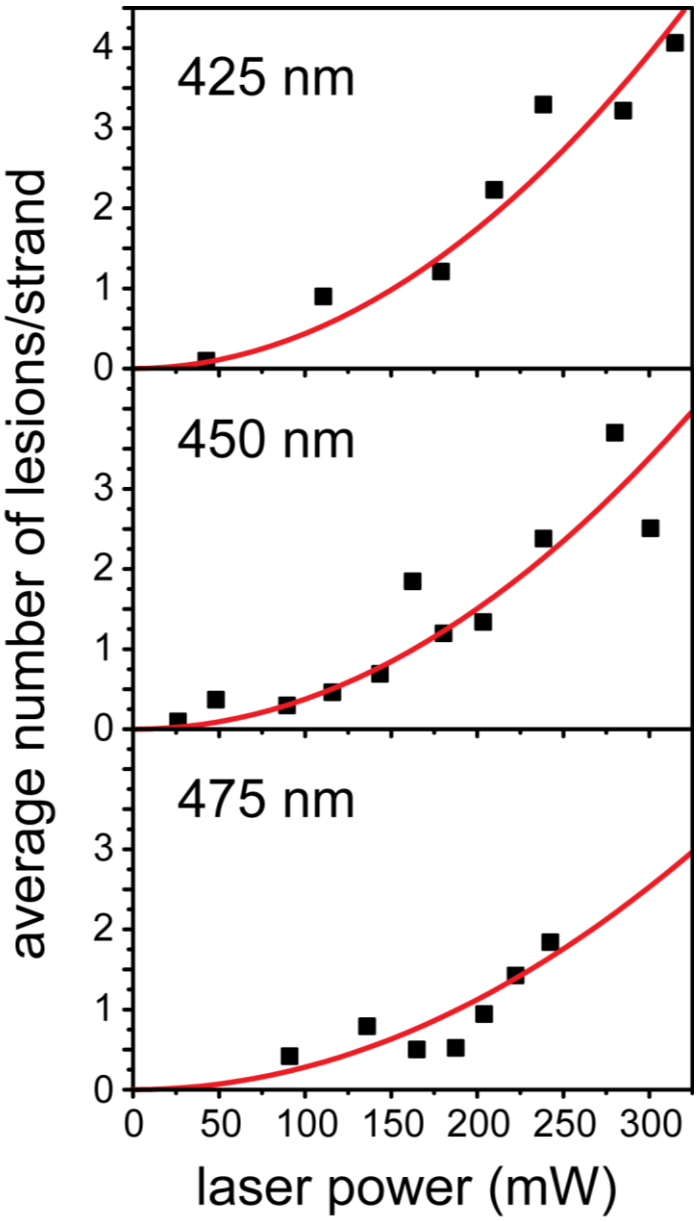

Figure 2.5- Power-dependent damage produced by irradiation of DNA samples with focused femtosecond pulses at 425 nm, 450 nm and 475 nm. The data were fit to a quadratic function of power to determine the two-photon absorption cross-section, as described in the text.

imaging fluorescent beads. The beam FWHM was found to be 1.77±0.27 μm. This value is larger than the minimum (diffraction-limited) beam profile because we chose to slightly underfill the objective lens back aperture (to achieve maximal laser power at the sample) and because of a slight beam asymmetry introduced by the frequency doubling crystal. The beam radius at 450 nm was determined to be 1063 nm, and the ratio of the theoretical to measured beam radius at 450 nm was used to estimate the 425 nm and 475 nm beam radii to be 1004 nm and 1122 nm respectively.

The relationship between $\langle I^2 \rangle$ for a pulsed laser source and the average power is given by [19]:

$$\langle I^2 \rangle = \frac{g_p \langle I \rangle^2}{R * \tau_p} = \frac{g_p \langle P \rangle^2}{R * \tau_p (\pi * \omega^2)^2} \quad (5)$$

Where g_p is the temporal laser pulse shape (assumed to be Gaussian-Lorentzian, for which g_p = 0.66), R is the laser repetition rate, and τ_p is the pulse duration. Equation 3 can be rewritten by combining Equations 4 and 5 to yield:

$$\frac{Lesions}{Strand} = \sigma_2 \left(\frac{2 * g_p * \Phi_D * n}{\pi^2 * R * \tau_p * \omega^3 * s} \right) \langle P \rangle^2 \quad (6)$$

(This expression includes a conversion from power expressed in watts to photons/s.) The two-photon absorption cross section can thus be determined by equating the quadratic coefficient in Equation 6 to the coefficients from the quadratic fits of the data.

The two-photon cross section of DNA per nucleotide extracted from our data is 2.58 (± 0.47) GM at 425 nm, 2.36 (± 0.46) GM at 450 nm, and 1.86 (± 0.48) GM at 475 nm (where 1 GM = 10^{-50} cm^4 s $photon^{-1}$). Our measured values are an order of magnitude larger than the

previously reported two-photon absorption cross section of pBR322 at the longer wavelength of 532 nm, which was 0.06 GM [25]. The discrepancy could be due to the increased sensitivity of our assay in comparison to the previous study. Regardless, these data indicate a trend of increasing two-photon absorption as the incident wavelength is blue shifted away from twice the 260 nm one-photon absorption maximum of DNA, which may be consistent with the blue shift noticed in the absorption spectra of many fluorphores under two-photon excitation. Alternately, it may indicate that the transitions excited by two-photon absorption in this wavelength region are similar to those excited by far-UV light. Overall this value is low for small aromatic compounds. As a common reference, rhodamine 6G has a maximal two-photon cross section of ~150 GM, while our two-photon cross section was only ~2 GM [41].

It is important to note that our experiments used a low NA objective to irradiate the largest area possible, thus requiring high average laser powers. This would not be the case in a typical high resolution imaging experiment. Given that the maximum average power output of our laser system was 300 mW, and using the measured beam profile at 450 nm, our peak irradiance used to induce two-photon photodamage was approximately 0.704 TW/cm^2. In contrast, a typical cell imaging system would employ a higher NA objective lens for maximum resolution, which would generate similar peak irradiances at 4-5 mW (assuming 1.2NA).

Having characterized the two-photon damage at several wavelengths, it is useful to illustrate the differences that exist between our *in vitro* system and a more typical *in vivo syst*em. We consider our experiments to be a measure of direct two-photon damage, as our system is free of potential energy-transferring DNA sensitizers – both endogenous ond

exogenous. Using high peak irradiance levels (~ 0.7 TW/cm^2), we achieved only minimal DNA damage, while previous groups conducting *in vivo* experiments using three-photon irradiation to generate DNA damage have reported saturating amounts of damage at lower irradiation levels. In a study conducted by Meldrum *et al.*, *three*-photon irradiation of cells at 750 nm lead to *saturation* of photolesions formation at approximately 0.2 TW/cm^2[21]. Another three-photon absorption study performed by Trautlein*et al.*[22], in which cells were irradiated at 750 nm and 1050 nm, found *saturation* of photolesion formation at 0.3 TW/cm^2 and 0.9 TW/cm^2 respectively. Finally, work by Dinant*et al.*[20], found that similar CPD damage was created when cells were irradiated under three-photon conditions at an average power of 80 mW (estimated intensity ~ 3.5 TW/cm^2) as when cells sensitized with Hoechst dye were irradiated under single photon conditions at 405 nm at 18 mW. Since the cited studies used three-photon absorption, much greater photon intensities should have been required than the two-photon experiments we conducted. Given the intensity levels they required to induce maximal damage *in vivo*, we suspect that the cellular environment can contain endogenous sensitizers that mediate energy transfer to DNA, promoting lesion formation. Significantly, some of these studies employed cells that express GFP fusion proteins, which, given the high near-UV absorption cross section of GFP, may act as an exogenous sensitizing agent. This possible mechanism of damage sensitization could be analogous to the chromophore-assisted laser inactivation (CALI) method used to abolish protein function *in vivo*. This would be especially relevant if the fluorescently tagged proteins were often in close association with DNA [42], thus placing the DNA within the ~60 Å damage radius of the reactive oxygen species generated during GFP-excitation [43]. Thus, it is important to consider the presence of possible absorbing species, since cells can be naturally

sensitized by endogenous chromophores or deliberately sensitized through the incorporation of light absorbing species or intercalating dyes. Additionally, the *in vivo* studies detected photolesions through immunostaining methods [20, 22], which presumably require much higher levels of DNA damage to be detected than our QPCR assay; this further highlights the difference in damage levels obtained between our methods.

Finally, it is important to note that at high laser intensities, DNA damage can be induced indirectly by radical species generated through optical breakdown of the aqueous solvent, as opposed to the direct formation of photoproducts by two-photon absorption. However, we do not believe this mechanism is a likely explanation for the photodamage we observed experimentally. Most often, the optical breakdown of water is reached by multiphoton absorption of ~4.6 eV UVB (266 nm) photons at ~TW/cm^2 intensity levels. In an investigation by Fan *et al*., the optical ionization threshold for water has been estimated to lie between 6.5-10 eV[44], and is thus readily reached through a two-photon process involving UV light. In our experiments, the highest energy wavelength used was 425 nm, corresponding to 2.9 eV, which is below the ionization threshold if two-photon absorption is to be considered. While higher order multiphoton absorption of 425 nm light could lead to optical breakdown, this should manifest itself as a tertiary or higher order power dependence of DNA lesion formation with respect to power, a trend that was not observed in the course of our experiments. Further, in the same study [44], the ability to reach optical breakdown in water was determined for a similar irradiation system, a 580 nm laser operating with a pulse duration of 100 fs. They found that the minimum laser intensities to achieve ionization is 11.1 TW/cm^2. This system closely models

our apparatus (with a pulse duration of 140 fs) and indicated that our maximum laser intensity of ~ 0.7 TW/cm^2 is insufficient to produce solvated electrons in the visible.

Conclusion

We developed a sensitive PCR assay for DNA damage, and determined the quantum yield for one-photon DNA photodamage at 254 nm. Irradiation of DNA with focused, ultra-short visible pulses yielded a second order dependence of photolesion formation on incident light intensity, confirming the ability of two-photon absorption to cause UV-like photochemical damage. The two-photon absorption cross section of DNA was determined to vary from 2.6 – 1.9 GM in the range of 425 – 475 nm. Further research that extends the range of irradiation wavelengths will be required to determine the full multiphoton absorption spectrum of DNA.

REFERENCES

(1) Patrick, M. H.; Rahn, R. O. *Photochemistry and Photobiology of Nucleic Acids;* Academic Press: New York, 1976; Vol. II, pp 35-96.

(2) Gorner, H. *J. Photochem. Photobiol. B, Biol.* **1994***, 26*, 117-139.

(3) Cremer, C.; Cremer, T.; Fukuda, M.; Nakanishi, K. *Hum Genet* **1980***, 54*, 107-110.

(4) Essers, J.; Vermeulen, W.; Houtsmuller, A. B. *Curr. Opin. Cell Biol.* **2006***, 18*, 240-246.

(5) Mone, M. J.; Volker, M.; Nikaido, O.; Mullenders, L. H. F.; Zeeland, A. A. v.; Verschure, P. J.; Manders, E. M. M.; Driel, R. v. *EMBO Rep.* **2001***, 2*, 1013-1017.

(6) Mone, M. J.; Bernas, T.; Dinant, C.; Goedvree, F. A.; Manders, E. M. M.; Volker, M.; Houtsmuller, A. B.; Hoeijmakers, J. H. J.; Vermeulen, W.; Driel, R. v. *Proc. Natl. Acad. Sci. U. S. A.* **2004***, 101*, 15933-15937.

(7) Katsumi, S.; Kobayashi, N.; Imoto, K.; Nakagawa, A.; Yamashina, Y.; Muramatsu, T.; Shirai, T.; Miyagawa, S.; Sugiura, S.; Hanaoka, F.; Matsunaga, T.; Nikaido, O.; Mori, T. *J. Invest. Dermatol.* **2001***, 117*, 1156-1161.

(8) Imoto, K.; Kobayashi, N.; Katsumi, S.; Nishiwaki, Y.; Iwamoto, T.; Yamamoto, A.; Yamashina, Y.; Shirai, T.; Miyagawa, S.; Dohi, Y.; Sugiura, S.; Mori, T. *J. Invest. Dermatol.* **2002***, 119*, 1177-1182.

(9) Limoli, C. L.; Ward, J. F. *Radiat. Res* **1993***, 134*, 160-169.

(10) Rogakou, E. P.; Boon, C.; Redon, C.; Bonner, W. M. *J. Cell Biol.* **1999***, 146*, 905-915.

(11) Lukas, C.; Falck, J.; Bartkova, J.; Bartek, J.; Lukas, J. *Nat. Cell Biol.* **2003***, 5*, 255-260.

(12) Lukas, C.; Melander, F.; Stucki, M.; Falck, J.; Bekker-Jensen, S.; Goldberg, M.; Lerenthal, Y.; Jackson, S. P.; Bartek, J.; Lukas, J. *EMBO Journal* **2004***, 23*, 2674-2683.

(13) Bekker-Jensen, S.; Lukas, C.; Melander, F.; Bartek, J.; Lukas, J. *J. Cell Biol.* **2005***, 170*, 201-211.

(14) Bekker-Jensen, S.; Lukas, C.; Kitagawa, R.; Melander, F.; Kastan, M. B.; Bartek, J.; Lukas, J. *J. Cell Biol.* **2006***, 173*, 195-206.

(15) Lan, L.; Nakajima, S.; Oohata, Y.; Takao, M.; Okano, S.; Masutani, M.; Wilson, S. H.; Yasui, A. *Proc. Natl. Acad. Sci.* **2004***, 101*, 13738-13743.

(16) Kruhlak, M. J.; Celeste, A.; Dellaire, G.; Fernandez-Capetillo, O.; Muller, W. G.; McNally, J. G.; Bazett-Jones, D. P.; Nussenzweig, A. *J. Cell Biol.* **2006***, 172*, 823-834.

(17) Lakowicz, J. R.; Gryczynski, I. *Nonlinear and Two-Photon Induced Fluorescence;* Topics in Fluorescence Spectroscopy; Plenum Press: New York, 1997; Vol. 5.

(18) Denk, W.; Strickler, J. H.; Webb, W. W. *Science* **1990***, 248*, 73-76.

(19) Zipfel, W. R.; Williams, R. M.; Webb, W. W. *Nat. Biotechnol.* **2003***, 21*, 1369-1377.

(20) Dinant, C.; Jager, M. d.; Essers, J.; Cappellen, W. A. v.; Kanaar, R.; Houtsmuller, A. B.; Vermeulen, W. *J. Cell Sci.* **2007***, 120*, 2731-2740.

(21) Meldrum, R. A.; Botchway, S. W.; Wharton, C. W.; Hirst, G. J. *EMBO Rep.* **2003***, 4*, 1144-1149.

(22) Trautlein, D.; Deibler, M.; Leitenstorfer, A.; Ferrando-May, E. *Nucleic Acids Res.* **2010***, 38*, e14.

(23) Elf, J.; Li, G.; Xie, X. S. *Science* **2007***, 316*, 1191-1194.

(24) Slade, K. M.; Baker, R.; Chua, M.; Thompson, N. L.; Pielak, G. J. *Biochemistry (Washington)* **2009***, 48*, 5083-5089.

(25) Hefetz, Y.; Dunn, D. A.; Deutsch, T. F.; Buckley, L.; Hillenkamp, F.; Kochevar, I. E. *J. AM. CHEM. SOC.* **1990***, 112*, 8528-8532.

(26) Douki, T.; Court, M.; Sauvaigo, S.; Odin, F.; Cadet, J. *J. Biol. Chem.* **2000***, 275*, 11678-11685.

(27) Wulff, D. *Biophysical Journal* **1963***, 3*, 355-362.

(28) Chen J., Kadlubar F. F.,Chen J.Z. *Nucleic Acids Res.* **2007**, 1-12.

(29) Rahn, R. O.; Stefan, M. I.; Bolton, J. R.; Goren, E.; Shaw, P.; Lykke, K. R. *Photochem. Photobiol.* **2003***, 78*, 146-152.

(30) Rahn, R. O. *Photochem. Photobiol.* **1997***, 66*, 450-455.

(31) Kalinowski, D. P.; Illenye, S.; Houten, B. V. *Nucleic Acids Res.* **1992***, 20*, 3485-3494.

(32) Ayala-Torres, S.; Chen, Y.; Svoboda, T.; Rosenblatt, J.; Houten, B. V. *Methods* **2000***, 22*, 135-147.

(33) Santos, J. H.; Meyer, J. N.; Mandavilli, B. S.; Houten, B. V. In *Quantitative PCR-based measurement of nuclear and mitochondrial DNA damage and repair in mammalian cells;* Henderson, D. S., Ed.; Methods in Molecular Biology, vol. 314: DNA Repair Protocols; Humana Press: Totowa, NJ, 2006; Vol. 314, pp 183-199.

(34) Ponti, M.; Forrow, S. M.; Souhami, R. L.; D'Incalci, M.; Hartley, J. A. *Nucleic Acids Res.* **1991***, 19*, 2929-2933.

(35) Boas, M. L. *Mathematical Methods in the Physical Sciences;* John Wiley & Sons, Inc.: 1983; , pp 793.

(36) Gut, I. G.; Hefetz, Y.; Kochevar, I. E.; Hillenkamp, F. *J Phys. Chem.* **1993***, 97*, 5171-5176.

(37) Jiang, Y.; Rabbi, M.; Kim, M.; Ke, C.; Lee, W.; Clark, R.; Mieczkowski, P. A.; Marszalek, P. E. *Biophysical Journal* **2009***, 96*.

(38) Douki, T. *Journal of Photochem. Photobiol. B: Biology* **2006***, 82*, 45-52.

(39) Schreier, W. J.; Schrader, T. E.; Koller, F. O.; Gilch, P.; Crespo-Hernandez, C. E.; Swaminathan, V. N.; Carell, T.; Zinth, W.; Kohler, B. *Science* **2007***, 315*, 625-629.

(40) Lakowicz, J. R., Ed.; In *Principles of Fluorescence Spectroscopy;* Springer: New York, 2006; .

(41) Albota, M., Xu, C., Webb,W.W. *Applied Optics* **1998***, 37*.

(42) Tanabe, T.; Oyamada, M.; Fujita, K.; Dai, P.; Tanaka, H.; Takamatsu, T. *Nature Methods* **2005***, 2*, 503-505.

(43) Liao, J. C.; Roider, J.; Jay, D. G. *PROC. NATL. ACAD. SCI. U.S.A.* **1994***, 91*, 2659-2663.

(44) Fan., C. H., Sun., J., Longtin, J.,P. *Journal of Applied Physics* **2002***, 91*, 2530-2536.

CHAPTER 3

QUANTIFICATION OF DYE-MEDIATED PHOTODAMAGE DURING SINGLE-MOLECULE DNA IMAGING

"Bad times have scientific value. These are occasions a good learner would not miss"

-Ralph Waldo Emerson

Overview:

Single-molecule fluorescence imaging of DNA-binding proteins has enabled detailed investigations of their interactions. However, the intercalating dyes used to visually locate DNA molecules have the undesirable effect of photochemically damaging the DNA through radical intermediaries. Unfortunately, this damage occurs as single-strand breaks (SSBs), which are visually undetectable but can heavily influence protein behavior. We investigated the formation of SSBs on DNA molecules by the dye YOYO-1 using complementary single-molecule imaging and gel electrophoresis based damage assays. The single-molecule assay imaged hydrodynamically elongated lambda DNA, enabling the real-time detection of double-strand breaks (DSBs). The gel assay, which used supercoiled plasmid DNA, was sensitive to both SSBs and DSBs. This enabled the quantification of SSBs that precede DSB formation. Using the parameters determined from the gel-damage assay, we applied a model of stochastic DNA damage to the time-resolved DNA breakage data, extracting the rates of single-strand breakage

at two dye staining ratios and measuring the damage reduction from the radical scavengers ascorbic acid and β-mercaptoethanol. These results enable the estimation of the number of SSBs that occur during imaging and are scalable over a wide range of laser intensities used in fluorescence microscopy.

Introduction

Investigations of DNA-binding proteins and their substrate interactions have benefited greatly from the level of detail afforded by single molecule imaging (SMI) techniques. It has recently become possible to interrogate nonspecific protein-DNA interactions, which are difficult to study with bulk experiments, by directly observing the interaction of individual proteins with immobilized strands of DNA using high resolution optical microscopy [1]. In the majority of such experiments, the proteins of interest are fluorescently labeled and tracked relative to a DNA substrate that is located by the use of intercalating dyes. Paramount among the assumptions made in such experiments is that the protein-DNA interaction under consideration is not perturbed by the presence of the intercalating dye [2]. While in some cases this can be verified experimentally [3], the photochemical effect of the dye on the nucleic acid substrate is often neglected.

Once excited, fluorescent dyes may undergo intersystem crossing and interact with ground state oxygen molecules, generating highly reactive singlet oxygen and fluorophore radicals [4]. These damaging species can attack DNA to produce various forms of oxidative radical photodamage, including strand breaks [5, 6]. Individual damage events typically cleave only one strand of the DNA sugar-phosphate backbone [7, 8]; the accumulation of many single-

strand breaks (SSBs) leads to double-strand cleavage [9]. Since many proteins involved in DNA replication and repair bind to single-stranded DNA [10-12], the presence of SSBs induced by photoexcitation of intercalating dyes could strongly bias protein-DNA interactions. Unfortunately, SSBs cannot be visualized directly in SMI experiments, so their impact on the observed protein dynamics is often assumed to be minor in the absence of significant double-strand cleavage.

In light of the deleterious effect SSBs may have on protein-DNA interactions, we have investigated the rate of single-strand photocleavage in SMI experiments. Since it is not possible to detect this form of damage using SMI, we conducted parallel SMI and ensemble experiments under similar conditions to fully assess photoexcited dye-induced DNA cleavage. The single-molecule experiments employed a fluorescent microscope to image the double-strand photocleavage of flow-stretched DNA substrates tethered on a passivated glass slide in a microfluidic flow cell. The ensemble experiments used a traditional gel electrophoresis assay to quantify both single-strand and double-strand photocleavage of a plasmid DNA sample that had been irradiated by an unfocused laser beam. In both sets of experiments, the DNA was labeled with the intercalater YOYO-1, a cyanine dye with extremely high DNA affinity [13], and was irradiated by the same 488 nm laser. Additionally, the ability of two radical scavenging systems to protect DNA from photodamage was investigated.

The SMI and ensemble experiments both quantified the rate of double strand cleavage, allowing us to establish that the rate of damage measured by the ensemble assay could be extrapolated to the SMI experiments by adjusting for the excitation light flux. However, the

ensemble gel electrophoresis assay was required to monitor the formation of SSBs since this quantity is undetectable in the SMI experiments. We fit the results of both experiments using a kinetic model that assumes SSBs are distributed randomly along the DNA strand and double-strand cleavage occurs only when two SSBs are sufficiently close. Based on the cleavage rates extracted from the fits, we are now able to estimate the number of SSBs that may arise from typical SMI irradiation conditions. We conclude that these conditions produce a significant amount of DNA photodamage, which has been largely neglected but should be considered to properly interpret the outcome of these experiments. This conclusion is particularly important for research concerning DNA repair enzymes, in which damage sites are hypothesized to act as surface energy minima, trapping enzymes until the next step in the repair pathway.

Materials & Methods

All chemicals and materials are Fisher Brand unless otherwise noted.

1. *Observing double-strand photocleavage using flow-stretched DNA*

    a. *Surface functionalization, microfluidic chamber fabrication, and DNA substrate preparation*

The procedure to passivate and functionalize the coverslip for DNA attachment was based on a previously reported method [14]. Coverslips were ethanol rinsed, sonicated in chloroform for 5-minutes, ethanol rinsed again, dried, and soaked in Piranha solution (1/3 20% hydrogen peroxide, 2/3 sulfuric acid) for a minimum of 30-min. Following drying for 1-hr at 150°C to remove physisorbed water, the glass was submerged in aminopropyltriethoxysilane solution (10% APTES/90% anyhydrous acetone) for 10-minutes with agitation. After heating to

110°C to cure the surface, the coverslips were coated with a solution containing methoxy poly(ethylene glycol) succinimidyl-valerate (mPEG-SVA) and biotin-PEG-SVA (Laysan Bio.) dissolved in coupling buffer ($NaHCO_3$, 100 mM, pH 8.2). The solution composition was biotin-PEG (2 mg)/mPEG (150 mg) in coupling buffer (1 mL), with 100 µL deposited on each coverslip. This solution was incubated on the coverslips for 5-7 hours to allow adequate time to couple the PEG-SVA to the APTES layer.

For microfluidic chamber fabrication, quartz microscope slides (Finkenbeiner Glass Inc.) were drilled and Nanoport tubing connectors (Upchurch Scientific) affixed to the slide with quick-dry epoxy. Double-stick tape (3M) was used as spacer between the slide and PEGylated coverslip surface, with sealant (rapid-dry nail polish) applied to the edges of the coverslip to prevent leaks.

The DNA substrates were prepared from 48 kbp-lambda DNA (Promega) by sticky-end filling with biotinylated-dUTP using the Klenow fragment of DNA polymerase I (3'→5' exo^-, New England BioLabs), following a protocol from Smith et al. [15]. Lambda DNA stock was diluted in polymerase buffer (0.17 µg/µL, 115 µL total volume) and heated to 65°C for 10 minutes to melt the sticky ends. The dNTP solutions were added (6.4 µL each, 10 µM) followed by the addition of the Klenow fragment (10 Units). The sample was heated to 37°C for 30 minutes, followed by heat inactivation at 75°C for 20 min. The DNA was purified by dialysis in TE buffer (pH 8) for 24 hr.

b. DNA staining and injection for SMI

Prior to DNA injection into the flow cell, the sample was stained by mixing YOYO-1 dye (Invitrogen) with the biotinylated lambda at a specific dye to nucleotide molar ratio of 1:4 or 1:10. The DNA was used at a concentration of 10 pM. At all times, solutions containing the DNA substrates were handled with wide-bore pipette tips to reduce the incidence of shearing. These experiments used dye to nucleotide ratios of 1:4 and 1:10, and three working buffer systems adjusted within pH 7.7±0.2: TE (10 mM Tris, 1 mM EDTA), TE/β-mercaptoethanol (5% v/v) , and TE/ascorbic acid (10 mM).

After conditioning a flow cell chamber with blocking buffer (4 mM Tris-HCl, 0.1 mM EDTA, 0.2 mg/mL BSA) 300 µL of the DNA solution was injected into the flow cell at a rate of 25 µL/min using a KD Systems Syringe pump. Solutions entered the flow cell by withdrawal from a reservoir sealed to one of the Nanoports (a large pipette tip glued to the Nanoport). Unbound DNA was removed from the chamber by flushing with soaking buffer (10 mM Tris-HCl, 1 mM EDTA, 10 mM NaCl) at a rate of 40 µL/min for 5 mins. Generally, the DNA sample concentration was sufficient to bind between 10-50 spatially resolved DNA strands per field of view (136 x 136 µm). At this point buffer flow was terminated and imaging experiments could begin.

c. Single-molecule imaging

All experiments were carried out on a homebuilt inverted microscope. Samples were excited by a 488 nm diode Coherent sapphire laser, the output of which was focused onto the back aperture of an Olympus 60X/1.2NA water immersion objective lens configured for wide-field imaging. The laser power was controlled by neutral density filters. The fluence at the

objective was determined by restricting the illumination field with an iris, measuring the field diameter with a Ronchi ruling slide (Edmund Optics), and measuring the laser power with a calibrated power meter. The iris was opened for imaging all DNA samples. The emission from the YOYO-1 was filtered with a 575/150 nm bandpass filter (Chroma) before detection by an EMCCD camera (Hamamatsu ImageEM, model C9100-13).

For each experiment, the buffer flow was terminated, the tethered DNA substrates were brought into focus, the beam blocked to prevent photodamage prior to observation, and a nearby location selected at random to begin imaging. Imaging of the selected region was continued until the majority of the elongated strands in the field of view had been cleaved. The image collection frame rate was varied to match the timescale of the cleavage events, typically between 30-2 Hz for collections between 10 to 300 secs. For each buffer-sample condition two replicate flow cells were tested, with 6-11 replicate regions imaged at each laser intensity.

d. Radical scavenger buffer preparation

For each radical scavenger tested, the scavenger was added to the working buffer and pH-adjusted. Ascorbic acid was used at 10 mM, while β-mercaptoethanol (BME, sealed ampoules) was used at 5 % (v/v) immediately before irradiation. The same scavenger concentrations were used for the ensemble studies.

2. Single-molecule image processing

MATLAB programs (Appendix B) were written in-house to determine the number of intact DNA molecules in each frame of the SMI experiments. The first frame (t = 0 seconds) of

each image sequence was used to identify the regions that initially contain extended DNA strands. The initial image was blurred with an asymmetric Gaussian filter and subtracted from itself. After performing this background subtraction the image was slightly smoothed using an asymmetric Gaussian filter extended along the direction of flow. The image was then thresholded and skeletonized, a morphological process that reduces objects to a single connected line of pixels, thus rendering the elongated DNA strands as lines. The result was then overlaid in false color with the image so that the user could manually select which features corresponded to DNA strands. All well-spaced DNA strands were selected and subsequent image processing steps were automated. The skeletons of selected features were dialyzed to create a binary mask, which was applied to every subsequent frame before thresholding. The number of objects longer than 12 μm was recorded for each thresholded frame to produce breakage curves.

3. Ensemble DNA damage assay

a. Bulk DNA sample preparation

The plasmid pBR322 DNA (New England BioLabs) was diluted in TE working buffer (10 mM Tris-HCl, 1 mM EDTA, pH 7.5) to a concentration of 20 ng/μL. The dye YOYO-1 (Invitrogen) was added to the diluted plasmid to a final concentration of 3.14 or 1.26 μM. The YOYO-1 molar concentrations correspond to dye to nucleotide ratios of 1:4 or 1:10. For experiments using YO-PRO1 dye, the concentration of the dye was doubled to maintain equivalent molar ratios of fluorophores (YO-PRO1 to nucleotide molar ratios were therefore 1:2 and 1:5).

b. Bulk sample irradiation

Irradiations were performed by placing a 60 µL sample on a coverslip and illuminating it from above by a 488-nm diode laser beam. The laser power was measured prior to each trial with a calibrated power meter. The beam was expanded to 2 cm in diameter to fully encompass the sample with a uniform intensity of 4.7×10^{-3} W/cm^2. To ensure that the sample drop did not absorb enough light intensity to provide a protective effect to the lower fluid layer, the absorbance through the drop was estimated. Given the extinction coefficient of YOYO-1 dye (98,900 $M^{-1}cm^{-1}$ as reported by Invitrogen) and an approximated sample drop thickness of 0.306 cm (assuming a hemispherical drop of 60 µL), the minimum laser light transmittance would be approximately 80% , sufficiently low to ensure the drop thickness was uniformly irradiated. Irradiations ranged from 0.5 to 30.0 mins, with a new sample drop irradiated for each time point. Three control samples were included for every set of conditions tested- native plasmid, plasmid that was stained with dye but not irradiated, and plasmid that was irradiated but not stained with dye.

c. Gel electrophoresis

All samples were run in triplicate on a 1.2% agarose gel in 1x TAE buffer at 3.5 V/cm for 5 hours. To ensure consistent quantification of samples illuminated for various times, a YOYO-1 destain step was performed after electrophoresis [16, 17]. This dye was removed by washing the gel in 1 L of destain buffer (0.1 M NaCl, 10 mM SDS) for 5 hours, followed by 2 L of 1x TAE for 16 hours to remove SDS. The gel was then stained with ethidium bromide and imaged.

4. Ensemble damage assay: Ascorbic acid mediated DNA degradation

The potential of ascorbic acid solutions to mediate strand breakage of DNA was investigated with pBR322 DNA(1 µL, 1µg/mL) diluted in TE/ascorbic acid buffer (50 µL, 10 mM, pH 7.5). The dilutions were performed at 10-minute intervals for 120 minutes before separation by gel electrophoresis, to determine the time-dependence of breakage. No dye was used in this study.

5. Gel quantification

Analysis of the gel images to quantify the relative fraction of supercoiled, nicked and linear DNA in each lane was performed using MATLAB scripts written in-house (Appendix A). One-dimensional Gaussian functions were locally fit to each peak from the pixel intensities of line scans of each lane. A linearly slopping baseline was included to account for non-uniformities in the background. The relative fraction of each component was determined from the area of the Gaussian fits; the area of the nicked and linear forms was multiplied by a correction factor of 0.8 [18] to correct for the preferential affinity of ethidium bromide for relaxed DNA.

Results

1. Double-strand photocleavage of individual DNA molecules

We applied SMI to observe double-strand photocleavage of individual dye-labeled DNA molecules over time. Lambda DNA molecules were end-labeled with biotin, stained with YOYO-1 intercalating dye, and injected into a microfluidic flow cell whose surfaces had been functionalized with a PEG/biotin-PEG layer. The PEG coating minimizes nonspecific electrostatic interactions between the DNA and the glass, while the relatively small

subpopulation of biotin-terminated PEG provides binding sites for the DNA termini through biotin-streptavidin-biotin linkages. The hydrodynamic flow during injection elongates DNA molecules that are initially bound to the surface at only one end, causing them to become double-tethered in an extended configuration. This arrangement allows us to image the photocleavage of each molecule by applying laser-based widefield microscopy in the absence of buffer flow, eliminating potential complications caused by hydrodynamic tension. Fig. 3.1-a displays images recorded at several times during the course of a typical dataset, with the time resolution determined by the image capture rate. The initial image consists of vertical lines that correspond to double-tethered DNA strands and smaller spots that are DNA strands which have been mechanically sheared during handling. Individual DNA molecules accumulate SSBs as a result of irradiation, but this form of damage is not visually perceptible in the images. However, the accumulation of at least two single photocleavage events that break opposite DNA strands in sufficiently close proximity leads to double-strand cleavage, which transforms a vertical line into two points in subsequent frames, as seen in Fig. 3.1-d. We analyzed each image sequence using an automated MATLAB script that counts the number of extended DNA strands in each frame, resulting in time-dependent breakage curves that describes the dataset (Fig. 3.1-b). The breakage curves were then fit with a kinetic damage model to determine the rate of single strand breakage, vide infra (Fig. 3.1-b, solid lines).

We applied this SMI procedure to characterize double-strand DNA photocleavage under a variety of conditions, including two dye-staining ratios and several buffer compositions. A qualitative inspection of the breakage curves (Fig. 3.1-b) indicates that SSBs accumulate rapidly,

as shown by the early onset of double-strand cleavage in the tethered samples, and the rate of DNA photodamage is strongly dependent on incident laser intensity and intercalater

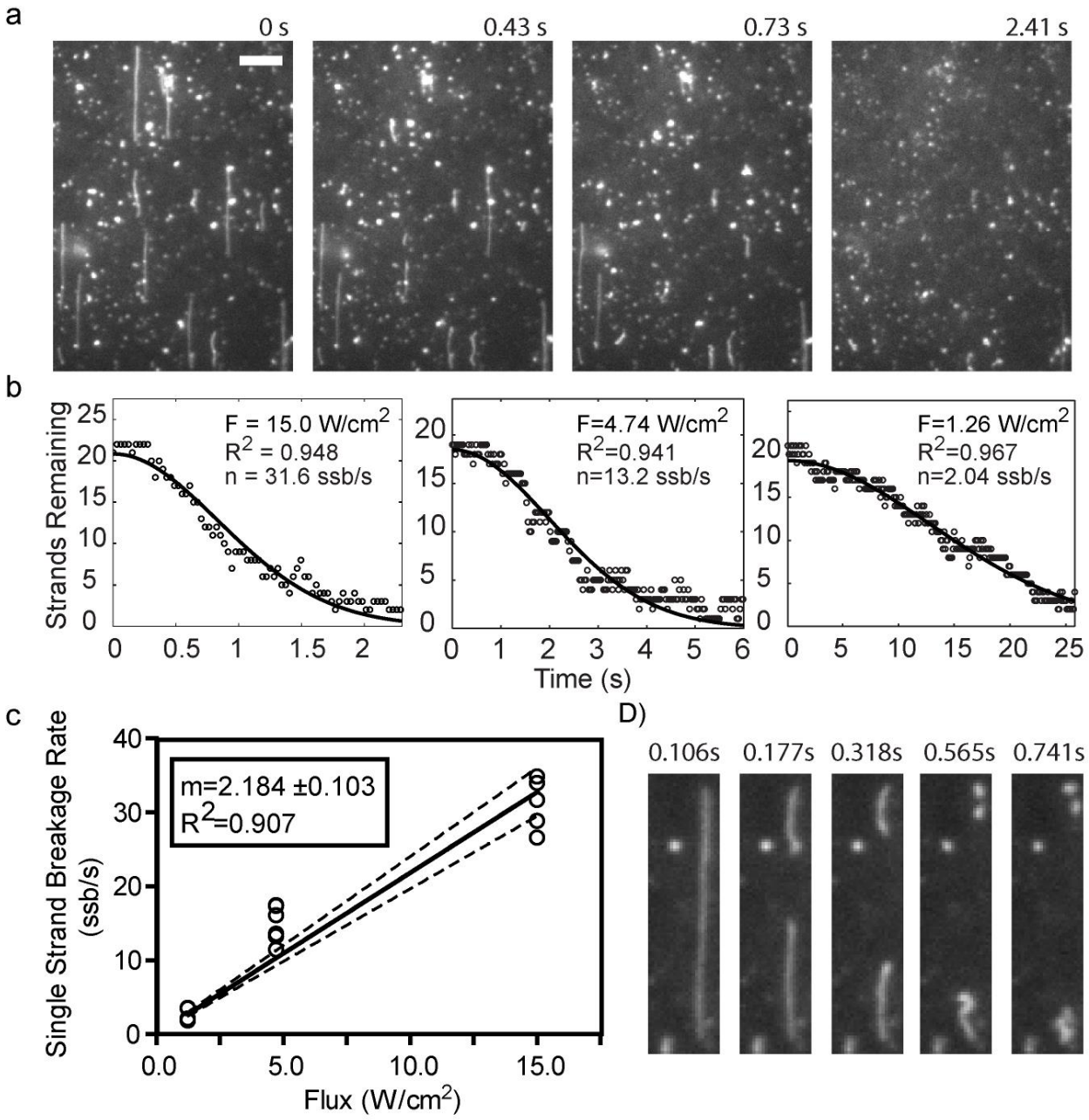

Figure 3.1- SMI strand cleavage assay and damage quantification for flow-stretched, YOYO-stained lambda DNA at a dye to nucleotide ratio of 1:4.(a) Movie stills depicting time resolved strand breakage. The sample was irradiated continuously with a 488 nm laser at an intensity of 15.0 W/cm^2. The DNA had been deposited as described in the text by flow in the vertical direction, but there was no flow while imaging. Bright spots at t=0 are mechanically cleaved

strands damaged during injection into the flow cell. Scale bar = 10 μm (upper right corner) **(b)** The quantification of the intact strands over time is displayed for three different irradiation intensities. The breakage curves have similar profiles and all display the early plateau region of SSB induction, but note that the time axes are different. The three different regions all had similar DNA strand density. The solid lines are fits to the stochastic DNA damage model, as outlined in the text. **(c)** The linear regression of the single strand breakage rates (n) as a function of laser flux provides a characteristic slope describing each breakage condition and can be used to estimate the breakage rate at any flux. **(d)** Time resolved breakage and recoiling of a single strand from panel A, confirming that strands are only tethered by their endpoints.

concentration. These trends are quantified below, after presenting results of related ensemble experiments and introducing the model used to extract cleavage rates by fitting the breakage

2. Ensemble study of single and double-strand photocleavage

Because the SMI experiments cannot detect SSBs directly, we also employed a bulk, electrophoresis-based assay to quantify the accumulation of SSBs preceding the formation of a double-strand cleavage. Supercoiled plasmids were used for the ensemble measurements because the structural forms that result from strand breaks can be separated by electrophoresis: the presence of one or more SSBs constitutes a plasmid form termed "nicked" that has a reduction in both the degree of supercoiling and the electrophoretic mobility, while the linear form that results from double-strand cleavage has an intermediate mobility. The buffers and staining ratios that we investigated using this ensemble assay were similar to those used in the SMI experiments to facilitate comparisons of the results. However, the irradiation conditions required to produce a sufficient amount of damaged DNA for detection in an ensemble assay were quite different than the SMI experiments. The ensemble assay requires several orders of magnitude more sample than the SMI assay, so a larger amount of solution must be irradiated with a defocused laser beam.

We therefore labeled supercoiled pBR322 plasmid DNA with YOYO-1 or YO-PRO1 at specific nucleotide:dye ratios and irradiated 60 µL drops of this solution with an unfocused laser beam. Aliquots of samples irradiated for various amounts of time were analyzed by electrophoresis (Fig. 3.2) and the resulting gel images were quantified using MATLAB (Fig. 3.2-

b). To confirm that the observed DNA damage resulted from a photosensitizing process and not from

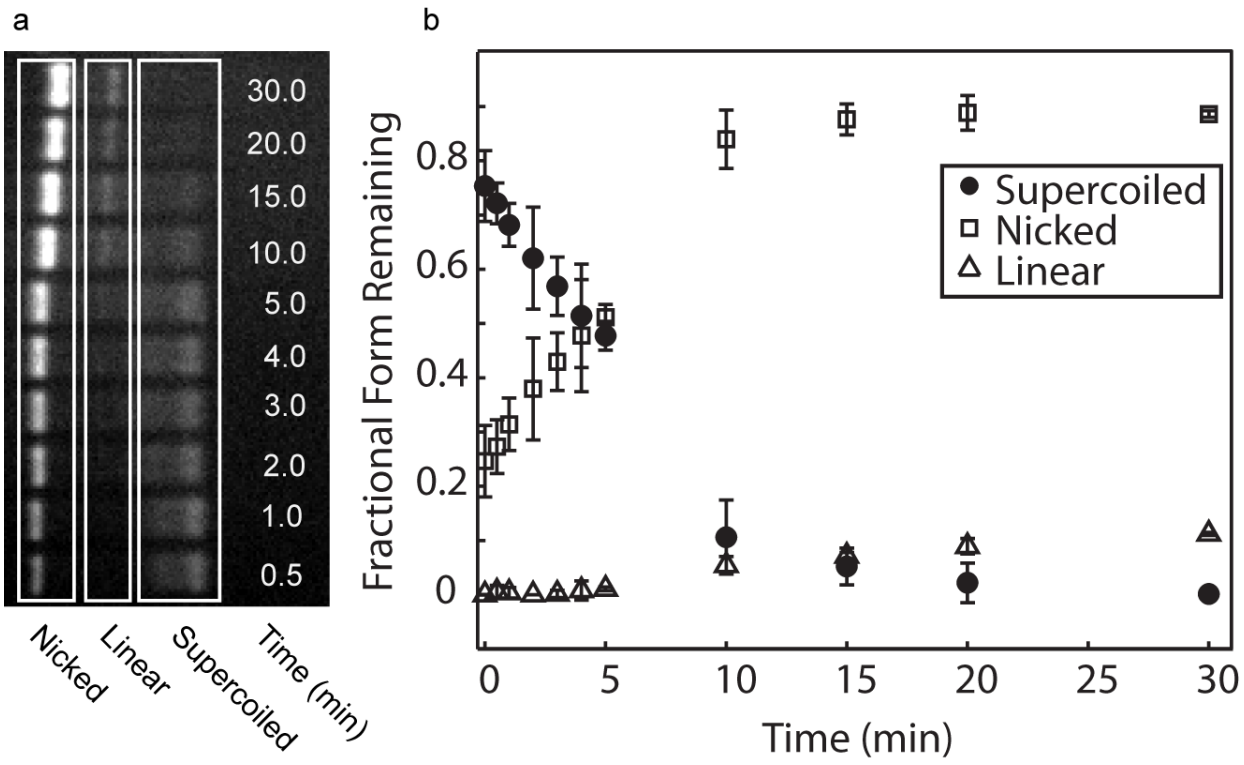

Figure 3.2- Ensemble breakage assay and damage quantification.(a) Gel image of YOYO-stained plasmid DNA samples (dye staining ratio of 1:10) that had been irradiated by an unfocused 488 nm laser for the times indicated. **(b)** Quantification of each DNA population over time, determined by fitting the intensity of each band in the gel to a Gaussian profile.

dye intercalation, we tested two control samples: one stained by the intercalating dye but not exposed to laser irradiation and one exposed to laser irradiation but not dye stained. These samples did not show any resulting damage compared to native plasmid, which has been observed to consistently be greater than 85% intact to start. As expected, short-duration irradiations resulted in a partitioning of the DNA between the intact supercoiled form and a growing damaged population comprised of plasmids containing one or more single strand breakage sites (collectively referred to as nicked). At longer time intervals, the supercoiled form was completely depleted as the nicked population dominated, to be eventually consumed as sufficiently numerous single strand breaks were accumulated to cause double-strand breaks, generating linearized plasmids (Fig. 3.2-a). The time-dependent partitioning of the various DNA forms for each sample-buffer condition was fit by the same model used for the SMI experiments, as discussed below.

3. Kinetic modeling of DNA strand cleavage

a. Modeling for DNA cleavage

To extract more meaningful information about DNA photocleavage and determine the relationship between single and double strand breaks from the SMI and gel data, we applied a complex model for DNA damage developed by Cowan et al. [19]. The model was originally intended to describe the action of a DNA nicking enzyme that creates SSBs in random locations on a supercoiled DNA plasmid, relaxing the supercoil. Double-strand cleavage linearizes the plasmid only when two nicks on opposite strands are sufficiently close (Fig 3.3-a). This process

is statistically equivalent to the mechanism of dye-mediated radical photocleavage, and it can be

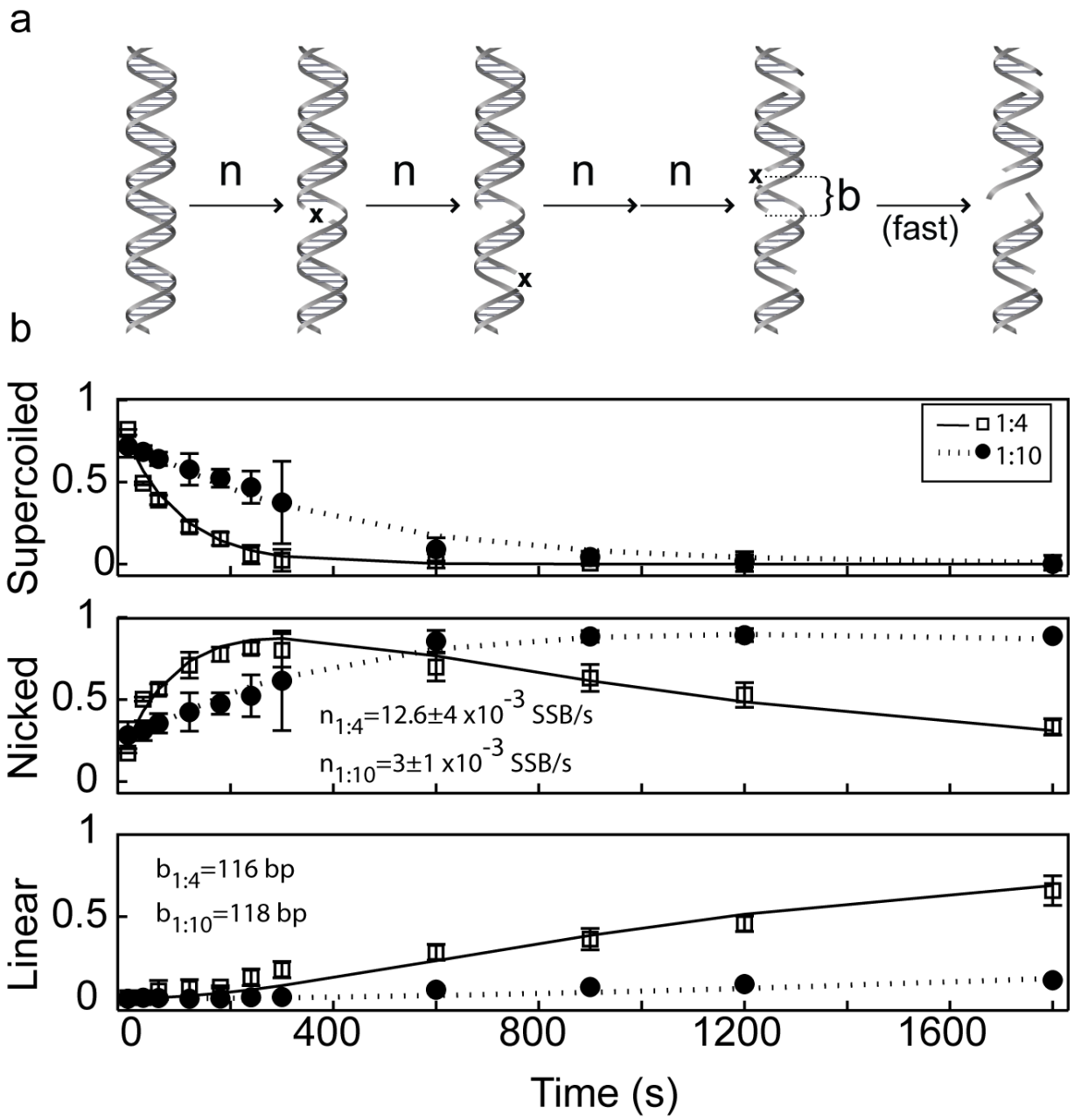

Figure 3.3- Stochastic DNA damage model and fitting of the ensemble data.(a) Illustration of the process in which a DNA strand suffers many single-strand breaks before cleaving. The rate of single strand breakage (n) is constant, and many breaks may occur (shown by x on the strand) before two are proximal enough to allow melting of the DNA strand between these breaks. This separation distance is defined as *b*. The *b*-value could only be determined for the

plasmid samples, because only the gel assay enabled quantification of the nicked population. **(b)** Fitting of the plasmid populations at the two dye to nucleotide ratios to the stochastic damage model. As expected, the higher staining ratio caused more rapid photodamage. The solid line corresponds to the model applied to the 1:4 dye ratio, while the dashed line corresponds to the 1:10 dye ratio. All three DNA populations were fit simultaneously to optimize the n and b values, but the single strand breakage rate (n, SSB/s) is determined primarily from the transition of the supercoiled to nicked populations, while the intra-strand breakage separation (*b*) distance is determined primarily from the rise of the linear population.

used to model single- and double-strand cleavage of either plasmid or linear DNA. After outlining the assumptions of the model and presenting the resulting equations derived byCowan *et al.* for the time-dependent population of each fraction, we use this model to fit the experimental data.

The DNA damage model is premised on the stochastic formation of single strand breaks: the number of single-strand nicks generated during each time period are statistically independent and follow a Poisson distribution. The key assumption of the model is that the damaging agent does not discriminate between sites along a DNA molecule and the formation of one SSB does not influence the formation of another. Furthermore, there is a characteristic distance between SSBs below which the attractive force exerted by the intervening hydrogen bonds is overcome by the entropic coiling force of the DNA polymer, leading to the formation of a double-stand cleavage. These assumptions are sufficient to derive the probability that a molecule can accumulate a certain number of SSBs without double-strand cleavage, leading to expressions for the time-dependent population of undamaged (U), nicked (N) and broken (B) fragments. These expressions are quoted below, but the reader is referred to [19] for a full derivation.

As the initial transition from a supercoiled to nicked plasmid obeys first order kinetics, the loss of the undamaged form is described by an exponential decay in time (t):

$$U(t) = e^{-nt} \tag{1}$$

Where *n* is the single-strand breakage rate. The nicked population conforms to a Poisson distribution of SSBs. The fraction that does not have SSBs close enough to form a DSB is given by:

$$N(t) = 2e^{-\frac{nt}{2}} - 2e^{-nt} + (nt) * X \quad (2)$$

$$X = \sum_{m=1}^{\infty} e^{-(\lambda t)(1+mb)/2}[(\lambda t)(1-mb)_{+}/2]^{2m-1}/2m! \quad (3)$$

Where*b* is the spacing between SSBs that leads to a double-strand cleavage, expressed as a fractional length of the DNA molecule. The subscript + following the difference (1-m*b*) indicates that if the quantity becomes less than unity, use the value zero; this truncates the summation at m=1/b. Finally, the broken population (due to double-strand cleavage) can be approximated:

$$B(t) \approx b^{-1}\left(e^{ntb/2} - 1\right)\left(ntX - Y + e^{-nt/2} - e^{-nt}\right) \quad (4)$$

where

$$Y = \sum_{m=1}^{\infty} e^{-\frac{(nt)(1+mb)}{2}}\left[\frac{(nt)(1-mb)_{+}}{2}\right]^{2m-1}\left[2m + \frac{(nt)(1-mb)}{2}\right]/(2m!) \quad (5)$$

The broken population includes strands that have undergone only one double-strand cleavage event. It is not possible to derive an analytic expression for this population, but Cowan et al.[19]derive lower and upper bounds, and recommend the use of Eq. 4 as a good approximation. Their full model accounts for further fragmentation of the broken population, but due to experimental limitations, these fragmented strands are not readily detectable in either SMI or gel studies. We account for this discrepancy by normalizing U, N and B by their sum at each time point.

b. Fitting ensemble data

Applying the model, we determined the SSB rates and characteristic break separation distances for each sample in the ensemble studies by varying the *n* and *b* parameters to fit the time-dependent fractions of supercoiled, nicked and linearized fragments (Fig. 3.3-b). We note that optimization of these two parameters is somewhat uncorrelated since the *n* value is primarily determined by the decay of the supercoiled species while the *b* value is determined primarily by the growth of the linear species; this does not apply for the SMI data because the undamaged and nicked populations are indistinguishable. Therefore, the value of the *b*-parameter determined by fitting the gel data is applied in fits of the SMI data under the same conditions.

Fig. 3.4-a compares the SSB rates measured for the two dye ratios and different solvent conditions. The damage rates measured in the ensemble assay are not proportional to the dye staining ratio. This may reflect the transition in dye binding modes between the two intercalater ratios, with the externally bound dye capable of mediating greater damage, possibly the result of an increased access to sensitize dissolved oxygen [16].

Table 1 presents the characteristic distances that produce double-strand cleavage from SSBs. To facilitate application to other DNA segments, they are listed as separation in base pairs by multiplying the fractional value of *b* obtained from fits by the length of the DNA plasmid. These base pair separation distances are assumed to apply to all genomic DNA molecules. The values obtained for YOYO-stained DNA are over 100 bp, which might seem somewhat surprising because single-strand fragments of this length should remain stably

bound for long times at room temperature. However, these values do not represent true separation distances because they do not explicitly account for the dye spacing.

We suspected that the use of a bis-intercalating dye biases the formation of closely-spaced strand breaks by linking two damage-causing fluorescent moieties close together along

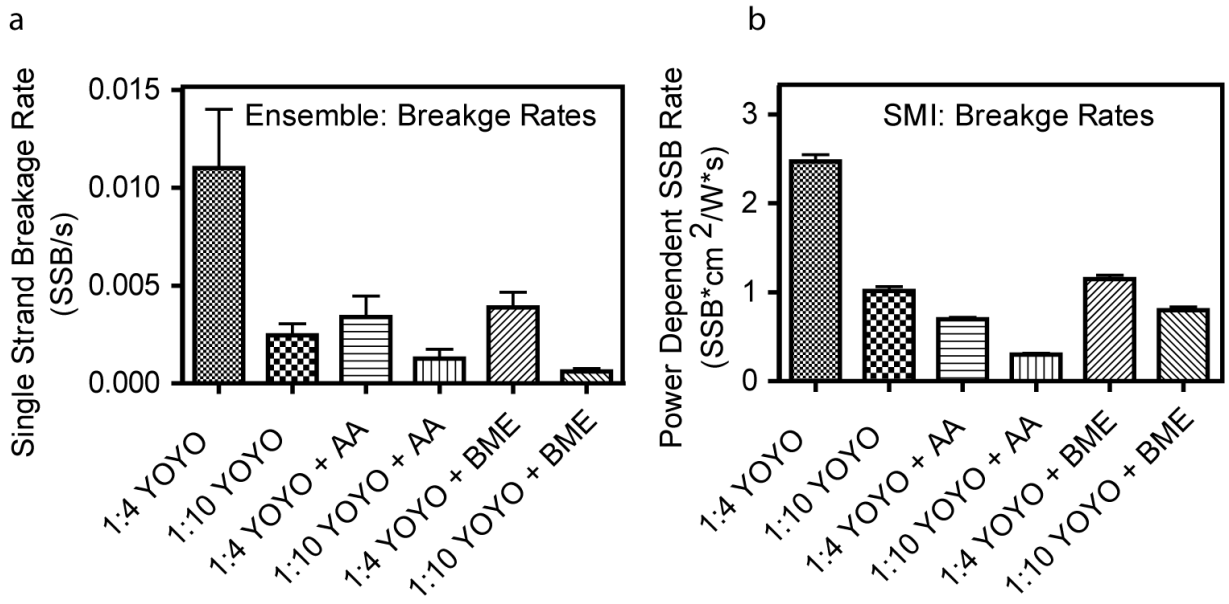

Figure 3.4- Comparison of the single-strand breakage rates (n) obtained by fitting results of the ensemble (A) or the SMI (B) damage assays to the stochastic DNA damage model.

the DNA chain. Therefore, we tested if the b-values determined were being inflated through the use of a bis-intercalater by determining the breakage rates and interstrand spacing using the dye monomer YO-PRO1. To account for the total number of potentially intercalated fluorophores, the concentration of YO-PRO1 dye was doubled to match that of YOYO-1 at each staining ratio. We reasoned that by decoupling the damaging agents, which are free to intercalate randomly along the chain, the observed *b*-values would decrease. This was confirmed, as the monomeric dye exhibits more physically realistic values of 15 and 27 bp for samples with higher and lower dye concentrations respectively. Unexpectedly, the monomeric dye induces a 3-fold greater single strand damage rate than the YOYO dimer at the same fluorophore concentration, potentially reflecting their relative photostability [20, 21].

c. Fitting double-strand photocleavage of flow-stretched DNA

Having determined the critical b-vales, the same model was then used to quantify the rate of single strand breakage in the SMI studies by fitting the breakage curves for each condition tested (fits overlaid with data in Fig. 3.1-b). Since the SMI experiments only reveal double-strand cleavage events, Eq. 4 and 5 were applied. As opposed to the gel assay, these experiments are only sensitive to double-strand cleavage, making it difficult to determine the *b*-value independently from the single-strand cleavage rate. Therefore, the values of *b* obtained in the ensemble studies were applied to their respective SMI conditions; only the single-stand cleavage rate, *n (SSB/s)*, was varied in the SMI data fitting procedure. The breakage rates were found to be very reproducible at all conditions tested (Fig. 3.4), including replicate measurements performed in different flow cells, and the data showed good agreement with the model (most datasets exhibiting a R^2-value > 0.90).

Table 3.1- Characteristic parameters describing the single strand breakage rates by imaging condition

Staining Ratio	b-value (bp)	Ensemble n (SSB/s)x10^{-3}	SMI Power Dependant n (SSB*cm^2/W*s)
1:4 YOYO	144±104	11±3	2.47±0.08
1:2 YO-PRO1*	15±13	32±13	-
1:4 YOYO + BME	114±75	4±1	1.15±0.04
1:4 YOYO + Ascorbic Acid	65±71	3±1	0.696±0.024
1:10 YOYO	118±111	3±1	1.02±0.05
1:5 YO-PRO1*	32±31	7±3	-
1:10 YOYO + BME	59±1800	0.6±0.2	0.798±0.038
1:10 YOYO + Ascorbic Acid	91±632	1.3±0.5	0.299±0.016

* The monomeric dye concentration was doubled relative to the bis-intercalater to maintain an equivalent number of fluorophores

We applied this SMI assay to characterize double-strand DNA breakage under a variety of conditions, including two dye-staining ratios and several buffer compositions. For each condition, we measured breakage curves as a function of several laser intensities. The breakage rate at each laser flux was then fit with a linear regression to extract a characteristic slope of the SSB rate versus laser intensity (Fig. 3.1-c). The regressions were performed over all the data points (not the averages at each irradiance) and forced through the origin (fits had an average R^2-value of 0.93). These slope quantities permit comparisons between different conditions and more importantly enable determination of the breakage rate at any flux. As opposed to the ensemble study, the rate of damage was observed to be proportional to the dye concentration, with a 2.5-fold increase in dye staining resulting in ~2.5 increase in the damage rate.

d. Effect of scavengers

The mitigating effects on DNA photodamage of the DNA protectants BME (5% v/v), and ascorbic acid (10 mM) were investigated by incorporation into the DNA buffers. We tested the primary scavenging systems in both the SMI and ensemble assays. From the ensemble damage assay, ascorbic acid and BME show damage rate reduction by 2.9-fold and 2.0-fold for the 1:4 and 1:10 dye to nucleotide ratio samples respectively (Fig. 3.4). Similar reductions were observed for the SMI studies, with ascorbic acid providing a slightly greater protective effect than BME.

e. Extrapolation between SMI conditions and ensemble studies

To validate using the results determined from the ensemble studies being applied to fit the SMI data, we extrapolated the damage rates measured for the SMI conditions to those measured in the gel studies by scaling for the incident laser flux (Fig. 3.5). The linear regressions were performed only on the SMI results. The regression was then extrapolated over two orders of magnitude to lower laser intensities, providing a comparison to the gel results. We found the estimated damage rates match very closely to the experimentally measured rate and most are within the 95% confidence intervals of the linear regressions. This indicates the relationship between laser flux and breakage rate holds over four orders of magnitude.

4. Degradation of DNA by ascorbic acid

Although beneficial in reducing photodamage, we also noted that addition of ascorbic acid to the sample buffer can have a deleterious effect on DNA substrates. We found that incubation of DNA plasmids in solutions containing ascorbic acid introduced enough SSBs to convert 60% of the plasmids to their nicked form in a 2-hr period at room temperature (Fig. 3.6). This was not observed for the TE or BME containing buffers. Thus, while able to impart a protective effect on DNA and reduce strand breaks in the presence of radical species, the scavenger itself is capable of inflicting significant damage on the DNA. This finding was only possible due to the use of the bulk assay which is sensitive to all three forms of the DNA plasmid, as the SMI work cannot resolve the single strand breaks.

We confirmed that this phenomenon is a dye independent process and is accelerated at elevated temperatures. Further, ascorbic acid samples from two leading suppliers- Fischer and Roche, were both confirmed to produce the same result. While the method by which this damage occurs is uncertain, two explanations are likely. The first is that ascorbic acid is

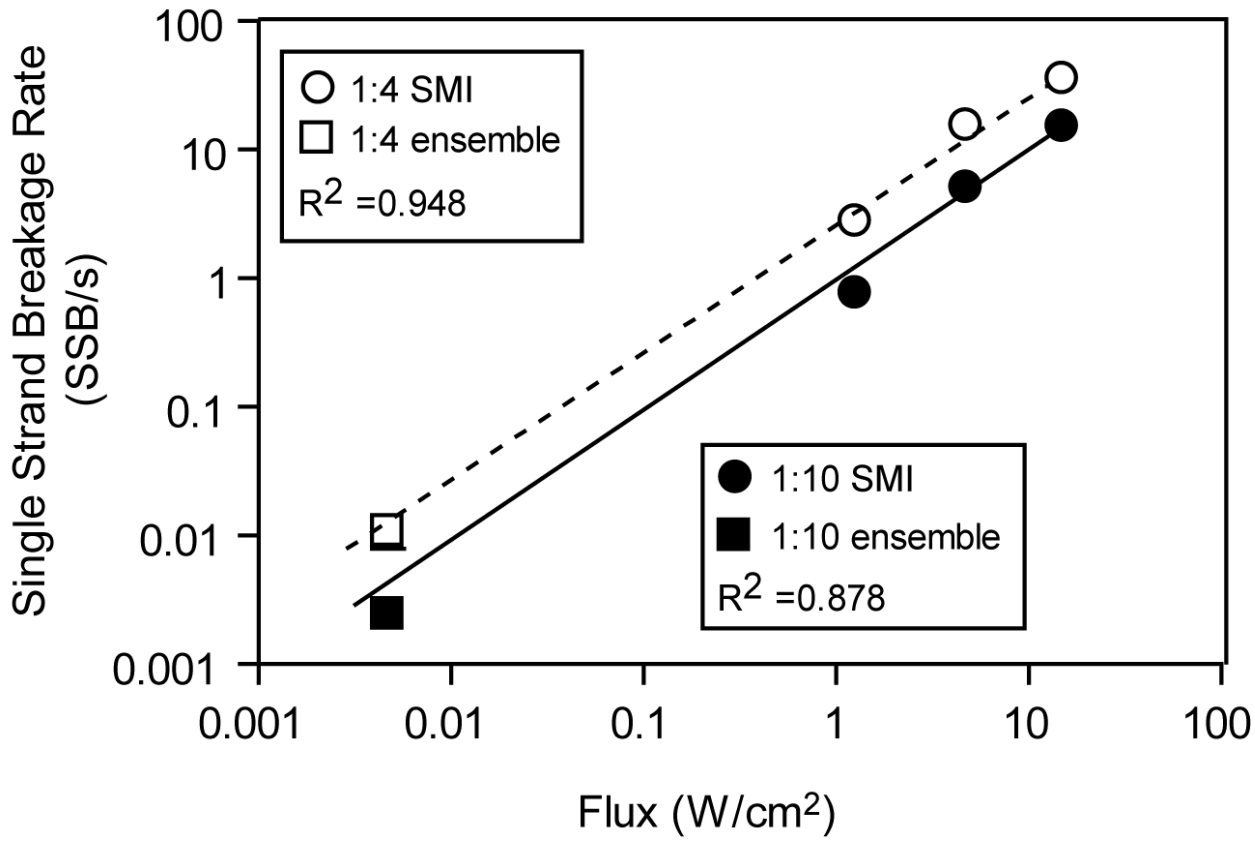

Figure 3.5- Extrapolation of the intensity-dependent SMI single-strand breakage rates to the laser intensity used for the ensemble measurements. The single strand breakage rates for the 1:4 (open circles) and 1:10 (closed circles) YOYO dye staining ratios are shown on a logarithmic scale with corresponding linear regressions. The linear regressions were performed only on the SMI data. Each data point represents an average value of 10-17 sample regions; error bars are too small to be discerned at this scale. The R^2-values shown apply to regressions performed over the entirety of the SMI data set (47 points for the 1:4 set and 46 points for the 1:10 set). The ensemble single strand breakage rates are plotted on the same scale and fall near the best-fit line, indicating that the linear regression fits to the SMI assay maintain linearity over 4-orders of magnitude in laser intensity. This confirms the applicability of extending values determined in the ensemble assay to the modeling of the SMI data.

participating in a Fenton reaction, reducing transition metal ions that have participated in generating hydroxyl radicals from a $3^+ \rightarrow 2^+$ state. This activity, previously reported *in vitro*, regenerates the 2+ transition metal ion, typically Fe or Cu, enabling them to catalytically generate hydroxyl radicals in the water and singlet oxygen [22, 23]. While the ion chelator EDTA was included in the buffer preparation, reports have indicated some metals, particularly iron, retain redox activity despite chelation [24]. An inductively-coupled plasma mass spectrometry analysis of the ascorbic acid and buffer systems used indicated the presence of Fe (II) at 2.5 ppb and Cu(II) at 0.1 ppb. The second explanation is that the natural oxidation of ascorbic acid by dissolved oxygen leads to the formation of damaging radical species.

Discussion

The purpose of our study is to quantify single-strand photocleavage during SMI experiments because this form of damage is not readily apparent. We do note that the SMI DNA breakage curves all share the common feature of an initial plateau before decaying (Fig. 3.1-b). This plateau corresponds to the induction period during which single strand breaks are accumulating but are insufficient in frequency to cause double-strand cleavage. The stochastic DNA damage model used to fit the data captures this feature (which corresponds to the lag time of the growth of the linear species in Fig. 3b). While an initial plateau also results from a more simplistic model based on two consecutive first-order reactions that assumes only two single-strand cleavage events are required to produce a double-strand break, the duration of the induction period with this model was not sufficient to fit the data. This simple model was also unable to fit the gel data.

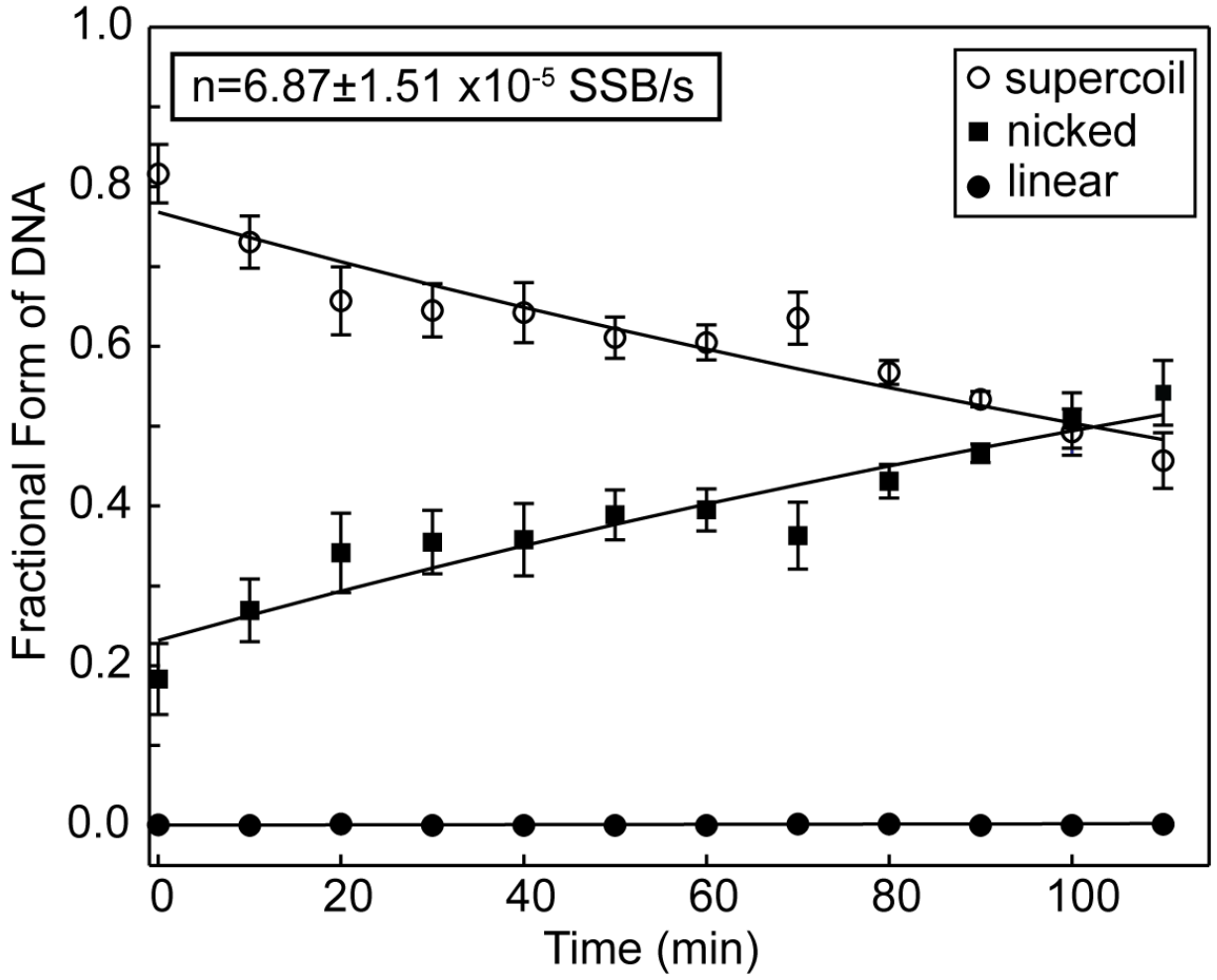

Figure 3.6- Ascorbic acid mediated DNA Damage. Quantification and fitting (solid lines) of the strand breakage caused by incubation of DNA plasmids in the absence of dye or irradiation with buffer containing ascorbic acid (10 mM) at 20°C. Over 20% of the plasmids accumulate single-strand breaks in a 2-hour time-span. Control experiments performed with TE and TE/BME but without ascorbic acid do not exhibit strand breaks on the same time scale.

From the observed cleavage of elongated molecules coupled with the basepair separation values determined using the ensemble damage assay, we determined the rate of single strand breakage at each laser power and dye ratio. Using the data in Table 1 along with the Equations 1-5, it is possible to estimate the number of SSBs in an intact DNA molecule observed using SMI. The measured breakage rates indicate that in all cases, a large number of undetectable single-strand damage sites are formed before the observable double-strand cleavage occurs. For example, we estimate that the intact strands in the Fig. 1a that have been imaged for only 0.73 secs have an average of 22 SSBs, and 75 SSBs by the 2.41 secs time point.

Our results highlight the utility of the stochastic DNA damage model to predict the accumulation of SSBs before double strand cleavage occurs, but application of this model to DNA stained with dimeric fluorophores is not perfect. The optimized values obtained for the *b*-parameter from fits to the YOYO-stained ensemble data were larger than 100 basepairs, which may sound surprising if the *b*-value is interpreted as the actual separation distance (bp) between SSBs that cause double-strand cleavage. A simplistic treatment of typical oligonucleotide melting temperatures indicates room temperature melting should occur for single strand breaks approximately 10 basepairs apart (applying the Wallace Rule, $T_m \approx 3°C*(bp)$ [25]). This discrepancy is rationalized however, by considering how the incorporation of bis-intercalating damage agents violate the assumption of randomly positioned damage, as compared to a mono-intercalating damage agent. In the case of the mono-intercalater, damage is caused at intervals approximately equal to the spacing of the intercalater. Thus many damage events are required to cause two in close proximity. However, by tethering the damage agents, single strand breaks are accumulated in much closer proximity than would be

expected if the damage agents were randomly spaced along the DNA backbone. This causes double strand breaks to occur at a greater frequency than expected for the same number of SSBs. Since the same number of single strand breaks have occurred but resulted in more double strand cleavages, it seems as if the interstrand break spacing required to produce a double-strand break is quite large. In actuality, a small basepair separation is still required to melt the double helix, but application of the stochastic model to dimeric dyes manifests itself as a inflated *b*-value.

Our experiments used dye to nucleotide ratios of 1:10 and 1:4. Previous studies have indicated that the former (lower) dye ratio, fully intercalates into the double helix, while the latter (high) dye ratio saturates the intercalation sites and additionally binds though nonspecific electrostatic interactions along the DNA backbone [20]. These staining ratios are well above the threshold required to visualize the full contour of extended lambda [26], but they mimic the staining action of incorporating dye directly into the working buffer used during an experiment [17]. Of greater importance, the use of high staining ratios was necessary to ensure we were able to observe double strand breaks. Consider a very low staining ratio, in excess of 100 nucleotides: 1 dye molecule. In this situation it becomes likely that the damage agents will be spaced further apart than the distance required for two single strand breaks to cause a double strand cleavage. This would result in the extensive formation of single strand breaks along the DNA molecules without ever being detectable as a rupture of the molecule. This possibility is highlighted by the findings for the YO-PRO1 mono-intercalater. This dye was found to have a greater rate of single strand breakage than its dimeric counterpart. Yet the small interaction distance between single strand breaks means it becomes easy to stain at a ratio that spaces

these dyes further apart than the minimum distance to cause a double strand cleavage. In effect this dye biases damage towards single strand breaks and enables a distinction between the single strand breakage rate and the double strand breakage rate. In many ways this represents the worst case scenario for an experiment in which single strand breaks can disturb protein interactions. Therefore, dimeric dyes at moderate concentrations are optimal for SMI experiments on protein-DNA interactions that could be perturbed by the presence of SSBs.

Conclusion

This work was undertaken to quantify the damage mediated by common fluorescent DNA intercalaters on DNA substrates during imaging experiments. While we tested YOYO-1, these findings are applicable to any DNA intercalater, including those used to image nucleic acids in live cells. We determined the breakage rates of DNA using a gel based assay, gaining information about the separation of strand breaks required to linearized the molecule and ability of radical scavengers to reduce damage rates. These findings were applied to the study of flow stretched lambda DNA, and this is the first work to report breakage rates both with and without radical scavenging systems. These breakage rates can be used to estimate the prevalence of single strand nicks occurring on a DNA molecule over the course of a typical optical imaging experiment, which are undetectable by optical methods. Such information is vitally important in the consideration of data obtained concerning DNA-repair protein interactions as many proteins recognize strand breaks as binding sites [10, 27]. In such circumstances, the unintentional formation of protein trap sites could strongly bias the movement and interaction times of DNA-protein complexes. Further, we observed that

ascorbic acid mediates DNA degradation in the absence of dye, which should be considered before using it as a radical scavenger.

REFERENCES

(1) van Mameren, J.; Peterman, E. J. G.; Wuite, G. J. L. *Nucleic Acids Res.* **2008***, 36*, 4381-4389.

(2) Finkelstein, I. J.; Visnapuu, M. L.; Greene, E. C. *Nature, 2010, 468, 983-987* .

(3) Bianco, P. R.; Brewer, L. R.; Corzett, M.; Balhorn, R.; Yeh, Y.; Kowalczykowski S.C.; Baskin, R.,J. *Nature 409, 374-378 (2001)* .

(4) Schweitzer, C.; Schmidt, R. *Chem. Rev.* **2003***, 103*, 1685-1757.

(5) Teoule, R. *Int. J. Radiat. Biol.* **1987***, 51*, 573-589.

(6) Ward, J. F. *Int. J. Radiat. Biol.* **1990***, 57*, 1141-1150.

(7) Guo, H.; Tullius, T. D. *Proc. Natl. Acad. Sci. U. S. A.* **2003***, 100*, 3743-3747.

(8) Siddiqi, M. A.; Bothe, E. *Radiat. Res , 112*, 449-463.

(9) Patrick, M. H.; Rahn, R. O. *Photochemistry and Photobiology of Nucleic Acids;* Academic Press: New York, 1976; Vol. II, pp 35-96.

(10) Houten, B. V.; Croteau, D. L.; Vecchia, M. J. D.; Wang, H.; Kisker, C. *Mutat. Res.* **2005***, 577*, 92-117.

(11) Friedberg, E. C. *Nature* **2003***, 421*, 436-440.

(12) Caldecott, K. W. *Nature Reviews: Genetics* **2008***, 9*, 619-631.

(13) Gurrieri, S.; Wells, K. S.; Johnson, I. D.; Bustamante, C. *Anal. Biochem. 249, 44-53 (1997)* .

(14) Tanner, N. A.; van Oijen, A. M. *Methods Mol Biol. 2009;521:397-410.* .

(15) Smith, S. B.; Cui, Y.; Bustamante, C. *Science (1996): Vol. 271 no. 5250 pp. 795-799* .

(16) Akerman, B., Tuite, E. *Nucleic Acids Res.* **1996***, 24*, 1080.

(17) Kad, N.M., Wang, H., Kennedy, G.G., Warshaw, D.M., Van Houten,B. *Mol. Cell* **2010***, 37*, 702-713.

(18) Shubsda, M. F.; Goodisman, J.; Dabrowiak, J. C. *J. Biochem. Biophys. Methods 34 (1997) 73-79* .

(19) Cowan, R.; Christina, C. M.; Grigg, G. W. *J. Theor. Biol. (1987) 127, 229-245* .

(20) Larsson, A.; Carlsson, C.; Jonsson, M.; Albinsson, B. *J. Am. Chem. Soc. 1994, 116, 8459-8465* .

(21) Benson, S. C.; Mathies, R. A.; Glazer, A. N. *Nucleic Acis Research, 1993, vol.21, no.24, 5720-5726* .

(22) Saran, M.; Bors, W. *Radiat Environ Biophys* **1990***, 29*, 249-262.

(23) Filho, A. C. M.; meneghini, R. *Biochimica et Biophysica Acta, 781 (1984) 56-63* .

(24) Graf, E.; Mahoney, J. R.; Bryantm R., E., J.W. *J. Biol. Chem.* **1984***, 259*, 3620.

(25) Wallace, R. B.; Shaffer, J.; Murphy, R. F.; Bonner, J.; Hirose, T.; Itakura, K. *Nucleic Acids Res. 6, 3543 (1979).* .

(26) Graneli, A.; Yeykal, C. C.; Prasad, T. K.; Greene, E. C. *Langmuir* **2006***, 22*, 292-299.

(27) Friedberg, E. C.; Walker, G. C.; Siede, W. *DNA repair and mutagenesis;* ASM Press: Washington D.C., 1995; .

CHAPTER 4

RNA POLYMERASE II SUBUNITS EXHIBIT A BROAD DISTRIBUTION OF MACROMOLECULAR ASSEMBLY STATES IN THE INTERCHROMATIN SPACE OF CELL NUCLEI

"Science is like sex: sometimes something useful comes out, but that is not the reason we are doing it"

-Richard P. Feynman

Overview:

Nearly all cellular processes are enacted by multi-subunit protein complexes, yet the assembly mechanism of most complexes is not well understood. The anthropomorphism "protein recruitment" that is used to describe the concerted binding of proteins to accomplish a specific function conceals significant uncertainty about the underlying physical phenomena and chemical interactions governing the formation of macromolecular complexes. We address this deficiency by investigating the diffusion dynamics of two RNA Polymerase II subunits, Rpb3 and Rpb9, in regions of live cell nuclei that are devoid of chromatin binding sites. We demonstrate that both unengaged subunits are incorporated into a broad distribution of complexes, with sizes ranging from free (unincorporated) proteins to those that have been predicted for fully assembled gene transcription units. In live cells, Rpb3 exhibits regions of stability at both size extremes connected by a continuous distribution of complexes. Corresponding measurements on cellular extracts reveal a distribution that retains peaks at the extremes but not in between,

suggesting that partially assembled complexes are less stable. We propose that the broad distribution of macromolecular species allows for mechanistic flexibility in the assembly of transcription complexes.

Introduction

A central question in modern molecular biology is the mechanism by which large, multi-subunit protein complexes assemble inside a cell. Essential cellular processes such as transcription [1], splicing [1, 2], and genome repair [3] are undertaken by massive assemblies involving many distinct molecular modules that efficiently carry out specific tasks. While "protein recruitment" is cavalierly viewed as the initial step in assembly, molecular-level details about how this process is initiated and through what intermediates such complexes form remain ambiguous [4]. Two primary models have emerged to explain how cellular machinery assembles to handle the dynamic demands they must meet [5]. One proposal is a top-down approach, in which the components of a macromolecular assembly bind one another prior to receiving an activation signal, forming a stable supra-assembly that is often called a molecular factory. Such a factory would be poised for efficient handling of cellular tasks but would be slow to traverse the cellular interior and poorly suited to respond to changing external stimuli. On the other extreme is a bottom-up approach, in which each component of the final molecular assembly diffuses through the cellular interior individually and stochastically encounters binding partners at the active site until the entire complex is amassed. This stochastic model would enable rapid movement of the smaller molecular modules within the cell, but the binding steps to form a full complex from individual components may limit the overall activation rate. Interestingly, proponents of both models invoke the crowded nuclear milieu as corroborating evidence,

either in support of factory domains or restrictive nuclear architecture [6, 7]. In an effort to distinguish between these paradigms, we decided to investigate the incorporation of individual components of the RNA Polymerase II (RNAPII) transcription complex in regions of live cell nuclei devoid of chromatin binding sites.

The present study specifically investigates RNAPII since it is responsible for mRNA production and occupies a critical position in the central dogma. While extensive *in vitro* molecular biology research has elucidated the mechanical intricacies of how the RNAPII complex transcribes template DNA, the advent of *in vivo* fluorescent labeling and the widespread use of fluorescent microscopy have enabled detailed observations of RNAPII complex interactions with chromatin in the native cellular environment [8-11]. Much work has been conducted to characterize RNAPII behavior in bacterial, insect, and mammalian systems; however, the majority focuses specifically on subunit assembly and interactions on chromatin, typically in the vicinity of DNA binding sequences. In studies using both RNAPI and RNAPII, polymerase subunits and transcription factors have been found to have distinct dynamics, arguing against preassembled complexes [8, 9, 12], though these results contradict some earlier work [13,14, 15]. Thus, it remains unresolved whether the assembly is stochastic [9] or stepwise [8, 16], with implications for a generalized framework of multi-component protein assemblies [17].

No previous investigations have characterized RNAPII component diffusion dynamics preceding chromatin interactions in cells and most studies have completely neglected the importance of diffusion. We postulated that measuring the diffusion dynamics of RNAPII components prior to chromatin binding could yield insights into the mode of assembly. We

sought to better understand the process of RNAPII complex assembly and nuclear mobility by investigating the dynamics of the Rpb3 and Rpb9 subunits in the interchromatin space (nucleoplasm devoid of chromatin) of cell nuclei using fluorescence recovery after photobleaching (FRAP).

We express fusions of Rpb3 and Rpb9, two subunits exclusive to RNAPII , with enhanced green fluorescent protein (GFP) in the polytene cells of *Drosophila melanogaster* larvae [18]. These polytene cells contain many copies of the genomic DNA that form large chromosomal bundles during interphase (Fig 4.1-a,b). By expressing RNAPII subunit-GFP fusions and H2B-mRFP tagged histones in polytene cells, we are able to optically resolve nuclear regions containing chromatin and restrict our analysis exclusively to the interchromatin space (Fig 4.1). This region is devoid of chromatin and therefore lacks DNA binding sites. We find the diffusion of both RNAPII subunits was non-Brownian and the recovery dynamics of the two subunits are different.

While non-Brownian diffusive behavior is often termed anomalous and attributed to molecular crowding [19], we propose a fundamentally different interpretation. Through a comparison to the mobility of unconjugated GFP (lacking a localization sequence) [20], which does exhibit Brownian diffusion, we determine that molecular crowding is not responsible for the observed diffusive behavior. Rather, both RNAPII subunits must participate in heterogeneous distributions of complexes with a broad range of sizes, from isolated subunits to fully assembled transcription complexes. We term this type of diffusive behavior *apparent*

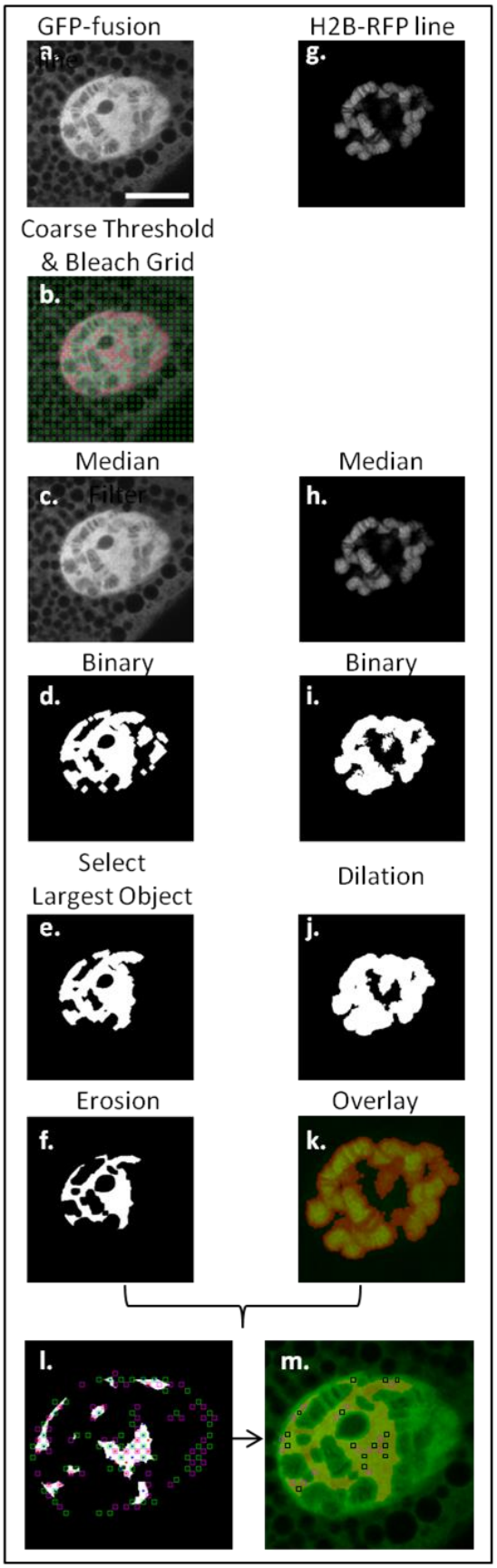
GFP-fusion
a.
H2B-RFP line
g.
Coarse Threshold
& Bleach Grid
b.
Median
Filter
c.
Median
h.
Binary
d.
Binary
i.
Select
Largest Object
e.
Dilation
j.
Erosion
f.
Overlay
k.
l.
m.

Figure 4.1- Image Collection and Automated Processing Methodology "Shotgun ptFRAP". The primary limitation of the ptFRAP method is the low SNR, requiring averaging over hundreds of individual bleach and control points. Collecting sufficient data necessitated an automated collection method in which an image of the sample is collected followed by the collection of ptFRAP curves at evenly spaced grid points in the sample. Only the subset of ptFRAP curves collected at grid points that meet the image selection criteria are used for subsequent analysis. (**a,g**) An initial image of both color channels is captured and used in subsequent thresholding operations. The GFP channel corresponds to the protein of interest, the RFP channel to the labeled polytene chromosomes. (**b**) A grid with 20 μm spacing is applied to the entire field of view. These grid points define the positions where FRAP data is collected. This is several times larger than the 300 nm PSF of the laser beam. A coarse threshold is applied to the GFP channel; only grid points contained within the thresholded region are collected (magenta boxes). Alternating points of the grid correspond to bleach and control datapoints. Post-processing steps are performed using MATLAB scripts developed in-house. (**c,h**) After data collection, a median filter is applied to both images to remove noise. (**d,i**) Threshold values are carefully selected for each image to capture the contours of the nuclear features. (**e,f**) In the GFP channel, the largest object in the field of view, corresponding to the nucleus, is retained. This eliminates any contributions from cytoplasmic signal. The binary mask is processed to remove sharp edge features then eroded 500 nm from every periphery to eliminate grid points in the vicinity of cellular membranes. (**j**) The polytene binary mask is dilated 300 nm to remove any grid points nearby the chromatin. (**k**) The mask (red can be seen overlaid with the image) confirms the entire region containing the polytenes will be excluded from analysis. (**l,m**) The

RFP channel mask is subtracted from the GFP channel mask; the resulting region corresponds to the interchromatin space. The open squares (green=control power, magenta=bleach power) indicate all grid points at which FRAP data is collected during the experiment, while squares enclosing dots indicate the grid points retained for analysis. The distribution of the retained grid points are inspected visually to verify the selection criteria have been met.

anomalous diffusion, in which non-Brownian behavior is observed by simultaneously probing many states of pre-formed complexes with different diffusion coefficients.

Materials and Methods

All chemicals are Fisher brand unless noted.

1. Fly Strains

Drosophila lines that express Rpb3-GFP, Rpb9-GFP or H2B-mRFP using the GAL4/UAS system have been described previously [22,21]. Fly lines containing transgenes for unconjugated GFP and Gal4-C147 were obtained from the Bloomington Drosophila Stock Center (lines #5430 and #6979 respectively). All GFP samples are enhanced green fluorescent protein. To simultaneously express H2B-mRFP with GFP or GFP fusions for dual color imaging, the homozygous line Gal4-C140; H2B-mRFP was first generated and then crossed to the appropriate GFP fusion transgenic line. Flies were raised using a standard cornmeal medium at room temperature; larvae were collected after 8-9 days. To prepare samples for imaging, wandering third-instar larvae were dissected in Grace's Insect Medium and intact salivary glands were used for imaging polytene cells. All imaging experiments were completed within one hour of dissection to maintain cell viability.

2. Salivary Gland Extract Preparation

To prepare polytene cellular extract samples of GFP and EGFP-Rpb3, 80 larvae were dissected and the glands placed on ice cold Tris-buffer (50 mM, pH 7.4). The glands were mini-centrifuged for 60 secs, the supernatant removed, and the glands re-suspended in ice cold lysis buffer (50 µL), followed by vortexing for 45 s and sonication for 30 mins to rupture the glands.

The lysis buffer consisted of Tris-HCl (50 mM, pH 7.4), NaCl (150 mM), NP-40 detergent (0.5% w/v), Pefabloc SC (1 mM in Tris buffer), leupeptin (2 µg/mL, in methanol), and pepstatin (2µg/mL, in methanol). After sonication in ice cold lysisbuffer, the sample was mini-centrifuged for 4 mins. The supernatant was used immediately for FRAP experiments.

3. Two-photon microscopy configuration and FRAP Procedures

Imaging and FRAP were done as described in our previous paper [20]. In brief, polytene cells were imaged with a 1.2NA/60x Olympus objective using a home-built laser scanning two-photon microscope. GFP and RFP were excited at 950 nm by a Chameleon Ultra II Ti:sapphire pulsed laser with a 140 fs pulse duration; the fluorophore emissions separated with a 570 short pass dichroic mirror. The GFP emission was collected with a 510/30 bandpass filter while RFP emission was collected with a 630/100 dichroic mirror. Quantitative bleaching studies were performed with a point-bleaching method (ptFRAP) developed previously in our laboratory, featuring an online image thresholding and data acquisition procedure followed by offline image analysis and data modeling. For all conditions studied, between 20-40 cells were analyzed; the number of datapoints collected and averaged are indicated in Figures 2 and 4. Data collection consists of two phases- recording bleach and control datapoints. Bleach points are established by photobleaching a diffraction limited volume (spot size of 300 nm diameter and 1 µm axial length) at a high laser power (bleach power) followed by recording the intensity of the spot during the diffusive recovery at a lower laser power (read power). Control points are established in the same manner but with the read power used in place of the bleaching power. A bleach depth of between 40-60% of the initial fluorescent intensity was achieved using a bleach power of 71.5 mW, while control measurements were taken at a read power of

11.5 mW (both values measured at the microscope objective using a calibrated power meter). For all proteins studied, FRAP recovery data was collected for 50 ms and data fitting was applied to datapoints collected starting at 80 µs post-bleach. The data was fit with a model for anomalous subdiffusion [20], which indicates the degree of anomlity and the diffusion coefficient (for normally diffusing species) or the transport coefficient of the diffusing species. The anomlity factor ranges between 1 and 0, with unity indicating Brownian diffusion. For detailed information on the microscope configuration, FRAP timing sequence, and fitting recovery data to an anomalous subdiffusion model with a photophysics correction for observational photobleaching, (Daddysman and Fecko[20]).

Results

1. Automated "shotgun ptFRAP" data collection

We chose to study the transport properties of the RNAPII subunits Rpb3 and Rpb9 in the absence of chromatin binding sites or membrane perturbations by restricting the region of FRAP investigation to the interchromatin space of cell nuclei. We used a point-FRAP (ptFRAP) method to probe diffusion, which is an implementation where optical diffraction-limited spots are photobleached and the fluorescent recovery tracked in time with sub-millisecond resolution [20]. In contrast to the more common area-FRAP in which micron-sized features are photobleached[22], ptFRAP probes smaller sample regions and enables several orders of magnitude higher time resolution. To restrict the analysis of photobleaching recovery to the interchromatin space of polytene nuclei (avoiding both cellular membranes and chromatin regions) and prevent datapoints from overlapping in space during collection, we implemented an automated datapoint collection method termed "shotgun ptFRAP" (Fig. 4.1). The method

consists of a data collection program in which evenly spaced datapoints are collected across the entire cell nuclei (ie. the entire cell is "hit" Fig. 4.1-b), followed by a post-experiment screening step that retains only datapoints in regions of interest that match our selection criteria (Fig. 4.1-l,m). Thus all regions of the nuclei are probed over the course of the experiment and individual regions can be analyzed afterwards. This procedure enables over a thousand datapoints to be collected, without user bias, and are averaged into a single FRAP dataset.

2. Different recovery dynamics observed for RNAPII subunits

Rpb3 is the third largest RNAPII subunit, having a native mass of 35 kDa; the GFP- fusion construct has a mass of 62 kDa. Native Rpb9 is less massive at 14 kDa; the fusion construct has a mass of 41 kDa. Both tagged subunits are incorporated into active transcription complexes [23]and the subunits have high binding affinities for most of the ten remaining RNAPII subunits [24]. Additionally, RNAPII has strong affinities for transcription factors and promoter proteins, giving rise to a large distribution of complexes in which Rpb3 and Rpb9 may participate. Using the ptFRAP method, we compared the recovery dynamics of both subunits in the interchromatin space of polytene nuclei, which were then compared to the recovery of unconjugated GFP under the same conditions. The GFP acts as an inert protein with no binding partners in the nucleus and is only subject to molecular crowding (Fig. 4.2). We have previously shown that unconjugated GFP obeys Brownian diffusion in the interchromatin space [20] exhibiting a reduced diffusion coefficient due to nuclear viscosity. For this study, GFP serves as an approximate molecular mass standard to account for the effects of nuclear crowding as a reduction in the translational diffusion coefficient [25]. However, it is apparent that differences in the FRAP curves between the RNAPII subunits and GFP (Fig. 4.2-c) indicate that the transport of these former species is not well described by Brownian

:

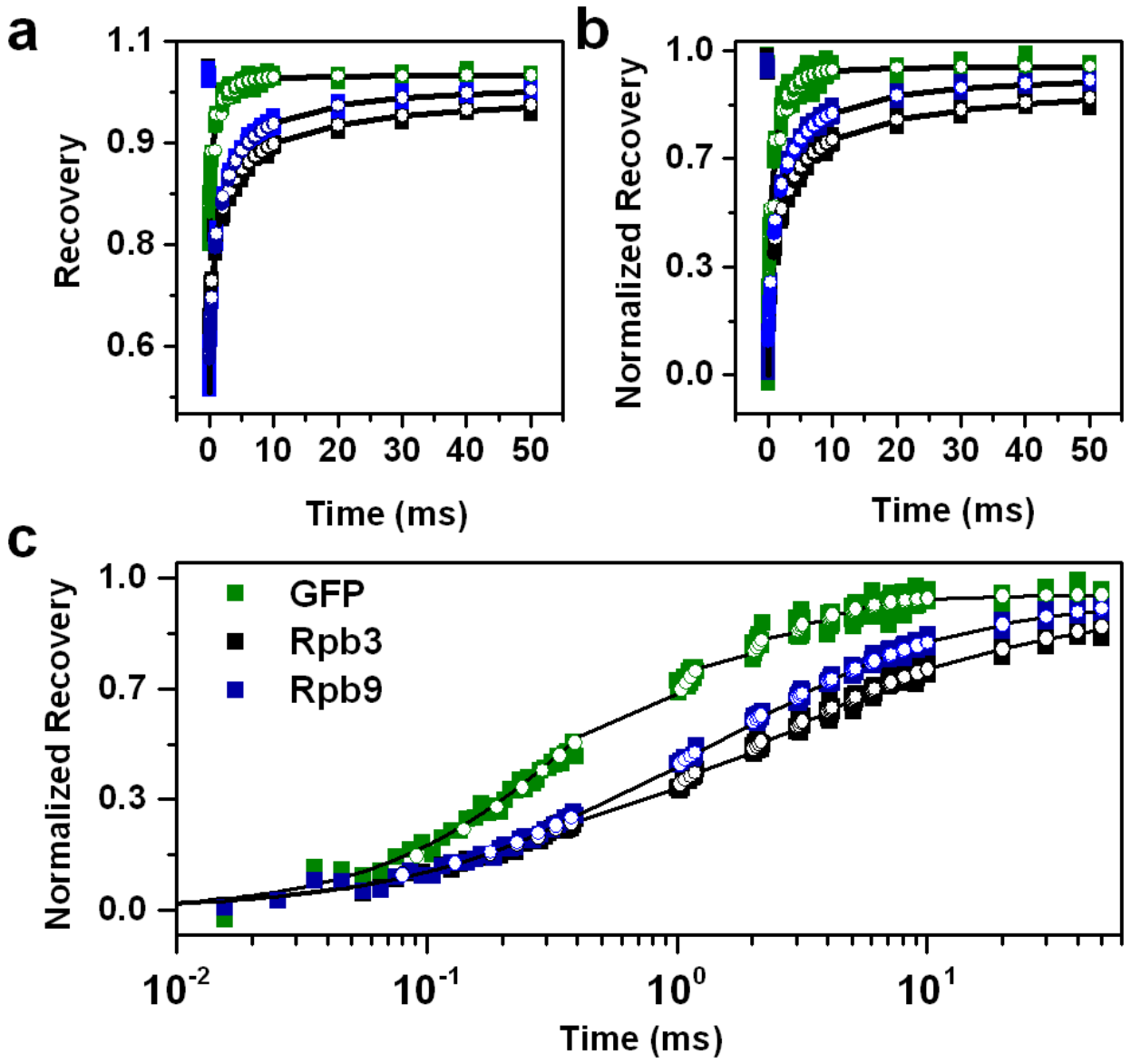

Figure 4.2- Comparison of in vivo subunit recovery dynamics.(a) The FRAP curves for the unconjugated GFP (green), the Rpb3-GFP (black), and Rpb9-GFP (blue) are shown. Data are plotted as closed squares, the best-fits to an anomalous diffusion model are shown as black lines, best-fits to the distribution model are shown are white circles. The data was collected with an intermittent collection technique that minimizes photobleaching while enabling long-duration interrogation. Numerous FRAP curves were averaged for each sample (GFP-1505 pts,

Rpb3-1694 pts, Rpb9-833 pts) to achieve a high SNR. All displayed data has been treated to a 10-point rolling average smooth to aid clarity but all fitting was performed on the un-treated datasets starting at the 80 μs time-point. (**b**) Evident from the immediate post-bleach datapoint, each protein exhibits a different bleach depth. This reflects a sample-specific protein expression level effect that significantly influences the bleach depth. To enable qualitative comparison of the FRAP recovery curves, we normalized the FRAP bleach depth for each sample to zero. The rescaled FRAP curves clearly indicate differences between the recovery profiles of GFP, Rpb3, and Rpb9. The recovery differences are striking given the similar molecular masses and identical nuclear environment. (**c**) For better comparison of short-time data, the rescaled recovery curves are displayed on a logarithmic time axis. Here, the differences in the slopes of recovery curves can be visualized: the flatter the slope, the greater the apparent anomlity factor.

diffusion. This result is striking given the similar masses of the three proteins and the weak dependence of diffusional mobility with molecular mass predicted by the Stokes-Einstein Equation.

Given the large differences between the recovery of GFP and the RNAPII subunits, we chose to initially fit the Rpb3 and Rpb9 FRAP curves with a model that allows for anomalous subdiffusion. Anomalous subdiffusion equations are often invoked to describe mass transport in which the mean squared displacement of each particle is sublinear with time, which can result from heterogeneity in the molecular environment

$$\langle \Delta r^2 \rangle = 6\frac{\Gamma}{\alpha}t^{\alpha} \quad (1)$$

The particle displacement is Δr, Γ is the transport coefficient, t is the time interval, and α is the anomlity value. The principle parameter describing anomalous diffusion is the anomlity value, bound between zero and unity, which indicates the magnitude of the deviation from Brownian behavior. An anomlity factor of unity corresponds to Brownian behavior (for which the transport coefficient is the diffusion coefficient); smaller values indicate progressively larger deviations. Such hindered molecular motion is often attributed to intracellular factors that retard the motion of a particle, such as binding to immobile traps, participation in viscoelastic complexes, and physical obstruction through labyrinthine corralling [26].

The ptFRAP model previously developed by our group [20] accounts for both anomalous diffusion[27] and a reversible photobleaching correction due to dark-state transitions of GFP during data collection. The FRAP signal is:

$$F(t) = F_0[1+\delta\exp(\frac{-t_{laser}}{\tau_{PP}})]\sum_{n=0}^{\infty}\frac{(-\beta)^n}{n!}\left[1+n\left(1+\frac{16\Gamma t^{\alpha}}{\alpha\omega_r^2}\right)\right]^{-1}\left[1+n\left(1+\frac{16\Gamma t^{\alpha}}{\alpha\omega_z^2}\right)\right]^{-1/2} \quad (2)$$

Here, F_0 is the pre-bleach fluorescence intensity, β is a factor related to the bleach depth, δ and t_{laser} are the reversible bleaching magnitude and timescale, and ω_r and ω_z are the size of the focused Gaussian beam in the radial and axial dimensions respectively. All of our data exhibited a near complete recovery on the 50 ms timescale indicating no immobile fractions. We fit the averaged FRAP curves according to Eq.2 (Fig. 4.2, black lines, see Supplementary Table 1 for fit parameters from individual datasets); the best fit parameters are compared (Fig. 4.3).

We found that both RNAPII subunit recoveries were well fit by the anomalous subdiffusion model. This is in contrast to the GFP recovery dynamics which were well fit by Brownian diffusion[20]. Since our GFP experiments have revealed that molecular crowding is not a source of anomalous diffusion and these experiments restricted the analysis to an identical nuclear environment devoid of RNAPII binding sites or membrane induced labyrinthine regions, we can infer that the observed subunit recovery is not true anomalous diffusion.

As another possible source of observed anomalous behavior, we considered that the simultaneous measurement of multiple diffusing species (a distribution) undergoing Brownian motion can produce an identical FRAP recovery profile to a single species undergoing anomalous diffusion [28]. We term this phenomenon *apparent anomalous diffusion*. Thus, we strongly believe that the subunits must be in a heterogeneous distribution of complexes resulting in the observation of apparent anomalous diffusion, as described in section 4.

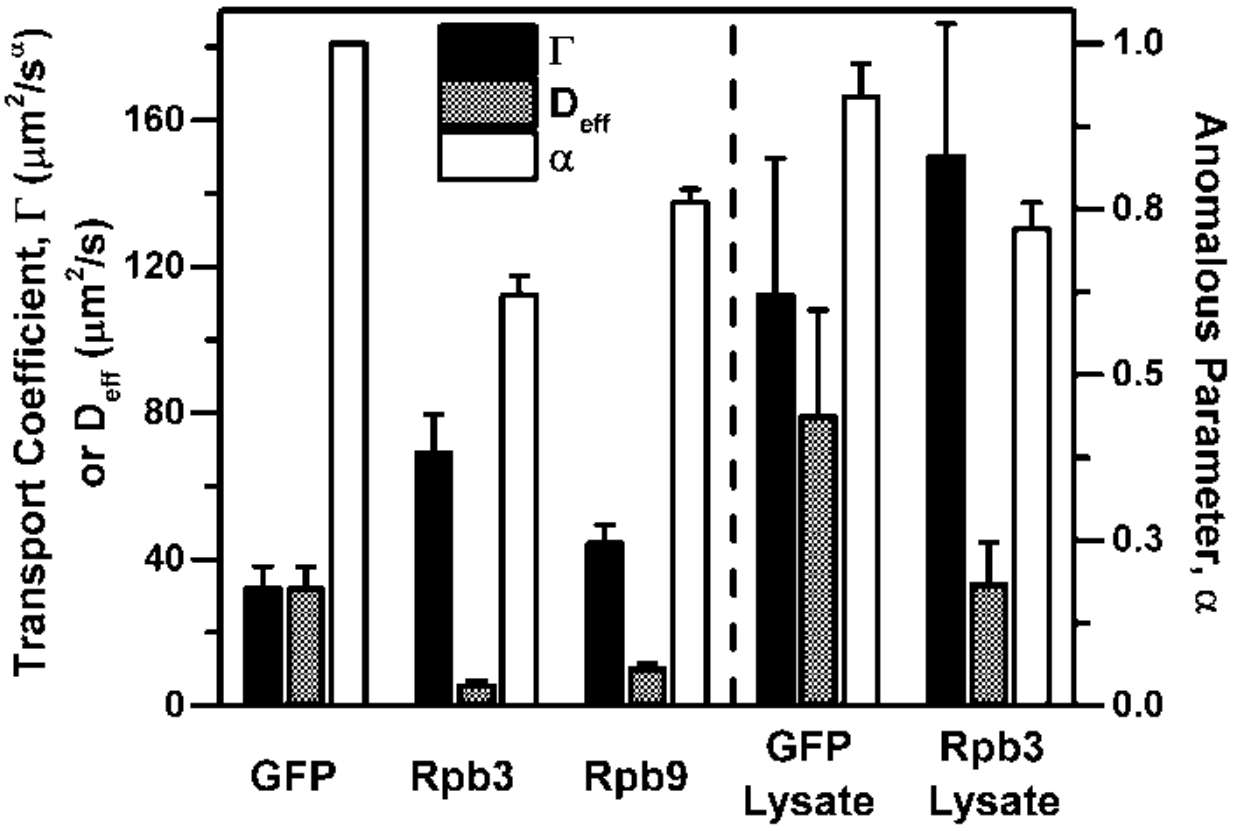

Figure 4.3- Summary of the best-fit apparent anomalous modeling parameters. The alpha value varies between zero and unity and is a measure of deviation from Brownian diffusion. The transport coefficient is measure of translational diffusion speed, the effective diffusion coefficient (D_{eff}) represents the diffusion coefficient if the particle obeyed Brownian diffusion. Error bars are shown at the 95% confidence interval. The GFP expressing line was found to diffuse normally with a diffusion coefficient of 32±6 $\mu m^2/s$. The RNAPII subunits showed apparent anomalous diffusion, with each exhibiting different diffusive kinetics. Rpb3 exhibited an apparent anomlity value of 0.62±0.03 while Rpb9 exhibited an anomlity value of 0.76±0.02. This reveals that the subunits are not bound in identical complexes. To the right of the dotted

line are the parameters for the *in vitro* lysate experiments. Within experimental error, the diffusion of GFP is found to be Brownian and of the same magnitude as GFP in dilute buffer. The Rpb3 lysate continues to indicate apparent anomalous diffusion.

3. Confirming the distribution of heterogeneous RNAPII subunit complexation states

We reasoned if the apparent non-Brownian transport persisted in dilute solution then the deviations from Brownian diffusion must be attributed to a distribution of complexes. To completely eliminate macromolecular crowding as a possible source of anomalous diffusion, we performed FRAP experiments on cellular lysates of the salivary gland polytene cells expressing either GFP or Rpb3 (Fig. 4.4). The cell lysates are whole cell preparations made by sonicating the salivary glands in a lysis buffer and extracting the soluble proteins. The cell contents were centrifuged and the supernatant used for FRAP experiments. A comparison of the fluorescent intensity between the lysates and the intact polytene cells revealed up to a 30-fold decrease insignal. We were unable to collect data on lysates made from Rpb9 due to extremely low sample signal.

The GFP lysate FRAP recovery indicated a normally diffusing species (Fig. 4.4). Further, the diffusion coefficient determined by the FRAP model described in Eq.2 of 79.1±30.0 $\mu m^2/s$, is in excellent agreement with the diffusion of free GFP (purified from bacteria) in solution, measured on our set-up as 84±6 $\mu m^2/s$ [20]. Thus our lysate preparation recapitulated a dilute solute environment by eliminating macromolecular crowding. We note that the GFP lysate yielded a slightly non-Brownian anomlity parameter (Fig. 4.3), which is the result of the very rapid

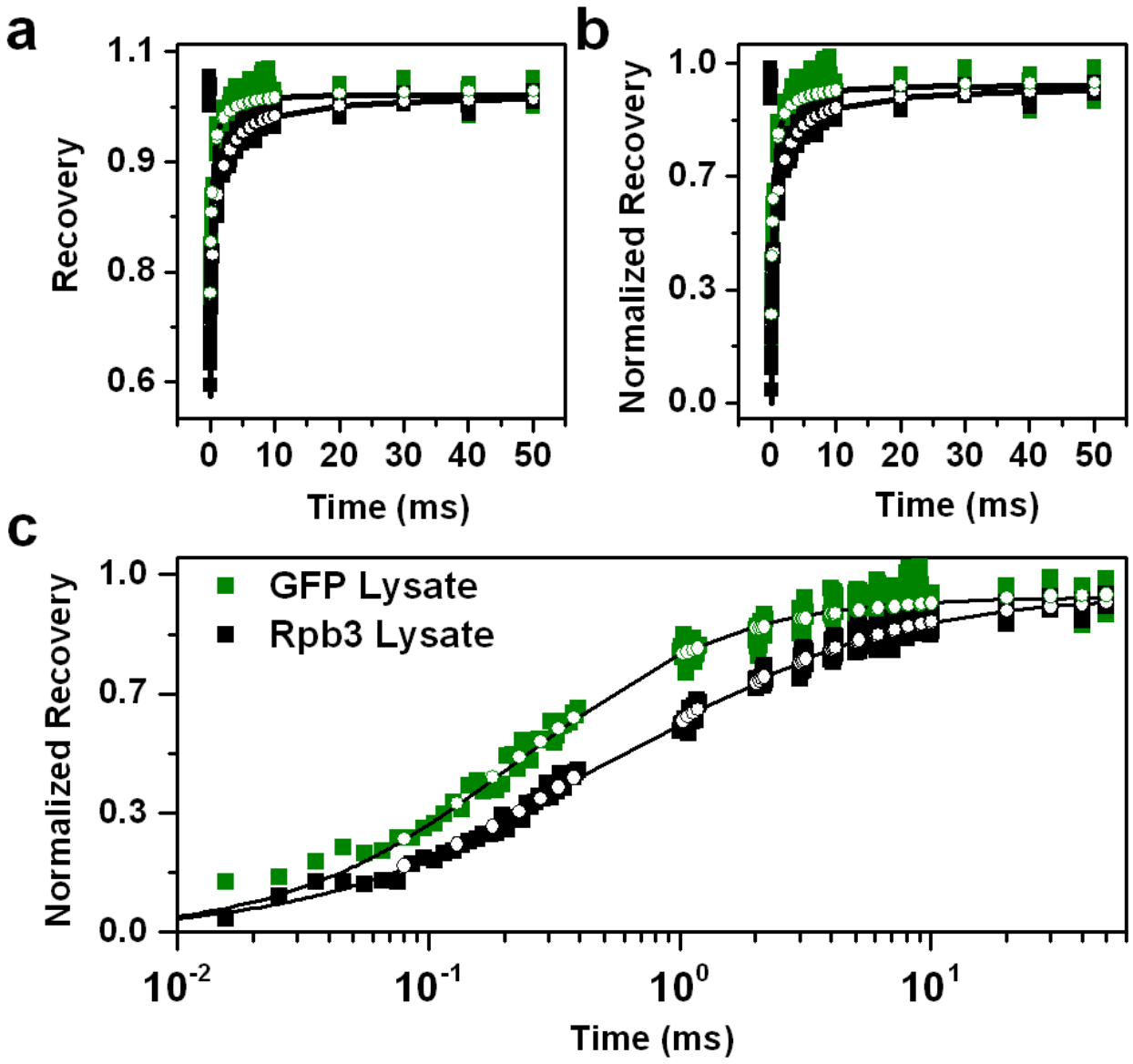

Figure 4.4- Comparison of in vitro subunit recovery dynamics. The FRAP curves for the GFP expressing control line and the Rpb3 subunit lysate experiments are shown. Numerous FRAP curves were averaged for each sample due to low signal intensity of the lysates (GFP- 6090 pts, Rpb3- 17420 pts) (**a,b,c**) Data are plotted as closed squares, the best-fits to an anomalous diffusion model are shown as black lines; best-fits to the Distribution Model are shown are white circles. (**c**) The flattened slope and slower recovery of the Rpb3 lysate is a clear indication that the sample is not undergoing Brownian diffusion.

recovery of the species coupled with low signal strength. Both of these factors reduce the accuracy and precision of the fitting algorithm.

Despite the highly dilute solvent environment, the Rpb3 lysate FRAP recovery reveals very different behavior (Fig. 4.4), displaying apparent anomalous diffusion (Fig. 4.3). Due to the lower viscosity of the lysate solvent, both the transport and effective diffusion coefficients, determined by Eq.1, are increased compared to Rpb3 diffusion *in vivo*. Further, the lysate recovery indicated a reduction in the measured anomlity value (Fig. 4.3). This reduction could stem from very large complexes no longer experiencing crowding effects [25] and reveals the degree of apparent anomlity resulting solely from the distribution of species in the absence of crowding effects. Alternatively, this could indicate the disintegration of complexes that coalesce *in vivo* but destabilize in the absence of molecular crowding.

4. Distribution modeling: decomposing apparently anomalous recovery curves into components exhibiting Brownian diffusion

In any FRAP measurement the observed signal is the sum of the signals from each species present in the sample. In a many component system, if the species have diffusion coefficients that are sufficiently different, it may be possible to distinguish distinct timescales in the recovery. More often, the observed signal takes a form that can appear as anomalous diffusion [29, 30]. In our experimental systems, we observed that GFP exhibits Brownian diffusion in the interchromatin space, but Rpb3 and Rpb9 do not. There is little reason to suggest that individual proteins similar in size to GFP would exhibit true anomalous diffusion. Therefore, we investigated the possibility that each protein species is incorporated into a heterogeneous size-

distribution of macromolecular complexes by applying a multi-component fit to the FRAP recovery that we term the *distribution model*.

The distribution model was implemented as [29]:

$$\mathcal{F}(t) = \sum_{i=1}^{m} c_i F(D_i, t, \alpha = 1) \quad (3)$$

The recorded FRAP recovery, $\mathcal{F}(t)$ is a linear combination of Brownian diffusion basis functions, $F(D,t,\alpha=1)$ that are given by Eq.2 with α=1 and a range of individual diffusion coefficients. The coefficient *c* of each species is allowed to float and the resulting output defines a distribution of species with various diffusion coefficients (the robustness of the distribution model is detailed in Appendix D).

The distribution model was first tested by fitting the *in vivo* FRAP recovery of unconjugated GFP for an underlying distribution (Fig. 4.5a, green). In agreement with the aforementioned fits to the anomalous diffusion model that indicated a single Brownian diffusing component, fits to the distribution model output collapsed to a Delta function, yielding a single diffusion coefficient of 27 $\mu m^2/s$ (peak 1). This is within 15% of our previously determined *in vivo* GFP diffusion coefficient [20]. Having validated the Distribution model (Appendix D), we applied it to the cellular FRAP recoveries of Rpb3 and Rpb9, along with the GFP and Rpb3 lysate data. In general, the breadth of the distribution for each sample qualitatively agrees with its degree of apparent anomalous diffusion. For example, the protein exhibiting less apparent anomlity, Rpb9, exhibits a distribution of species that have Brownian diffusion coefficients in a peak from about 10 through 30 μm2/s (Fig. 4.5-a, red), while the Rpb3

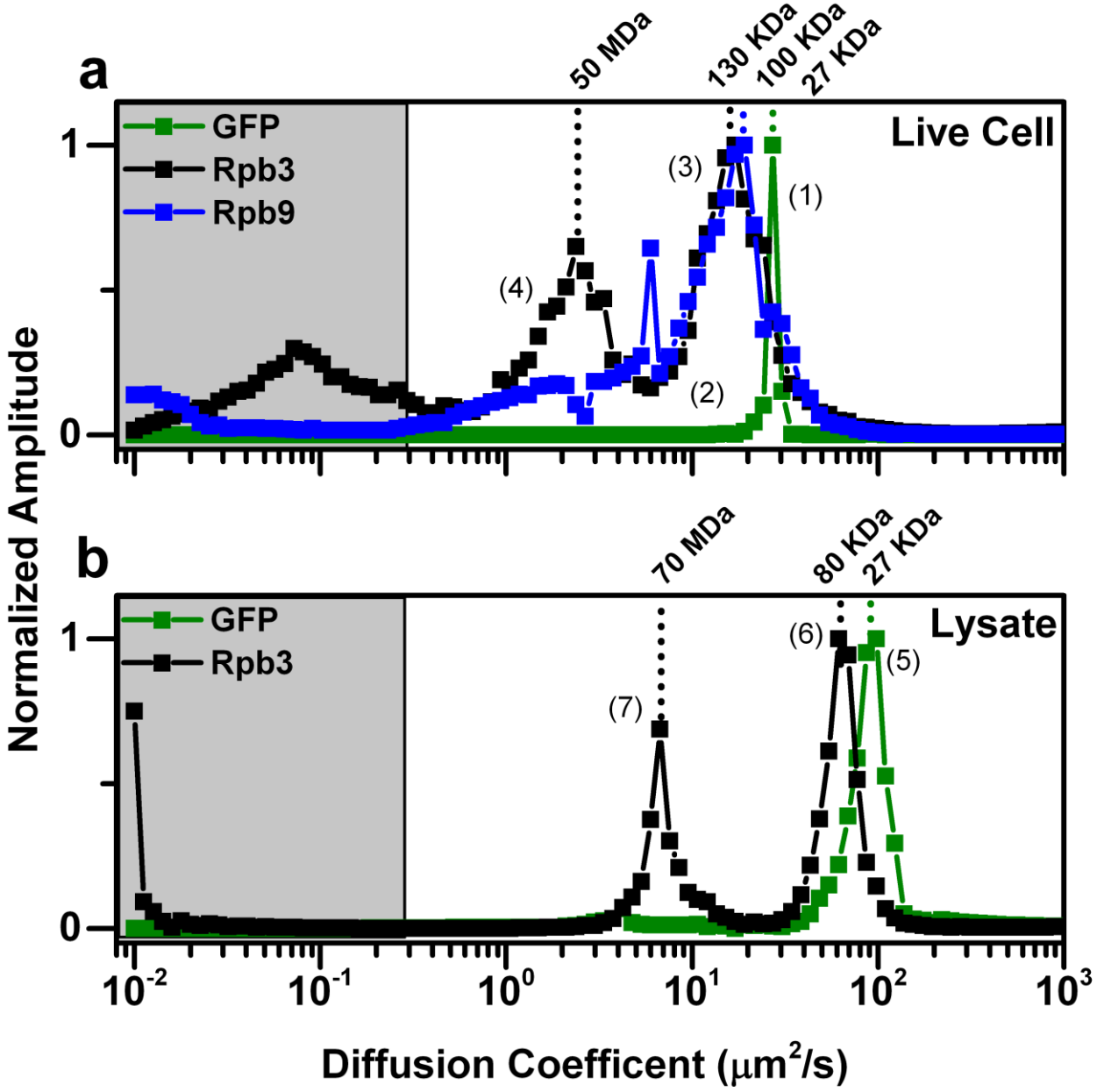

Figure 4.5- Brownian diffusion coefficient distributions. The distribution model (Eq. 4) was applied to *in vivo* (**a**) and (**b**) *in vitro* FRAP recovery curves. To implement the model, we defined 100 species with logarithmically spaced diffusion coefficients ranging from 0.20 to 1000 $\mu m^2/s$. This range of diffusion coefficients corresponds to a massive size range of species. Components with diffusion coefficients slower than 0.29 $\mu m^2/$ are below the limit of the

recovery threshold of our FRAP method. (**a**) The distribution of unconjugated GFP (green) collapses to a delta function with a diffusion coefficient of 27 $\mu m^2/s$. The observation of a single diffusing species demonstrates good agreement with the apparent anomalous diffusion model. The distributions for Rpb3 (black) and Rpb9 (blue) exhibit major peaks at 17 and 18 $\mu m^2/s$ respectively, corresponding at Stokes-Einstein predicted masses of 130±50 and 100±40 kDa respectively. These values are in good agreement with the predicted GFP-fusion construct masses. The Rpb3 distribution is bimodal, with the slower peak indicating a diffusion coefficient of 2 $\mu m^2/s$, mapping to a mass of 50±20 MDa. This peak indicates the presence of fully formed transcription factories. (**b**) The *in vitro* distribution for unconjugated GFP is narrow and indicates a diffusion coefficient of 92 $\mu m^2/s$, in good agreement with measurements of GFP in dilute buffer. The Rpb3 lysate distribution again reveals two well resolved peaks, corresponding to masses of 74±20 MDa and 82±24 kDa, similar to the peaks in the *in vivo* measurements.

exhibits a distribution that is even broader and more structured. However, much more information is contained in the shape of the distributions than is available from the anomlity parameter, as discussed below. Another notable observation about the distributions is that none contains diffusion components faster than unconjugated GFP.

The Stokes-Einstein Equation, which predicts the diffusion coefficient of a particle undergoing Brownian diffusion, can be re-arranged to estimate the relative diffusion coefficients of the proteins (assuming globular structures and the same viscosity) based on their molecular masses:

$$\frac{D_1}{D_2} = \left(\frac{M_2}{M_1}\right)^{1/3} \quad (4)$$

Here, *D* is the protein diffusion coefficient and *M* is the protein molecular mass. Using the molar mass of GFP and measured diffusion coefficient as a standard, the approximate mass corresponding to each diffusion component in the subunit distributions can be estimated using Eq. 4. The peak of the Rpb9 distribution (Fig. 4.5-a, peak 2) corresponds to a mass of 100±40 kDa, reasonable given the 41 kDa mass of the fusion construct (we confirmed that this is independent of protein expression level, Appendix D). The width of the distribution maps to species ranging in molecular mass from 27 kDa through 10^8 kDa. While the enormous upper limit on molecular mass should be viewed with incredulity, these results indicate that species are present ranging from unconjugated GFP through aggregates of multiprotein complexes. The upper mass limit defined by the distribution is unrealistically large and likely reflects components sufficiently large to be influenced by molecular crowding that undergo true anomalous diffusion.

In contrast to Rpb9, the Rpb3 subunit exhibited a wider and more structured distribution (Fig. 4.5-a, black). Interestingly, the distribution is bimodal, with two well-resolved peaks bridged by components of lower amplitude. As expected, the fastest components are bound by an upper limit of diffusion coefficients similar to unconjugated GFP. Assuming Stokes-Einstein, the "faster" peak (Fig. 4.5-a, peak 3) corresponds to a molecular mass of 130±50kDa, in good agreement with mass of the Rpb3-GFP fusion construct. The second, "slower" peak (Fig. 4.5-a, peak 4) corresponds to a mass of 50±20 MDa. The mass of a complete transcription complex [31] consisting of RNAPII and associated transcription factors has been estimated to be ~3 MDa; the mass of full transcription factories (aggregates of full transcription complexes and associated promoters) has been estimated up to ~38 MDa[1, 32]. Thus, the second major peak in the Rpb3 distribution is very close to the size of fully assembled gene transcription units [1, 31, 32]. Its presence indicates that these transcription units are present in the interchromatin space, in the absence of chromatin. We also note that the Rpb9 distribution exhibits a pronounced shoulder in the same range as the 50 MDa peak in the Rpb3 distribution.

The Rpb3 distribution also contains lower frequency components. Our FRAP method is insensitive to species slower than 0.29 $\mu m^2/s$ (Appendix D). These species are likely contributions to the distribution but the true amplitudes are uncertain. Importantly, the fit residuals are better than those produced by the anomalous diffusion model. The quality of the fits can be compared in Figure 4.2 and Figure 4.4, where the white circles indicate the distribution model fits, in comparison to the anomalous diffusion model fits in black.

As a comparison to the *in vivo* distributions obtained for GFP and Rpb3, we applied the Distribution model to the results of the lysate FRAP experiments, keeping the same number of components and the same bounds on diffusion coefficients (Fig. 4.5-b). By eliminating the stabilizing effects of macromolecular crowding, this analysis examines how the distribution of complexes is altered by a dilute solvent. The distribution for the GFP lysate (Fig. 4.5-b, green) indicates a narrow range of diffusion coefficients, with the major peak indicating a diffusion coefficient of 92 $\mu m^2/s$. This is within a 10% error of the previously determined diffusion coefficient of GFP in buffer solution (84±6 um^2/s) [20], confirming that the lysate provides a dilute environment that eliminates macromolecular crowding.

The results for the Rpb3 lysate (Fig. 4.5-b, black) are very similar to the distribution found *in vivo*, except shifted towards faster components due to the reduced solution viscosity. The lysate distribution indicates two major peaks, the "faster" peak at 65 $\mu m^2/s$ and the "slower" peak at 6.7 $\mu m^2/s$. These correspond to masses of 82±24 kDa and 74±20 MDa. Notably, the major peaks detected map to the same molecular masses as the *in vivo* fitting results, providing independent confirmation of the bimodal distribution. However, the lysate distribution differs from the *in vivo* distribution in two important locations. First, the middle range of diffusing components (inter-modal), between the two peaks is absent in the lysate distribution. This indicates that these protein complexes that are present in the crowded nuclear environment destabilize in the dilute solvent. These species, intermediate between complete and incomplete transcription factories have implications for the pre-assembly of transcription complexes. Their presence suggests that the formation of large protein assemblies proceeds through partially-assembled intermediates whose formation is favored in

the crowded nuclear environment. Second, the very slow components that are technically below our FRAP resolution limit are largely absent in the lysate distribution. This supports the suspicion that those components *in vivo* represent complexes sufficiently large to experience macromolecular crowding and truly exhibit anomalous diffusion.

Discussion

1. A new perspective for in vivo diffusion: apparent anomalous diffusion

Our experiments with RNAPII subunits sought to directly probe the nucleoplasm, devoid of chromatin, for evidence of the holoenzyme or larger transcription complexes. We determined that RNAPII subunits exhibit complex transport dynamics even in the absence of chromatin, that can be attributed to a staggeringly large distribution of assembly states, ranging from fully assembled transcription factories to unengaged subunits. The existence of such nuclear assemblies concerns one of the current fundamental dilemmas in modern biology- determining how large DNA-binding protein complexes assemble and subsequently find their binding sites. Recent studies have supported the theory that many DNA binding complexes encounter and bind to chromatin through a stochastic diffusion-mediated process, but little evidence exists to explain what governs the assembly of these multi-component complexes away from binding sites. Given the centrality of RNAP to transcription and possible mechanistic universality with regards to other large nuclear-localized complexes [33], this multi-subunit complex has been the subject of great scrutiny over the past decade.

Information about the assembly and interactions of large protein complexes can be obtained by investigating transport properties of individual components, since protein mobility

not in accordance with Brownian diffusion can indicate the presence of binding interactions or molecular hindrance [22, 34, 35]. Two types of passive transport are typically identified *in vivo*- Brownian motion and anomalous subdiffusion [28, 36, 37]. Given the widespread implementation of FRAP and FCS, it is interesting to note that with very few exceptions [28], the preponderance of eukaryotic proteins studied *in vivo* have been found to exhibit anomalous subdiffusion, while similar sized molecules studied in aqueous or viscous solvents typically have been found to obey Brownian motion [25, 28, 38, 39]

We compared the transport dynamics of the RNAPII subunits Rpb3 and Rpb9 to unconjugated GFP. Suspecting that the chromatin organization of typical eukaryotic cells could pose a potential interference to diffusion mobility, we avoided confounding structures present in the nuclear environment by choosing the polytene salivary glands of *Drosophila melanogaster* larvae as our model system. Our FRAP experiments performed with unconjugated GFP revealed that this inert protein is subject to Brownian diffusion. Nuclear molecular crowding was experienced as a change in viscosity resulting in a reduction of the diffusion coefficient of GFP from 84± 6 $\mu m^2/s$ in dilute solvent to 32±6 $\mu m^2/s$ in *Drosophila* cells. In contrast to GFP, we observed apparent anomalous diffusion for both RNAPII subunits. This is very surprising as the approximately two-fold increase in molecular mass of the fusion proteins relative to GFP would be expected to yield a very minor 1.2-fold change in diffusion coefficient based on Stokes-Einstein estimations (Eq. 4). This is hardly a large enough increase in size to make either subunit susceptible to extreme molecular crowding. Having eliminated all other contributions to anomalous diffusion, we have shown that molecular crowding is not a cause of anomalous diffusion for proteins in this size range. Therefore, we reason that the subunits are

actually engaged in distributions of complexes displaying an extremely large range of diffusion coefficients and therefore molecular sizes. We term this phenomenon *apparent anomalous diffusion*.

Apparent anomalous diffusion was suggested in the 1990s and experimentally confirmed to affect FRAP curves by using simple two component systems with inert solutes [29, 30, 40]. These previous groups demonstrated that multicomponent FRAP recovery curves of Brownian diffusing species can be represented by an anomalous fit, but this was not confirmed in a living system until now. Our experiments simultaneously probe the diffusion of assemblies with vastly different mobilities, from isolated subunits to possible aggregates of fully formed transcription units. Observed differences in the recovery dynamics of the two subunits (Fig. 4.2) indicates that they participate in different distributions of complexes (Fig. 4.5). This reflects differential affinities for the other RNAPII subunits and associated transcription factors, as well as suggesting that distribution width and subunit incorporation sequence are entwined.

We further explored the cellular transport behavior by performing FRAP experiments on *in vitro* lysates prepared from the GFP and Rpb3 polytene samples (Fig. 4.4). The diluted solvent abolished macromolecular crowding and ensured that the proteins did not experience crowding effects or find binding partners. This left only a distribution of diffusing species as the remaining source of perceived anomalous diffusion [37]. The results indicate that many of the Rpb3 complexes remained intact during the lysate preparation, since it still exhibited apparent anomalous diffusion (Fig 4.3).

It has been reported previously that the extent of anomalous diffusion can be used as a measure for environmental heterogeneity [19]. We argue that having shown that interchromatin space represents a homogenous diffusive environment, the degree of anomlity can instead be a proxy for the width of the distribution in which the tagged protein participates. This makes intuitive sense- if an anomlity factor of unity represents normal diffusion and therefore a single diffusing component, any departure from unity is describing an increasingly heterogeneous mixture. We found the Rpb3 subunit was associated with the highest degree of apparent anomalous diffusion (Fig. 3) indicating it participates in the widest size range of complexes (Fig. 4.5). The Rpb9 subunit was found to exhibit less apparent anomlity (Fig. 4.3), corresponding to a more narrow distribution (Fig. 4.5), while GFP, which does not interact with any other species, was found to show normal diffusion.

We applied a multi-component model to extract the underlying distributions of nuclear Rpb3 and Rpb9 to determine their participation in pre-assembled RNAPII complexes. The distribution model is advantageous as no *a priori* assumptions about the underlying distribution are made, thus protein complex sub-populations can be resolved. In reality, this model faces three limitations. The model assumes all component species obey Brownian diffusion- it is unable to resolve simultaneous diffusion of Brownian and anomalous species. Secondly, the application of the model is affected by the quality of the data. As reported by others [29, 30] the SNR of the data impacts the ability of the model to accurately resolve separate species, even in well resolved binary systems. Our implementation is sufficient to reliably predict two components at our experimental SNR, yet the potential complexity of the protein distributions means that discerning fine structure of sub-populations is difficult. Finally, our FRAP

implementation poses a resolution limit on how slowly diffusing a species we can accurately measure.

As anticipated, the comparison of the Rpb3 and Rpb9 distributions confirm that the greater the degree of apparent anomalous diffusion (Fig. 4.3), the wider the predicted distribution (Fig. 4.5-a). We can immediately detect that the Rpb3 subunit is involved in a wider array of complexes than Rpb9, with more of them involving very large molecular weight assemblies. The distribution modeling of the Rpb3 lysate reveals essentially the same structure, though shifted to faster diffusion components due to the reduced solvent viscosity. This provides two different experimental samples that confirm that same finding. Significantly, the more massive population is identical between both samples and corresponds to overlapping molecular mass ranges of 50±20 MDa *in vivo* and 70±20 MDa *in vitro*. Given the several mega-Dalton mass of a complete transcription complex [31] and the much larger mass of transcription factories [1, 32], this population represents a fully assembled transcription factory. Such complexes likely arise given the affinities between transcription complex subunits and the crowded cellular environment in which they dwell, meshing well with reports that transcription factories remain even in the absence of transcription [41].

While the envelope shape of Rpb3 associated complexes is preserved in the lysate preparation (Fig. 5), it is noteworthy that the majority of the *in vivo* distribution components lying between the major peaks are eliminated in the lysate distribution. These represent dynamic complexes that are stabilized in the crowded nuclear environment, where dissociation and re-binding is rapid due to partner proximity. In the dilute lysate solvent, once a complex of

low stability dissociates, rebinding is inhibited by the low concentration of binding partner. Further, the width of both peaks is similar to the width of the GFP peak. This indicates the remaining species show less dispersion. Finally, the lysate data does not exhibit the same structures at very slow diffusion coefficients (mapping to greater than a GDa), possibly an indication that Brownian diffusion was restored for very large complexes affected by macromolecular crowding.

2. RNAPII distributions indicate an intermediate assembly mechanism

Previous work has established the dynamic turnover of RNAPI and RNAPII associated proteins during transcription. It has been shown that four subunits of RNAPI as well as several preinitiation factors all exhibit unique diffusion properties even in the vicinity of chromatin and do not diffuse as an ensemble. Further, engaged RNAPII has been found to continuously exchange with nucleoplasmic RNAPII in transcriptionally active chromatin regions [8, 9, 16, 42, 43]. These findings have led to the developing consensus that complexes assemble at a promoter site through stochastic interactions. However, the continued evidence for the formation and stability of fully assembled transcription factories even in the absence of transcription throws uncertainty on the spatiotemporal formation of such assemblies [7, 9, 13, 21, 44]. Unfortunately, previous studies could not track the dynamics of the RNAP subunits prior to recruitment or localization.

Using our method which is sensitive to the diffusion, and therefore mass of a complex, but not to the activity state, our experiments have probed the dynamics of multiple subunits within the same binding complex, enabling us to observe the degree of pre-assembly. This is

significant as our analysis was restricted to the interchromatin space, representing a cellular location that we found to precede incorporation of all subunits into higher order assemblies but that follows subunit mRNA translation. Our work has shown that two subunits of RNAPII, including the central binding subunit Rpb3, exhibit different diffusion dynamics (Fig. 4.2). This casts doubt on *complete* pre-assembly of all RNAPII substituents prior to chromatin binding [6-13]. For both subunits, we detect a subpopulation of molecular complexes approaching a limit of a hundred mega-Daltons (Fig. 4.5), which corresponds to aggregates of fully assembled transcription factories. This indicates that transcription complex subunits have high affinities that experience enhanced stability conferred by the crowded cellular environment in which they dwell.

These distributions indicate that the formation of large protein complexes is driven by stabilizing interactions even in the absence of chromatin, yet this subpopulation does not account for all of the RNAPII subunits present within the interchromatin space. This has implications for large multi-complex assembly pathways, as stochastic protein-chromatin interactions can be reframed in terms of sampling interactions between complexes in various states of completeness. Such a model is at odds with the more static, top-down view of factory formation. While our results clearly indicate that large macromolecular complexes, such as transcription factories, are stable *in vivo*, the unanswered question is for how long they remain assembled. Most studies documenting transcription factories have relied on the appearance of punctate structures observed in fixed cells or on the purification of stable transcription complexes *in vitro*[4, 14]. Additionally, electron microscopy measurements that document the size of these complexes place an upper limit of <200 nm in diameter, still too small to accurately

resolve with optical microscopy on living cells [32]. These complicating factors, combined with our findings of the stability of large protein complexes *in vitro*, make it difficult to determine the longevity of these species.

As investigations into the dynamics of polymerase components and associated transcription factors reveal a conserved intrinsic turnover and universally accepted inefficiency of transcription initiation, the previously posited model of stochastic gene expression has gained traction [7, 44]. Mounting evidence indicates that RNAPII is not always recruited as a holoenzyme, though our findings clearly indicate that full transcription factories do form prior to RNAPII recruitment [44]. RNAPII is currently seen as assembling at a promoter through a multi-step process marked by efficient chromatin capture rates of up to 50% [9] but highly inefficient transcription initiation (<1%)[10], leading to an overall transient promoter interaction prior to elongation (which is unlikely if full transcription factories migrated throughout nucleus).

We believe our findings of RNAPII subunits existing in complex distributions lend validity to both models. Our essential finding is that transcription subunits form large, stable, and mobile complexes, indicating the true assembly behavior lies mid-way on a spectrum of pre-assembly. We measured diffusion coefficients for transcription factories in line with those determined for other proteins involved in nuclear macromolecular assemblies [42]. This suggests that large complexes are mobile (but slow) and can diffuse to binding sites, in contrast to static factory models in which chromatin must migrate to stationary factories. This integrates well with current observations, but helps to redefine the nature of assembly. Our results provide experimental evidence to considerations proffered by Phair and Misteli that protein complexes

can form stochastically, distal to their site of action, enabling rapid recruitment and dynamic responses to changes in binding partner availability [7, 42-44]. However, the large population of individual subunits and partially-formed complexes also allows for *de novo* assembly at gene loci.

As opposed to a hit-and-run model of polymerase factors encountering a chromatin binding site, our findings show that transcription complexes assemble to varying levels of completion in the interchromatin space removed from and prior to encountering chromatin. These partially formed assemblies, through diffusion, experience stochastic encounters with potential binding sites; the duration of the encounter depending on the completeness of the polymerase assembly. More complete RNAPII complexes, having a greater compliment of binding partners, form more stable chromatin interactions than less well developed sub-assemblies. As our distribution modeling shows, the majority of the subunits exist as incomplete assemblies, therefore the majority of chromatin interactions are likely aborted, leading to the inefficiency of transcription initiation. Our observation of a bias towards larger complexes exhibited by the more massive RNAPII Rpb3 (Fig. 4.5) subunit may reveal a measure of stepwise assembly. In this scenario, the larger subunits complex first, leading to stable chromatin-binding assemblies, forming nucleation sites for smaller subunit assemblies. Such a model ensures maximum flexibility in gene expression for different chromatin regions. The two assembly regimes we observe mean that fully formed transcription complexes, in the presence of open chromatin regions are likely to remain stably assembled and engage in high throughput transcription. These large structures experience slow diffusion and would remain relatively stationary, in alignment with transcription factory theory. Conversely, the smaller sub-

assembled modules, which account for a large fraction of the assembly states, are capable of rapid diffusion and permit protein recruitment to congested chromatin regions that experience lower basal transcription levels. The partial pre-assembly of the transcription complex enhances the efficiency of full complex assembly and is complimented by greater nuclear mobility than near-immobile transcription factories. Thus through a partially modular assembly mechanism the cell is endowed with a flexible response to changing transcription demands.

Additionally, while not the focus of this work, we have previously observed true anomalous diffusion due to confinement in the vicinity of the chromatin lattice even for small proteins [20]. Coupled with the findings of other researchers concerning the role of molecular crowding in gene expression [45,46,47], it stands to reason that large, partially assembled complexes, once in the vicinity of a promoter, sample increasingly frequent binding events due to molecular confinement and reduced mobility.

Conclusion

By applying FRAP in the polytene salivary glands of *Drosophila melanogaster* as a model system, showing for the first time that RNAPII exists in a large distribution of partially assembled complexes in the interchromatin space, including fully assembled transcription factories. By determining that the Rpb3 and Rpb9 subunits exhibit different diffusion properties, we confirm that RNAPII is a dynamic complex, though we detect a population of complete pre-assembled transcription factories prior to chromatin binding. Using GFP as an inert internal control protein, we have shown *in vivo* that the diffusion of the subunit distributions display apparent anomalous diffusion. This arises from the simultaneous

interrogation of multiple diffusing species using an ensemble measurement method. When considered individually, these complexes move primarily by Brownian diffusion throughout the crowded interchromatin space, experiencing a reduction in mobility due to the high viscosity but not experiencing molecular confinement. We confirmed the existence of these subunit assembly distributions through the use of cell lysates, in which apparent anomalous diffusion persisted in the absence of macromolecular crowding. The discovery of these partially assembled RNAPII complexes helps integrate current contradictory observations regarding the mode of transcription complex assembly. Our findings are consistent with the simultaneous action of a top-down and bottom-up assembly. While the exact nature of the species that initiate transcription cannot yet be determined, for the first time our data shows evidence for a distribution of pre-assembled complexes. Finally, the distribution of assembly states suggests that a partially modular mechanism of macromolecular assembly enables a flexible response to gene transcription.

REFERENCES

(1) Melnik, S.; Deng, B.; Papantonis, A.; Baboo, S.; Carr, I. M.; Cook, P. *Nature Methods* **2012***, 8*, 963.

(2) Kruhlak, M. J.; Lever, M. A.; Fischle, W.; Verdin, E.; Bazett-Jones, D. P.; Hendzel, M. J. *The J. Cell Biol.* **2000***, 150*, 41-52.

(3) Meister, P.; Poldevin, M.; Francesconi, S.; Tratner,Isabelle, Zarzov, Patrick; Baldacci, G. *Nucleic Acids Res.* **2003***, 31*, 5064.

(4) Hemmerich, P. *Zellbiologie* **2005***, 31*, 18.

(5) Houtsmuller, A. B.; Vermeulen, W. *Histochem Cell Biol* **2001***, 115*, 13.

(6) Cook, P. R. *J. Mol. Biol.* **2010***, 395*, 1-10.

(7) Misteli, T. *Science* **2001***, 291*, 843-847.

(8) Gorski, S. A.; Snyder, S. K.; John, S.; Grummt, I.; Misteli, T. *Mol. Cell* **2008***, 30*, 486-497.

(9) Dundr, M.; Hoffmann-Rohrer, U.; Hu, Q.; Grummt, I.; Rothblum, L. I.; Phair, R. D.; Misteli, T. *Science* **2002***, 298*, 1623-1626.

(10) Darzacq, X.; Shav-Tal, Y.; de Turris, V.; Brody, Y.; Shenoy, S. M.; Phair, R. D.; Singer, R. H. *Nat Struct Mol Biol* **2007***, 14*, 796-806.

(11) Yao, J.; Ardehali, M. B.; Fecko, C. J.; Webb, W. W.; Lis, J. T. *Mol. Cell* **2007***, 28*, 978-990.

(12) Chen, D.; Dundr, M.; Wang, C.; Leung, A.; Lamond, A.; Misteli, T.; Huang, S. *J. Cell Biol.* **2005***, 168*, 41-54.

(13) Schneider, D. A.; Nomura, M. *Proc. Nat. Acad. Sci. U.S.A.* **2004***, 101*, 15112-15117.

(14) Hannan, R.D., Cavanaugh, A.,Hempel, W.M., Moss, T., Rothblum, L. *Nucleic Acids Res.* **1999***, 27*, 3720.

(15) Grummt, I. *Genes Dev* **2003***, 17*, 1691.

(16) Kimura, H.; Sugaya, K.; Cook, P. R. *J. Cell Biol.* **2002***, 159*, 777-782.

(17) Politi, A.; Moné, M. J.; Houtsmuller, A. B.; Hoogstraten, D.; Vermeulen, W.; Heinrich, R.; van Driel, R. *Mol. Cell* **2005***, 19*, 679-690.

(18) Yao, J.; Zobeck, K. L.; Lis, J. T.; Webb, W. W. *Methods* **2008***, 45*, 233-241.

(19) Weiss, M.; Elsner, M.; Kartberg, F.; Nilsson, T. *Biophys. J.* **2004***, 87*, 3518-3524.

(20) Daddysman, M. K.; Fecko, C. J. *J Phys Chem B* **2013***, 117*, 1241-1251.

(21) Yao, J.; Munson, K. M.; Webb, W. W.; Lis, J. T. *Nature* **2006***, 442*, 1050-1053.

(22) Sprague, B. L.; McNally, J. G. *Trends in Cell Biol.* **2005***, 15*.

(23) Zobeck, K. L.; Buckley, M. S.; Zipfel, W. R.; Lis, J. T. *Mol. Cell* **2010***, 40*, 965-975.

(24) Acker, J.; de Graaff, M.; Cheynel, I.; Khazak, V.; Kedinger, C.; Vigneron, M. *J. Biol. Chem.* **1997***, 272*, 16815-16821.

(25) Seksek, O.; Biwersi, J.; Verkman, A. S. *J. Cell Biol.* **1997***, 138*, 131-142.

(26) Sokolov, I. M. *Soft Matter* **2012***, 8*, 9043-9052.

(27) Brown, E. B.; Wu, E. S.; Zipfel, W. R.; Webb, W. W. *Biophys. J.* **1999***, 77*, 2837-2849.

(28) Dix, J. A.; Verkman, A. S. *Ann. Rev. Biophys. Biomol. Struct.* **2008***, 37*, 247.

(29) Periasamy,N., Verkman, A.S *Biophys. J.* **1998***, 75*, 557.

(30) Gordon, G., W.; Chazotte, B.; Wang, X. F.; Herman, B. *Biophysical Journal* **1995***, 68*, 766.

(31) Wilson, C. J.; Chao, D. M.; Imbalzano, A. N.; Schnitzler, G. R.; Kingston, R. E.; Young, R. A. *Cell* **1996***, 84*, 235-244.

(32) Eskiw, C.; Fraser, P. *J. Cell Sci.* **2011***, 124*, 3676.

(33) Hager, G.; Elbi, C.; Becker, M. *Genes Dev.* **2002***, 12*, 137.

(34) Mueller, F.; Wach, P.; McNally, J. G. *Biophys. J.* **2008***, 94*, 3323-3339.

(35) Feder, T. J.; Burst-Mascher, I.; Slattery, J. P.; Baird, B.; Webb, W. W. *Biophysical Journal* **1996***, 70*, 2767-2773.

(36) Malchus, N.; Weiss, M. *J. Fluo.* **2010***, 20*, 19.

(37) Stasevich, T. J.; Mueller, F.; Michelman-Ribeiro, A.; Rosales, T.; Knutson, J. R.; McNally, J. G. *Biophys. J.* **2010***, 99*, 3093-3101.

(38) Fushimi, K.; Verkman, A. S. *J. Cell Biol.* **1991***, 112*, 719-725.

(39) Mika, J. T.; Poolman, B. *Curr. Opin. Biotechnol.* **2011***, 22*, 117-126.

(40) Hauser, G. I.; Seiffert, S.; Oppermann, W. *J. Microsc.* **2008***, 230*, 353-362.

(41) Mitchell, J. A.; Fraser, P. *Genes Dev.* **2008***, 22*, 20.

(42) Phair, R. D.; Misteli, T. *Nature* **2000***, 404*, 604.

(43) Hager, G.; McNally, J. G.; Misteli, T. *Mol. Cell* **2009***, 35*, 741.

(44) Misteli, T. *Cell* **2007***, 128*, 787.

(45) Guigas, G.; Weiss, M. *Biophys. J.* **2008***, 94*, 90-94.

(46) Hancock, R. *J. Struct. Biol.* **2004***, 146*, 281-290.

(47) Minton, A. P. *Curr. Opin. Struct. Biol.* **2000***, 10*, 34-39.

CHAPTER 5

DETERMINING THE UNDERLYING DISTRIBUTIONS OF MULTIPLE SIMULTANEOUS DIFFUSING SPECIES FROM FRAP SIMULATIONS

"There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact."

-Mark Twain

Overview:

All recorded Fluorescence Recovery after Photobleaching (FRAP) signals are summations of the diffusive recovery profiles of all species in solution with the same fluorescent tag. Oftentimes FRAP recoveries are assumed to correspond to a single tagged species, and for many artificial systems this is a valid assumption. However, when considering biological systems, this assumption may break down, as fluorescently tagged proteins may form homo- or hetero-complexes *in vivo*. In such cases, the recorded FRAP profiles no longer correspond to the protein of interest directly, but encode information about the binding states of all possible complexes formed. The following work considers FRAP profiles for several biologically relevant distributions of complexes, and reports the accuracy of predicting the underlying distributions.

Introduction

FRAP microscopy is a powerful perturbative optical technique useful in interrogating the diffusive properties of biological systems[1]. In a typical FRAP experiment, a small region of

fluorescent intensity is abolished by a strong laser pulse, and the recovery of fluorescent intensity in the region due to the influx of unbleached fluorophores is monitored over time[2]. Fluorescently tagged proteins of interest can be introduced into a cell by exogenous uptake followed by a coupling reaction, or through endogenous genetic encoding. Thus it is currently possible to isolate any protein for investigation with application of a fluorescent tag[3]. FRAP experiments are generally nondestructive to the biological sample under consideration, can be performed in any region of the cellular interior, and can provide sophisticated insight into the kinetic properties of the protein under study[4]. In quantitative FRAP implementation, when the shape of the laser bleach pulse is well characterized, diffusion and binding models can be applied, enabling extraction of detailed information, such as the diffusion coefficient, type of diffusive process, and duration of binding events[5, 6]. Such experiments have been performed extensively in the nuclei and cytoplasm of many cellular samples, and are largely responsible for the current models of dynamic transcription factor binding and transient assembly of gene metabolism complexes[7-9].

A common assumption in nearly all FRAP experiments is that a single fluorescently tagged protein is bleached during the experiment and the recorded recovery profile corresponds exclusively to the diffusion of that species. However, most proteins in biological systems do not exist in isolation, rather they dynamically participate in macromolecular assemblies, sampling a variety of binding configurations. In such cases, FRAP experiments record the summation of the recovery profiles of all the species containing the tagged protein of interest[10]. By decomposing such FRAP profiles into the underlying distribution of complexes, detailed biological information about the complexation states of the protein can be extracted[11].

Previous work by the Fecko lab implemented a distribution model capable of decomposing an experimental FRAP curve into an underlying distribution (Chapter 4). This work established that the subunits of RNA Polymerase II exists in a broad distribution, but also indicated that the accuracy of the model is subject to signal to noise restrictions. Here we explore the accuracy of this distribution model to predict underlying, biologically relevant distributions from simulated FRAP data corresponding to several experimental conditions. In particular, the reduction in accuracy with varying single to noise levels as well as capturing an incomplete FRAP recovery time course are considered. Both conditions are paired with simulated FRAP data representing a simple binary mixture and a gamma distribution of diffusing species.

Computations

1. The Distribution Model:

The distribution model assumes that the recorded FRAP recovery is composed of linear combinations of an underlying basis set of Brownian diffusing species. The contribution to the recorded signal by each species is scaled by the concentration of that species. The distribution model was implemented as[10]:

$$\mathcal{F}(t) = \sum_{i=1}^{m} c_i F(D_i, t, \alpha = 1) \quad (1)$$

The recorded FRAP recovery, $\mathcal{F}(t)$ is a summation of Brownian diffusion basis functions[5], $F(D,t,\alpha=1)$ and a range of individual diffusion coefficients. The coefficient *c* of each species is allowed to float and the resulting output defines a distribution of species with various diffusion coefficients. In this implementation, a basis set of 100 Brownian species with logarithmically

spaced diffusion coefficients from 0.01 to 100 $\mu m^2/s$ was generated. This provides for consideration of a wide range of diffusive components while limiting the computational burden. Further, the bleach depth of all components is assumed to be identical. The MatLab function *lsqnonlin* is used to establish a best-fit distribution to a simulated input FRAP curve.

2. Distributions of Diffusing Species:

The distribution model was used on two underlying distributions- a binary mixture and a continuous distribution defined by a gamma function. For the binary mixture, FRAP recovery profiles for two components of equal concentration with diffusion coefficients of 30 and 3 $\mu m^2/s$ were simulated. For these two profiles, three signal-to-noise levels were simulated, by adding 15, 35, or 50 dB white Gaussian noise using the MatLab *awgn* function (Fig. 5.1, top panels). The FRAP profiles were scaled by their relative concentrations (50%) and added together, resulting in simulated experimental FRAP curves that each encode two diffusing species.

To test a complex, biologically relevant distribution a gamma function (Eq. 2, below) was used to define the amplitudes of the components in the underlying distribution. The gamma function is asymmetric with a steep rise at large values of x (rapidly diffusing components) and a long, monotonically decreasing tail to small values of x (very slow diffusing components). This is well suited to model a biological distribution in which a single tagged protein is represented by the large x-value cutoff, while complexes of increasingly large size, with correspondingly lower diffusion coefficient, are represented by the tail to small values of x. The gamma function is defined as[12]:

$$F(x,k,\theta) = \frac{x^{k=1}e^{\frac{-x}{\theta}}}{\theta^{k}\Gamma(k)}; \text{ x>0, k,θ>0} \quad (2)$$

Where *x* is the dependent variable, *k* is the shape parameter and defines the width of the function, ϑ is the scale parameter defining the magnitude of the function, and Γ is the gamma function. Using this function, a distribution was created that defined the amplitudes of the 100 log-spaced input diffusion coefficients. The amplitude of each component in the distribution was used as the scaling value for each of the FRAP recovery profiles in the basis set; the summation of the basis set yielded a simulation of an experimental FRAP curve comprised of 100 individual species. Again, three single to nosie ratios (SNR) were considered. The entire basis set was modulated with either 15, 35, or 50 dB white Gaussian noise prior to scaling and summation. The result was three different experimental FRAP curves representing the same underlying distribution, but with different SNRs (Fig. 5.2, top panels).

For both simulated distributions, 100 FRAP curves were simulated for each SNR condition, and fit by the distribution model (Eq.1).

3. Incomplete FRAP Recovery Simulations

For both qualitative and quantitative FRAP implementations, it is important to capture the full extent of the FRAP recovery. Incomplete FRAP recoveries are typically designated as immobile fractions, and are thought to represent a portion of the protein population that does not diffuse and is immobile on the timescale of the experiment[7, 13, 14]. Unfortunately, inclusion of an immobile fraction in a quantitative FRAP analysis complicates data fitting. It is nearly impossible to distinguish between a Brownian recovery with an immobile fraction and an anomalous diffusive component

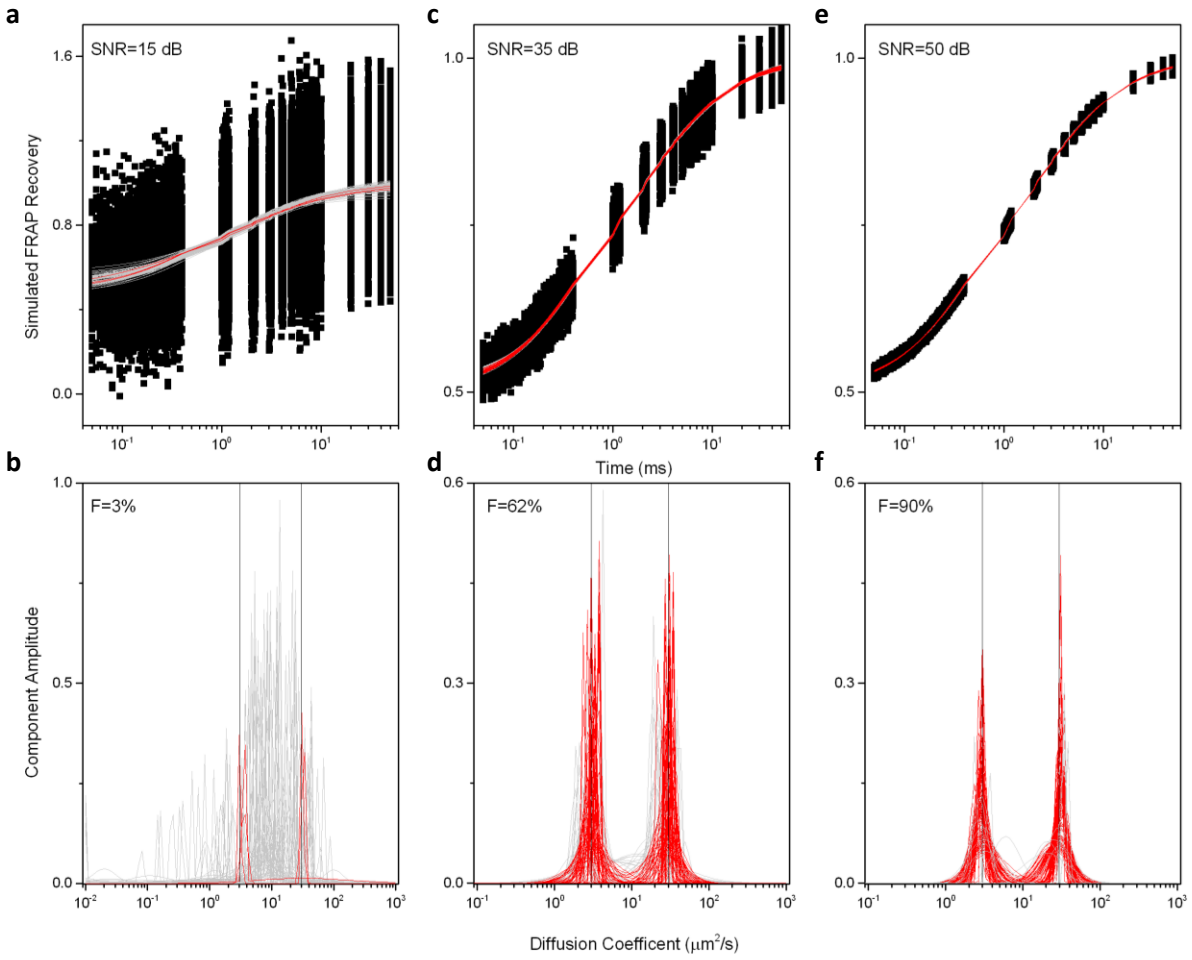

Figure 5.1- Extracting a binary mixture from a simulated FRAP curve at different SNR.(a,c,e) The top panels show simulated FRAP data with overlays of 100 different simulations for each condition. The best-fit lines resulting for each predicted distribution are overlaid, gray lines are failed distributions and red lines are passing distributions. **(b,d,f)** The bottom panels show the resulting distributions from fitting the simulated FRAP curves in the top panels. In each panel, black lines indicate the true diffusion coefficients of the input species (30 and 3 $\mu m^2/s$), gray lines the failing distributions, and red lines the passing distributions. The pass rate is indicated by the Fidelity number, *F*. The 50 dB SNR data can reliably be decomposed into the binary components. The 35 dB SNR data does not accurately predict both the amplitude and diffusion

coefficient, but always detected the bimodal structural of the input distribution. The 15 dB SNR data is not useful for analysis.

recovery of a more constrained (slow moving) species. Both results have vastly different biological interpretations, yet cannot be well resolved by data fitting[14, 15]. While an ideal experiment would have sufficient time resolution and duration to capture the full extent of the FRAP recovery, even for slow moving species, this is not always experimentally feasible.

To study such experimental realities, the FRAP recoveries generated from the two differing underlying distributions were truncated at either 90%, 85%, or 80% of the full recovery. These truncations represent an experimentally observed immobile fraction of 10%, 15%, or 20%, yet are artifacts resulting from the incomplete time course of the simulation (Fig.5.3, Fig. 5.4, Fig.5.5, and Fig.5.6, top panels). Notice that the truncations contain fewer datapoints as the immobile fraction increases. For both the underlying binary and gamma distributions that comprise the FRAP curves, data was simulated with the addition of 35 dB and 50 dB white Gaussian noise SNR. Again, 100 FRAP simulations were generated for each distribution at each SNR, and fit with the distribution model.

Results and Discussion

The distribution model was tested on FRAP curves corresponding to three different SNR, three different extents of recovery, and two different underlying distributions. The output distributions were judged on how accurately they represented the underlying distributions. For the binary mixtures, an accurate distribution was required to predict exactly two componentsand estimate the true diffusion coefficients and relative amplitudes to within a 20% error. For the gamma distributions, the output distributions were required to estimate both the scale and shape parameters (k and ϑ) to within 20% error. Fidelity scores were

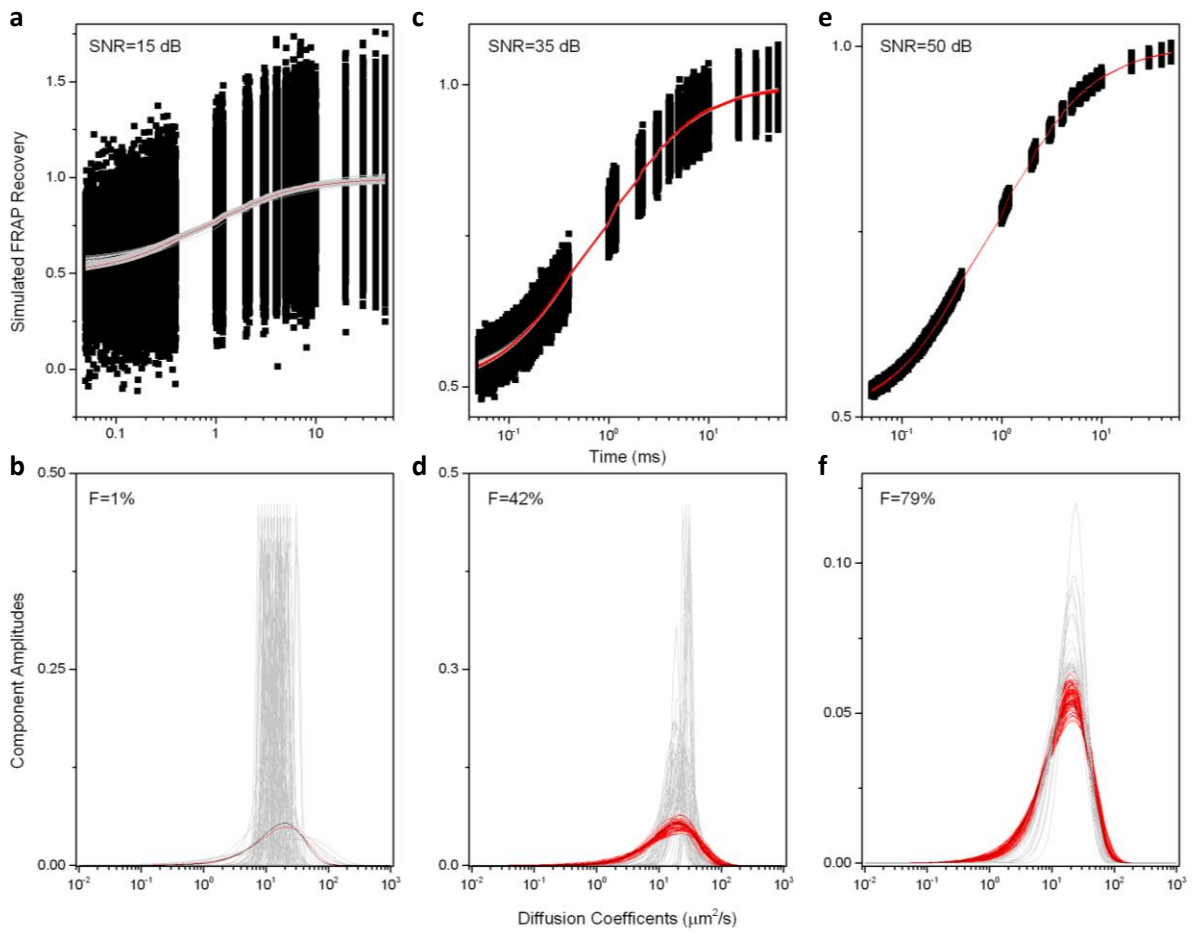

Figure 5.2-Extracting a gamma distribution from a simulated FRAP curve at different SNR. The top panels indicate the simulated FRAP data.The bottom panels indicate the passing (red) and failing (grey)distributions resulting from the data fitting. The input Gamma distribution is show are a black line. The 50 dB SNR data can be accurately decomposed into the underlying gamma distribution. Both of the lower SNR datasets pose fitting challenges, and middle-range components are over-selected. The 35 dB datasets indicates the width of the true distribution, but does not represent the amplitudes of the components properly.

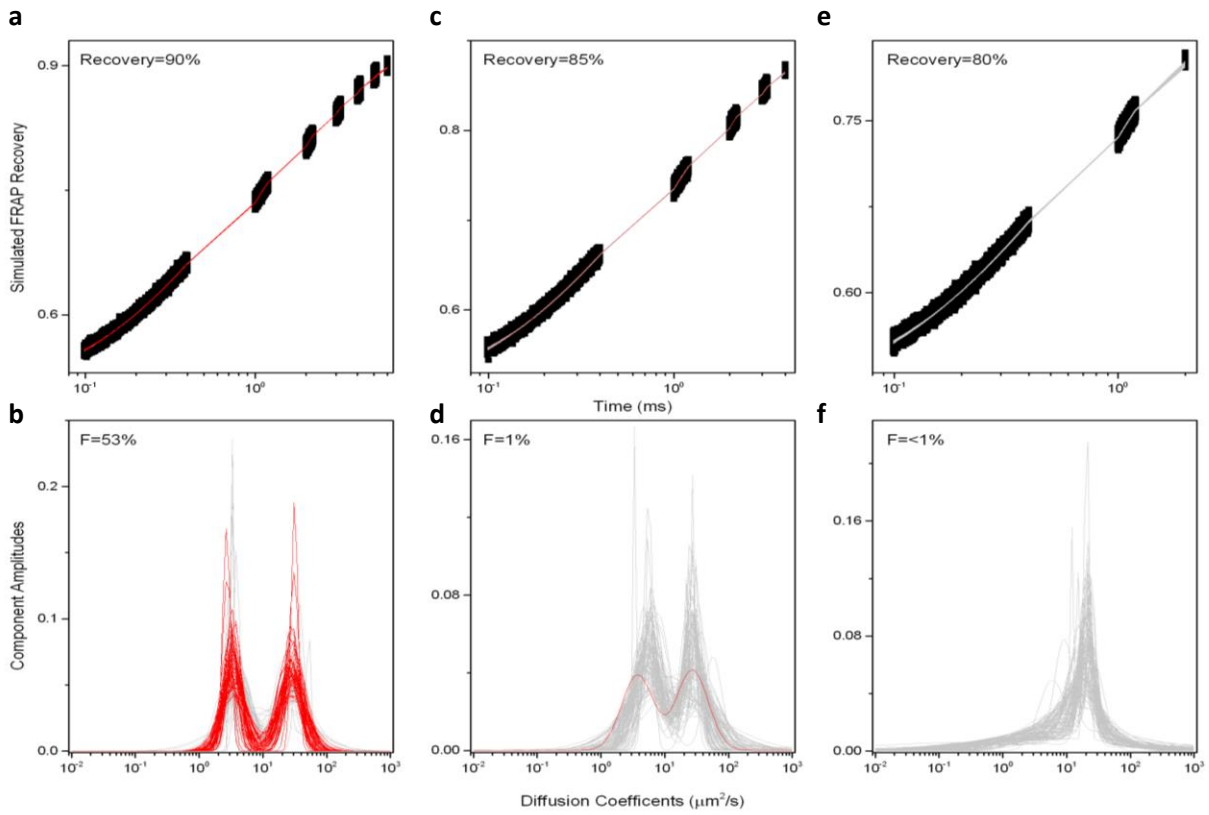

Figure 5.3- Inclusion of an artificial immobile fraction impairs fitting by the distribution model on datasets with 50 dB SNR.(a,c,e) The top panels depict the FRAP curves simulated from the binary mixture with diffusion? components at 30 and 3 $\mu m^2/s$, each at different recovery extent. By truncating the recovery, fewer datapoints were included. Excluding even a modest extent of the recovery (a)(10%) widens the output distributions and impairs datafitting. Missing 15% of the recovery (c) still preserves the overall bimodal structurel, but abolishes any accuracy in the amplitude determination. Once 20% of the recovery is missed (e), the output is unreliable, as indicated by the lack of passing (red) distributions (f) and low diffusion components are present in the output that do not exist in the input distribution.

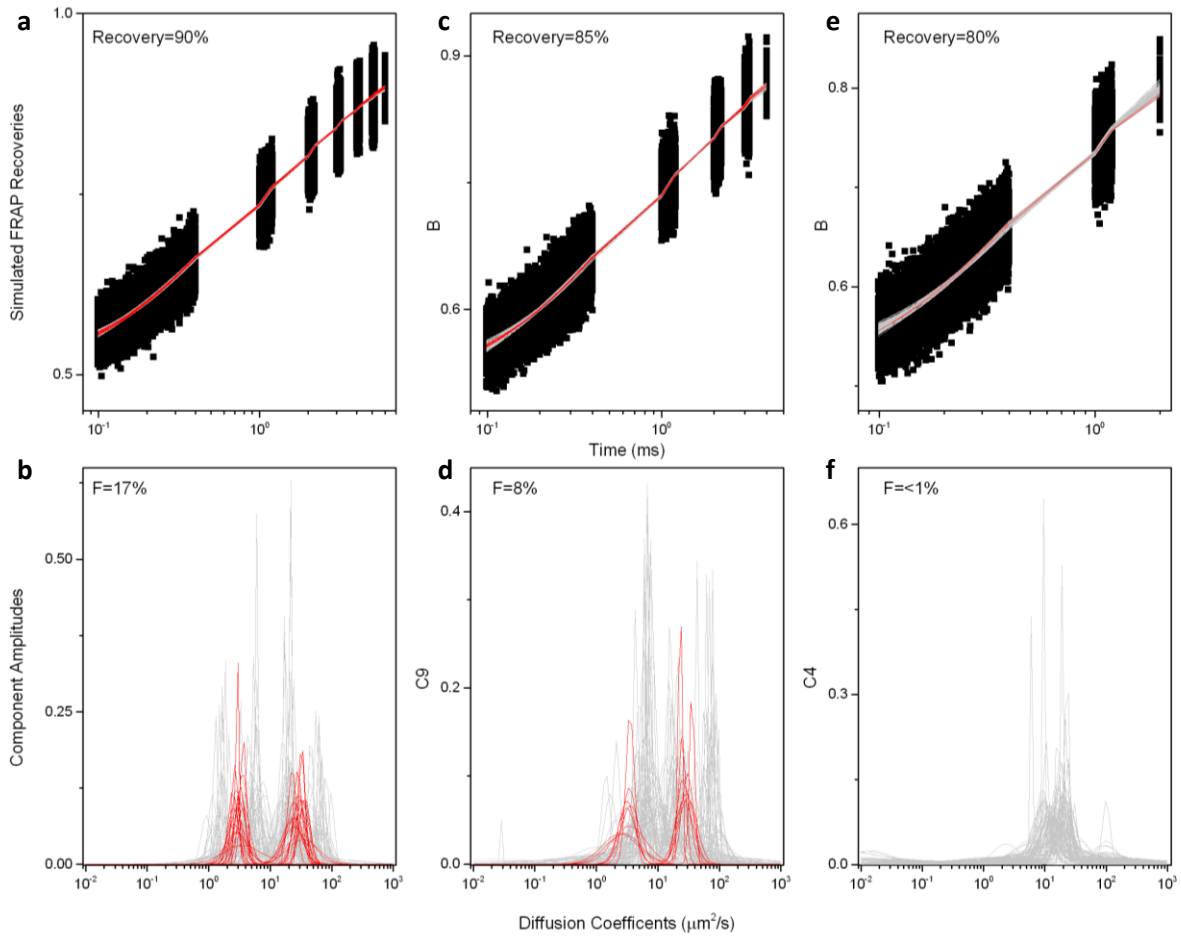

Figure 5.4- Inclusion of an artificial immobile fraction impairs fitting by the distribution model on datasets with 35 dB SNR.(a,c,e) The top panels depict the FRAP curves simulated from the binary mixture with components at 30 and 3 $\mu m^2/s$, each at different recovery extent. **(b)** Excluding even a modest extent of the recovery (10%) (a) destroys the accuracy of the predicted distributions. **(d)** Once 15% of the recovery is missed (c), slow diffusion components are present in the output that do not exist in the input distribution. **(f)**The low (slow?) diffusion components are apparent in the most truncated dataset, and completely distort any biological interpretation.

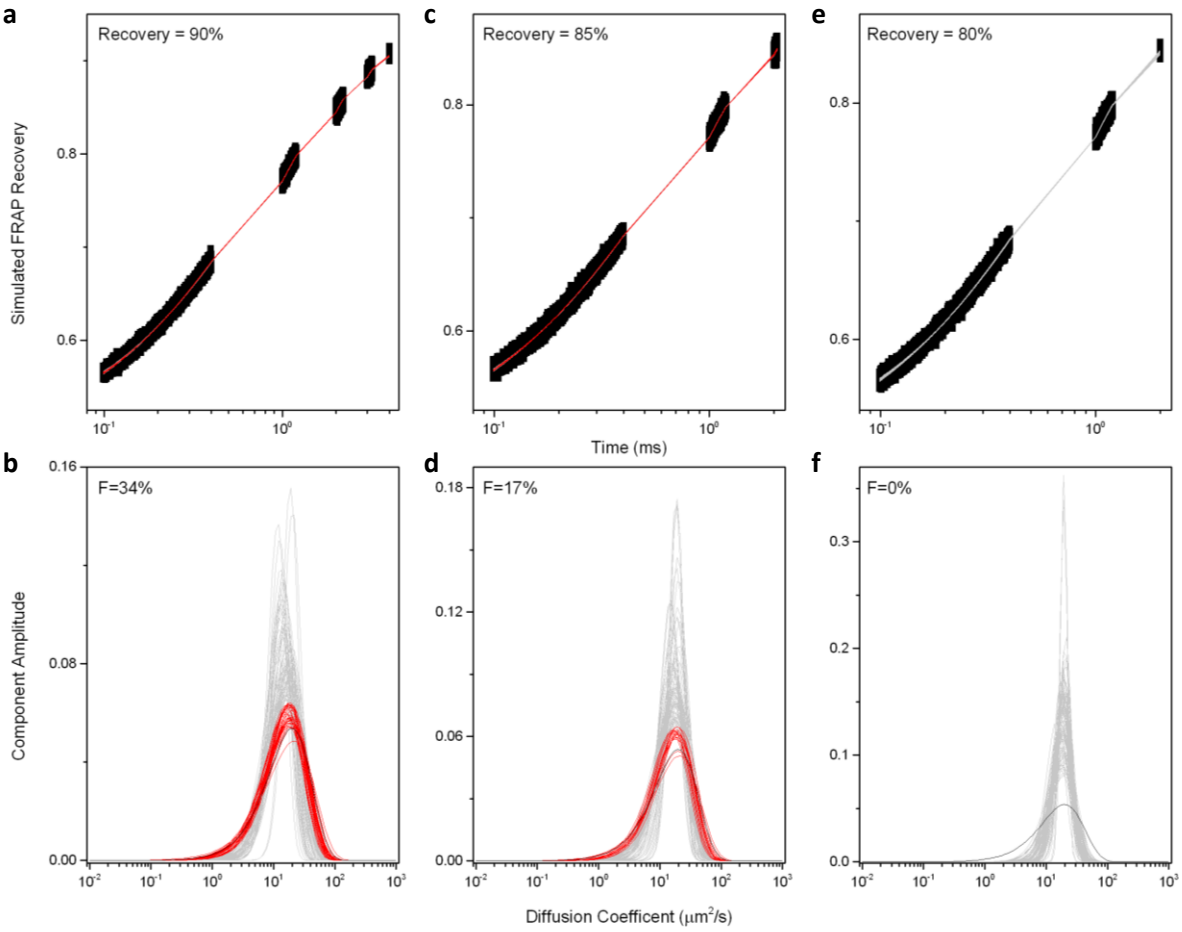

Figure 5.5- Results of extracting the underlying distribution from a gamma function input with the inclusion of an artificial immobile fraction at 50 dB SNR. **(a,c,e)** The top panels depict the FRAP curves simulated from the gamma distribution, each at different recovery extent as explained in 5.1. **(b,d,f)** The predicted outputs always retain the structure of a gamma distribution, but systematically under-estimate the width of the distribution. These outputs could be confused for a single component fit generated from data with a low SNR.

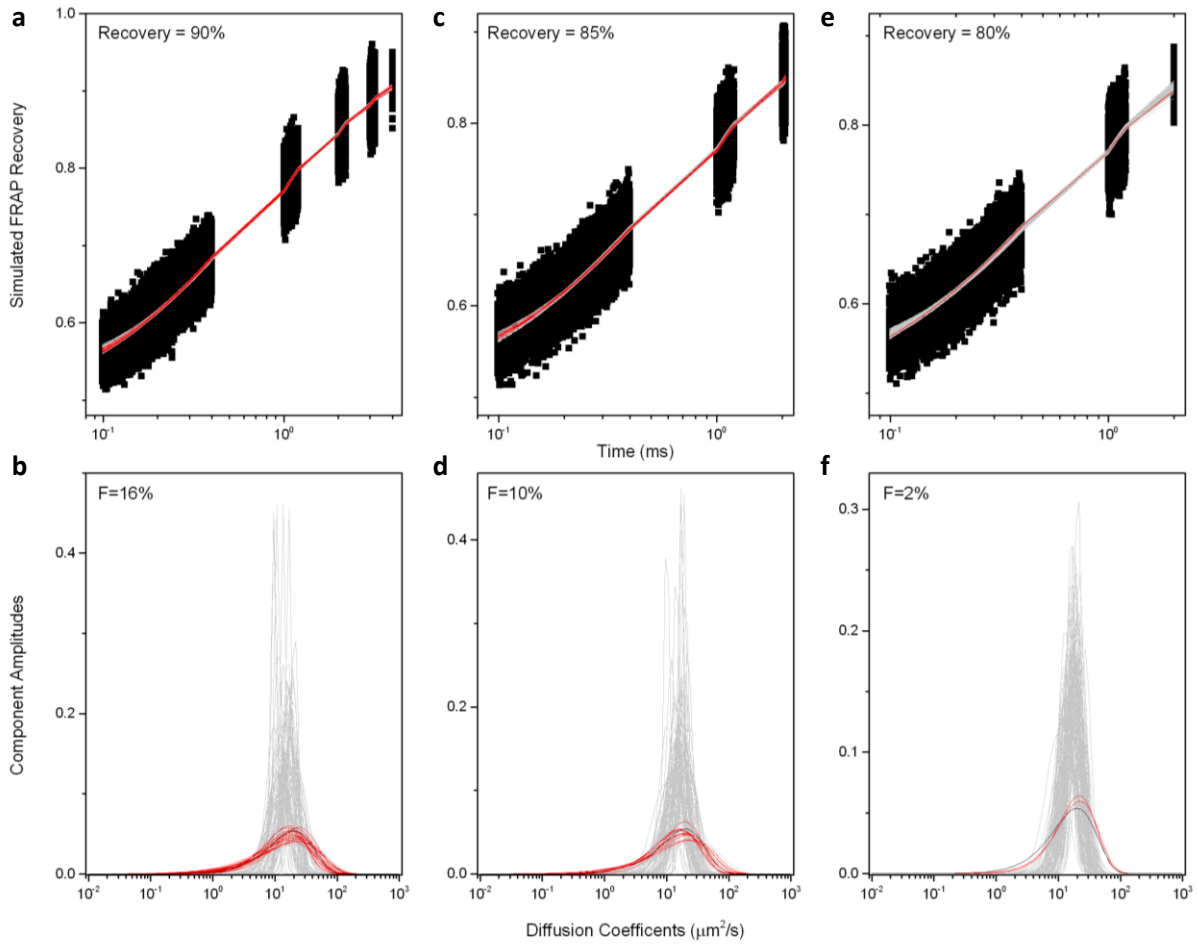

Figure 5.6- Results of extracting the underlying distribution from a gamma function input with the inclusion of an artificial immobile fraction at 35 dB SNR.(a,c,e) The top panels depict the FRAP curves simulated from the gamma distribution, each at different recovery extent.**(b,d,f)** The predicted outputs barely retain the structure of a gamma distribution, and systematically under-estimate the width of the distribution.

assigned to each simulated condition and represent the percent of predicted distributions that pass out of the 100 simulations.

1. Accuracy of Predicting a Binary Mixture

Prior work has established that our distribution model accurately determines the diffusion coefficient of a single component Brownian diffusing species at reasonable experimental SNR (Chapter 4). The next simplest condition is extracting the diffusion coefficients and amplitudes of a binary mixture. It was found that the SNR of the underlying FRAP data has a strong impact on the fidelity of the predicted distribution (Fig. 5.2, bottom panels). A moderate SNR of at least 35 dB is required to accurately predict the binary components 62% of the time, while near noiseless data at a SNR of 50 dB accurately predict the components 90% of the time. Interestingly, the 35 dB FRAP curve always predicts two components whereas the noisy 15 dB data did not, but the accuracy of the diffusion coefficient or relative amplitude is impacted by the noise. In contrast, data with a greater SNR tightens up the width of the predicted distributions. Thus, even at the mid-noise condition of 35 dB, the distribution model is able to provide a useful description of the structure of the underlying distribution.

2. Accuracy of Predicting a Biologically Relevant Distribution

Having demonstrated that data collected at a mid-level SNR can be accurately decomposed into an underlying binary mixture, a more complex, biologically relevant distribution was considered. In this case, a gamma distribution was created that features a sharp decrease in component amplitudes starting at the diffusion coefficient of unconjugated

GFP in a cell nuclei[5]. In a cellular system with a protein tagged by GFP, the fastest possible diffusing species would correspond to a GFP molecule cleaved from the tagged protein. The distribution tails off to slower components, which likely correspond to very large simulated complexes (greater than 1GDa in mass).

Again, the results indicate a strong dependence on distribution accuracy with regards to the SNR of the data (Fig.5.2, bottom panels). The output distributions are systematically too narrow, and the accuracy of determination falls off much more rapidly with decreasing SNR than for the less complex binary mixtures. The 15 dB predicted distributions are not reliable, while the 35 dB predicted distributions reflect the overall shape, but not the amplitudes, of the true distribution. Thus the distributions on mid-noise conditions can give an approximation of the width of the distribution but not the true shapebut, 50 dB is required to provide accurate outputs..

3. Accuracy of Predicting a Binary Mixture with an Artificial Immobile Fraction

The ability to accurately predict the composition of a binary mixture with an artificial immobile fraction was tested at a 35 dB (Fig. 5.4) and 50 dB SNR (Fig. 5.3). To maintain a constant comparison to the previous simulation, in this implementation, the FRAP curve was truncated at increasingly early timepoints to limit the extent of recovery. This is analogous to recording the FRAP recovery of a slow-moving species that does not demonstrate a full recovery on the experimental timescale. As a baseline, the ability to extract information from "noiseless" simulations was investigated (Fig. 5.7). These simulations are constructed from the basis set, but exclude Gaussian noise. As indicated, both 90% and 85% recovery still accurately

represent the structure of the underlying distribution, but accuracy of the predicted values rapidly decreases.

At the highest SNR tested, the ability to extract two species from the underlying distribution remains high, though the accuracy is poor for both input distributions (Fig. 5.3, bottom panels). Both recovery extents indicate two primary species, with the 90% recovery indicating true baseline resolution between the peaks. Thus the structure of the distribution can be trusted, but the true values rapidly become inaccurate. When only 80% of the recovery is captured, the predicted distribution displays only artifacts, and aliases in monotontically decreasing components towards low diffusion coefficients not present in the initial distribution.

At the mid-level SNR simulations, the predicted distributions rapidly lose informational content (Fig. 5.4, bottom panels). The outputs are significantly broadened, and interestingly, when only 80% of the recovery extent is captured, a minor peak can be detected in the majority of the distributions at very slow diffusion coefficients. Again, this indicates that slowly diffusing components are aliased into the distribution, likely to suppress the final value of the predicted FRAP curves.

4. Accuracy of Predicting a Gamma Distribution with an Artificial Immobile Fraction

A similar analysis was then performed for FRAP curves simulated from a distribution defined by a gamma function. Again, a "noiseless" baseline analysis (Fig. 5.8) indicated that the accuracy of the distribution rapidly decreased with decreasing extent of recovery. The width of the predictions contains all the underlying components, but the predicted amplitudes are far from an accurate representation. For both SNR simulations (Fig. 5.5 and Fig. 5.6), the predicted

outputs are not accurate representations of the input distributions. The predictions are systematically more narrow and emphasize mid-range components to a greater extent than the

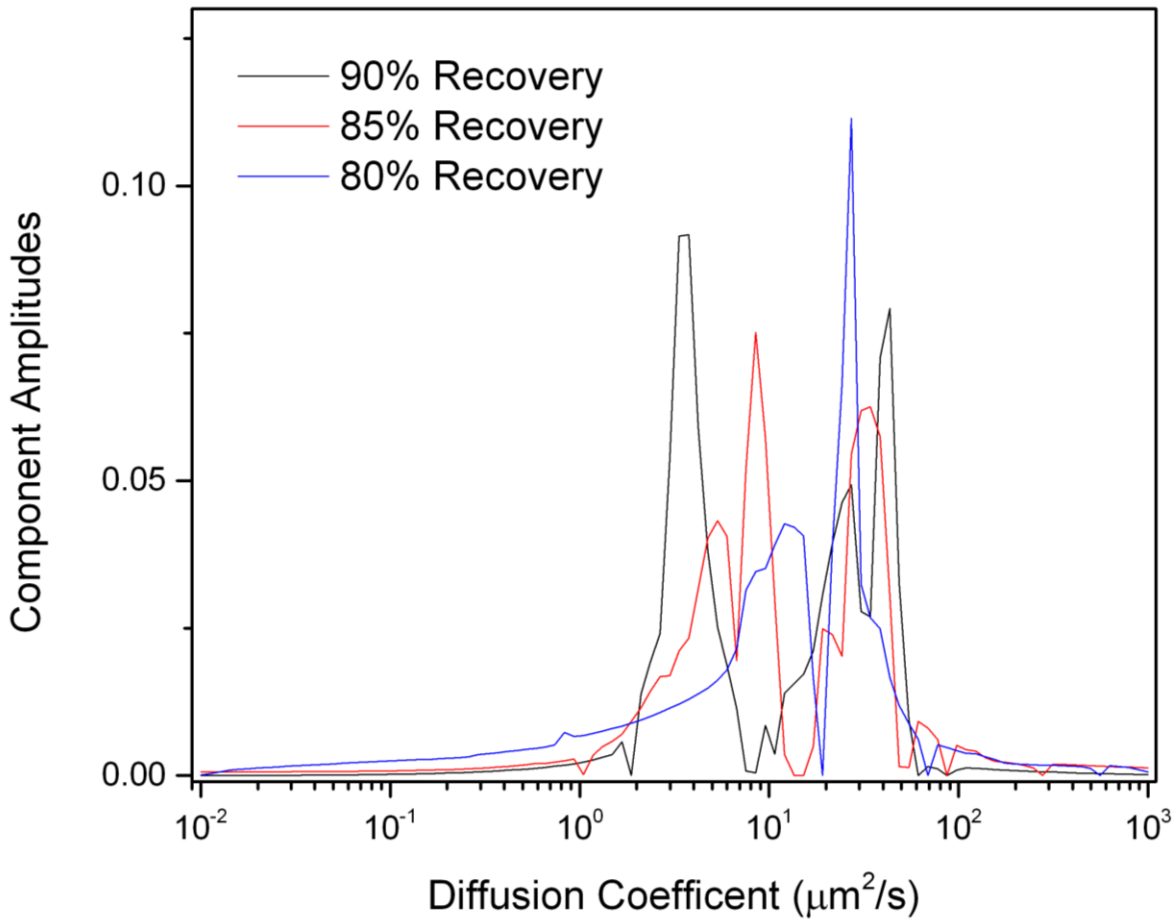

Figure 5.7- Effect of including an artificial immobile fraction on distribution fitting to a binary mixture without noise. Noiseless simulations were fit with the distribution model to determine the extent that missing part of the recovery would have on the output distribution. All the outputs are significantly wider than the true distribution, and false components are rapidly included in the output as the recovery extent decreases.

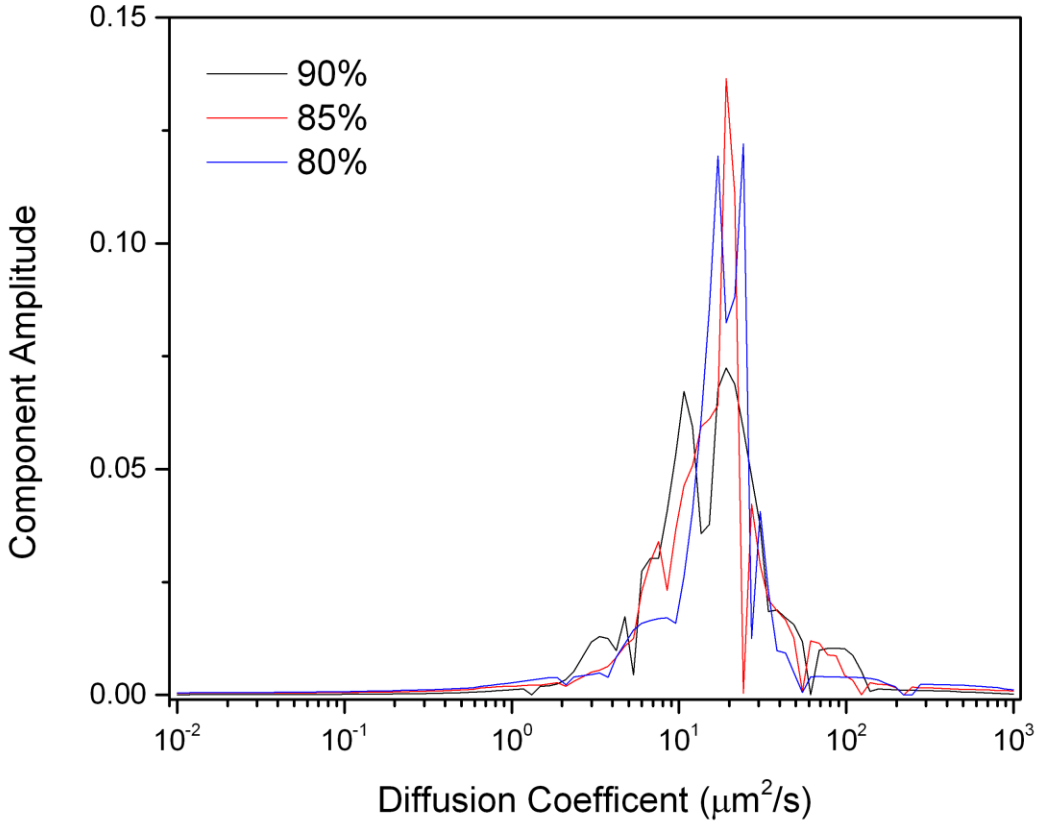

Figure 5.8- Effect of including an artificial immobile fraction on distribution fitting to a gamma distribution without noise. As information content about the recovery extent is lost, the distributions become more peaked and narrow.

true distribution. This indicates that for a distribution of any significant complexity, the entirety of the recovery is needed to accurately extract underlying components.

Conclusions

This analysis was conducted to investigate how robustly the distribution model could handle poor quality or incomplete datasets representative of "real-world" data. It was observed that data quality strongly impacts the accuracy of the model, and 35 dB or better SNR must be maintained for extraction of an underlying distribution. While relatively poor quality data could still encode the structure of the underlying distribution, for information regarding the true envelope of compound, a SNR of 35 dB or better is required. Further, it is essential to capture the full extent of the FRAP recovery. Inclusion of a possible immobile fraction will strongly interfere with determining the accurate distribution, and often aliases incorrect components into the final output.

This work should be strengthen by quantitatively determining how the distributions change in shape with respect to altered input conditions. Trends in features such peak resolution, peak width, and envelope structure can thenbe used to better assign a confidence rating to a predicted distribution. Additionally, the inclusion of non-Brownian components at the slow end of the distribution should be considered, as these likely occur in a cellular environment.

REFERENCES

(1) Axelrod, D.; Koppel, D. E.; Schlessinger, J.; Elson, E.; Webb, W. W. *Biophys. J.* **1976***, 16*, 1055-1069.

(2) Mueller, F.; Mazza, D.; Stasevich, T. J.; McNally, J. G. *Curr. Opin. Cell Biol.* **2010***, 22*, 403-411.

(3) Sprague, B. L.; McNally, J. G. *Trends in Cell Biology* **2005***, 15*.

(4) Sprague, B. L.; Pego, R. L.; Stavreva, D. A.; McNally, J. G. *Biophys. J.* **2004***, 86*, 3473-3495.

(5) Daddysman, M. K.; Fecko, C. J. *J Phys Chem B* **2013***, 117*, 1241-1251.

(6) Dundr, M.; Hoffmann-Rohrer, U.; Hu, Q.; Grummt, I.; Rothblum, L. I.; Phair, R. D.; Misteli, T. *Science* **2002***, 298*, 1623-1626.

(7) Gorski, S. A.; Snyder, S. K.; John, S.; Grummt, I.; Misteli, T. *Mol. Cell* **2008***, 30*, 486-497.

(8) Misteli, T. *Science* **2001***, 291*, 843-847.

(9) Houtsmuller, A. B. **2005***, 95*, 1292-1296.

(10) Periasamy,N., Verkman, A.S *Biophys. J.* **1998***, 75*, 557.

(11) Gordon, G., W.; Chazotte, B.; Wang, X. F.; Herman, B. *Biophys. J.* **1995***, 68*, 766.

(12) Hazewinkel, M., Ed.; In *Encylopedia of Mathematics ;* 2001; .

(13) Kimura, H.; Sugaya, K.; Cook, P. R. *J. Cell Biol.* **2002***, 159*, 777-782.

(14) Feder, T. J.; Burst-Mascher, I.; Slattery, J. P.; Baird, B.; Webb, W. W. *Biophys. J.* **1996***, 70*, 2767-2773.

(15) Kang, M.; DiBenedetto, E.; Kenworthy, A. *Biophys. J.* **2011***, 100*, 791-792.

APPENDIX A

QUANTIFICATION OF GEL ELECTROPHORESIS DATA USING FOUR GAUSSIAN PEAKS TO OVERCOME BACKGROUND HETEROGENEITIES

This section will detail the MATLAB code written to analyze the gel electrophoresis data presented in Chapter 3. After irradiating stained plasmid samples, gel electrophoresis was used to separate out the three different forms of DNA- undamaged supercoiled plasmid, nicked plasmid, and linearized plasmid. A minor fraction of multiple fragments was not usually considered. The relative amount of each species was quantified ratiometrically from the signal intensity of saved images of each gel, and a correction factor applied to account for the differential staining affinity of the different plasmid states. While a simplistic gel analysis can be performed using ImageJ, to obtain the best quantifications possible, three common complications were addressed with the following scripts. First, gel images often had non-uniform background intensities, which confounded the determination of each species within a gel lane. Second, the supercoiled and linear species often exhibited poor separation. Third, for closely resolved species, determining the lateral extent of the band intensity was often subjective. In response, each lane was fit as the sum of four Gaussian functions. Three Gaussian peaks corresponded to the three plasmid species, and the fourth was used to account for the background signal and enabled a correction to be applied, even if the background intensity was not uniform. To analyze each gel, three programs were written: gel_load2, lane_analysis2, and gel_analysis2b_MT

The analysis began by loading the gel image using *gel_load2*:

```
%loads gel and plots the intensity of each lane.
```

```
%change name and the lane positions (xstart01, ystart01, etc).

clear all

%file name (wihout .tif)
name='nc20 yo frame average';

%lane positions
xstart01=210;
xstart02=450;
xsize=180;

ystart01=433;
ysize01=7;
ydelta01=-19.2;

ystart02=430;
ysize02=7;
ydelta02=-19.2;

%nothing should need to be modified below this point

gel=imread([name '.tif']);

%first set of lanes
ystart1=ystart01;
ystart2=ystart01+round(ydelta01);
ystart3=ystart01+round(2*ydelta01);
ystart4=ystart01+round(3*ydelta01);
ystart5=ystart01+round(4*ydelta01);
ystart6=ystart01+round(5*ydelta01);
ystart7=ystart01+round(6*ydelta01);
ystart8=ystart01+round(7*ydelta01);
ystart9=ystart01+round(8*ydelta01);
ystart10=ystart01+round(9*ydelta01);
ystart11=ystart01+round(10*ydelta01);
ystart12=ystart01+round(11*ydelta01);
ystart13=ystart01+round(12*ydelta01);
ystart14=ystart01+round(13*ydelta01);
ystart15=ystart01+round(14*ydelta01);
ystart16=ystart01+round(15*ydelta01);
ystart17=ystart01+round(16*ydelta01);
ystart18=ystart01+round(17*ydelta01);
ystart19=ystart01+round(18*ydelta01);
ystart20=ystart01+round(19*ydelta01);
%second set of lanes
ystart21=ystart02+round(0*ydelta02);
ystart22=ystart02+round(1*ydelta02);
ystart23=ystart02+round(2*ydelta02);
ystart24=ystart02+round(3*ydelta02);
ystart25=ystart02+round(4*ydelta02);
ystart26=ystart02+round(5*ydelta02);
ystart27=ystart02+round(6*ydelta02);
ystart28=ystart02+round(7*ydelta02);
ystart29=ystart02+round(8*ydelta02);
```

```
ystart30=ystart02+round(9*ydelta02);
ystart31=ystart02+round(10*ydelta02);
ystart32=ystart02+round(11*ydelta02);
ystart33=ystart02+round(12*ydelta02);
ystart34=ystart02+round(13*ydelta02);
ystart35=ystart02+round(14*ydelta02);
ystart36=ystart02+round(15*ydelta02);
ystart37=ystart02+round(16*ydelta02);
ystart38=ystart02+round(17*ydelta02);
ystart39=ystart02+round(18*ydelta02);
ystart40=ystart02+round(19*ydelta02);

C=gel(:,:,:);
high=2^16-1;

%box1
C(ystart1-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart1+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart1-1:ystart1+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart1-1:ystart1+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box1=sum(gel(ystart1:ystart1+ysize01,xstart01:xstart01+xsize),1);

%box2
C(ystart2-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart2+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart2-1:ystart2+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart2-1:ystart2+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box2=sum(gel(ystart2:ystart2+ysize01,xstart01:xstart01+xsize),1);

%box3
C(ystart3-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart3+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart3-1:ystart3+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart3-1:ystart3+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box3=sum(gel(ystart3:ystart3+ysize01,xstart01:xstart01+xsize),1);

%box4
C(ystart4-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart4+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart4-1:ystart4+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart4-1:ystart4+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box4=sum(gel(ystart4:ystart4+ysize01,xstart01:xstart01+xsize),1);

%box5
C(ystart5-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart5+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart5-1:ystart5+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart5-1:ystart5+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box5=sum(gel(ystart5:ystart5+ysize01,xstart01:xstart01+xsize),1);

%box6
C(ystart6-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart6+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart6-1:ystart6+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart6-1:ystart6+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
```

```
box6=sum(gel(ystart6:ystart6+ysize01,xstart01:xstart01+xsize),1);

%box7
C(ystart7-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart7+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart7-1:ystart7+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart7-1:ystart7+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box7=sum(gel(ystart7:ystart7+ysize01,xstart01:xstart01+xsize),1);

%box8
C(ystart8-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart8+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart8-1:ystart8+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart8-1:ystart8+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box8=sum(gel(ystart8:ystart8+ysize01,xstart01:xstart01+xsize),1);

%box9
C(ystart9-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart9+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart9-1:ystart9+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart9-1:ystart9+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box9=sum(gel(ystart9:ystart9+ysize01,xstart01:xstart01+xsize),1);

%box10
C(ystart10-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart10+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart10-1:ystart10+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart10-1:ystart10+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box10=sum(gel(ystart10:ystart10+ysize01,xstart01:xstart01+xsize),1);

%box11
C(ystart11-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart11+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart11-1:ystart11+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart11-1:ystart11+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box11=sum(gel(ystart11:ystart11+ysize01,xstart01:xstart01+xsize),1);

%box12
C(ystart12-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart12+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart12-1:ystart12+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart12-1:ystart12+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box12=sum(gel(ystart12:ystart12+ysize01,xstart01:xstart01+xsize),1);

%box13
C(ystart13-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart13+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart13-1:ystart13+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart13-1:ystart13+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box13=sum(gel(ystart13:ystart13+ysize01,xstart01:xstart01+xsize),1);

%box14
C(ystart14-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart14+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart14-1:ystart14+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
```

```
C(ystart14-1:ystart14+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box14=sum(gel(ystart14:ystart14+ysize01,xstart01:xstart01+xsize),1);

%box15
C(ystart15-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart15+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart15-1:ystart15+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart15-1:ystart15+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box15=sum(gel(ystart15:ystart15+ysize01,xstart01:xstart01+xsize),1);

%box16
C(ystart16-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart16+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart16-1:ystart16+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart16-1:ystart16+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box16=sum(gel(ystart16:ystart16+ysize01,xstart01:xstart01+xsize),1);

%box17
C(ystart17-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart17+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart17-1:ystart17+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart17-1:ystart17+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box17=sum(gel(ystart17:ystart17+ysize01,xstart01:xstart01+xsize),1);

%box18
C(ystart18-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart18+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart18-1:ystart18+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart18-1:ystart18+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box18=sum(gel(ystart18:ystart18+ysize01,xstart01:xstart01+xsize),1);

%box19
C(ystart19-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart19+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart19-1:ystart19+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart19-1:ystart19+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box19=sum(gel(ystart19:ystart19+ysize01,xstart01:xstart01+xsize),1);

%box20
C(ystart20-1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart20+ysize01+1,xstart01-1:xstart01+xsize+1)=high*ones(1,xsize+3);
C(ystart20-1:ystart20+ysize01+1,xstart01-1)=high*ones(ysize01+3,1);
C(ystart20-1:ystart20+ysize01+1,xstart01+xsize+1)=high*ones(ysize01+3,1);
box20=sum(gel(ystart20:ystart20+ysize01,xstart01:xstart01+xsize),1);

%box21
C(ystart21-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart21+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart21-1:ystart21+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart21-1:ystart21+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box21=sum(gel(ystart21:ystart21+ysize02,xstart02:xstart02+xsize),1);

%box22
C(ystart22-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart22+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
```

```
C(ystart22-1:ystart22+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart22-1:ystart22+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box22=sum(gel(ystart22:ystart22+ysize02,xstart02:xstart02+xsize),1);

%box23
C(ystart23-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart23+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart23-1:ystart23+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart23-1:ystart23+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box23=sum(gel(ystart23:ystart23+ysize02,xstart02:xstart02+xsize),1);

%box24
C(ystart24-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart24+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart24-1:ystart24+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart24-1:ystart24+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box24=sum(gel(ystart24:ystart24+ysize02,xstart02:xstart02+xsize),1);

%box25
C(ystart25-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart25+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart25-1:ystart25+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart25-1:ystart25+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box25=sum(gel(ystart25:ystart25+ysize02,xstart02:xstart02+xsize),1);

%box26
C(ystart26-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart26+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart26-1:ystart26+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart26-1:ystart26+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box26=sum(gel(ystart26:ystart26+ysize02,xstart02:xstart02+xsize),1);

%box27
C(ystart27-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart27+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart27-1:ystart27+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart27-1:ystart27+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box27=sum(gel(ystart27:ystart27+ysize02,xstart02:xstart02+xsize),1);

%box28
C(ystart28-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart28+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart28-1:ystart28+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart28-1:ystart28+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box28=sum(gel(ystart28:ystart28+ysize02,xstart02:xstart02+xsize),1);

%box29
C(ystart29-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart29+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart29-1:ystart29+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart29-1:ystart29+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box29=sum(gel(ystart29:ystart29+ysize02,xstart02:xstart02+xsize),1);

%box30
C(ystart30-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
```

```
C(ystart30+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart30-1:ystart30+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart30-1:ystart30+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box30=sum(gel(ystart30:ystart30+ysize02,xstart02:xstart02+xsize),1);

%box31
C(ystart31-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart31+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart31-1:ystart31+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart31-1:ystart31+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box31=sum(gel(ystart31:ystart31+ysize02,xstart02:xstart02+xsize),1);

%box32
C(ystart32-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart32+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart32-1:ystart32+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart32-1:ystart32+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box32=sum(gel(ystart32:ystart32+ysize02,xstart02:xstart02+xsize),1);

%box33
C(ystart33-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart33+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart33-1:ystart33+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart33-1:ystart33+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box33=sum(gel(ystart33:ystart33+ysize02,xstart02:xstart02+xsize),1);

%box34
C(ystart34-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart34+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart34-1:ystart34+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart34-1:ystart34+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box34=sum(gel(ystart34:ystart34+ysize02,xstart02:xstart02+xsize),1);

%box35
C(ystart35-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart35+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart35-1:ystart35+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart35-1:ystart35+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box35=sum(gel(ystart35:ystart35+ysize02,xstart02:xstart02+xsize),1);

%box36
C(ystart36-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart36+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart36-1:ystart36+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart36-1:ystart36+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box36=sum(gel(ystart36:ystart36+ysize02,xstart02:xstart02+xsize),1);
%box37
C(ystart37-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart37+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart37-1:ystart37+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart37-1:ystart37+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box37=sum(gel(ystart37:ystart37+ysize02,xstart02:xstart02+xsize),1);

%box38
C(ystart38-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
```

```
C(ystart38+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart38-1:ystart38+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart38-1:ystart38+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box38=sum(gel(ystart38:ystart38+ysize02,xstart02:xstart02+xsize),1);

%box39
C(ystart39-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart39+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart39-1:ystart39+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart39-1:ystart39+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box39=sum(gel(ystart39:ystart39+ysize02,xstart02:xstart02+xsize),1);

%box40
C(ystart40-1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart40+ysize02+1,xstart02-1:xstart02+xsize+1)=high*ones(1,xsize+3);
C(ystart40-1:ystart40+ysize02+1,xstart02-1)=high*ones(ysize02+3,1);
C(ystart40-1:ystart40+ysize02+1,xstart02+xsize+1)=high*ones(ysize02+3,1);
box40=sum(gel(ystart40:ystart40+ysize02,xstart02:xstart02+xsize),1);

figure(1)
imshow(high/max(max(double(gel)))*C)

nolanes=40;
pix=1:(xsize+1);
options=optimset('MaxFunEvals',1e4,'MaxIter',1e4);

figure(2), clf
figure(3), clf

for j=1:nolanes
    eval(['data=box' int2str(j) ';'])
    figure(2)
    subplot(5,8,j)
    plot(pix,data)
    axis tight
    title(num2str(j))
end

clear jxstart*xsize*ystart*ysize*ydelta*
save(name)
```

The program assumes a 40-lane gel (two rows of 20 lanes), and applies a mask of 40 pre-spaced lanes to the entire gel (data was run in triplicate). User input is required to help align the mask. The signal from each lane is the summation of image intensity in the y-axis for the length of the lane (thus the intensity of each lane is recorded as a single line).

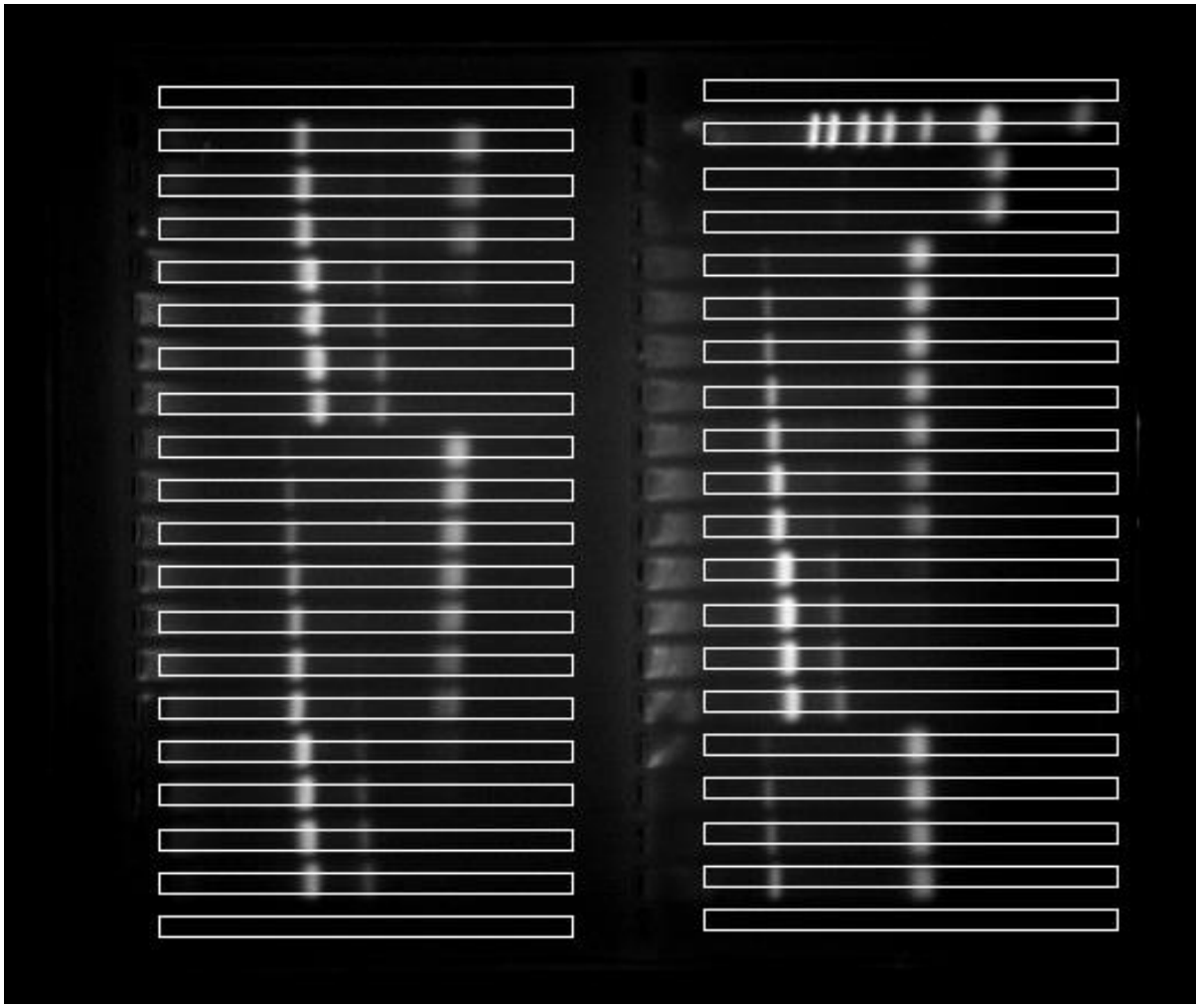

Figure A.1- Output of *gel_analysis2*, indicating the region of interest for each lane.Two sets of 20 boxes are superimposed over the gel image. The user modifies the initial starting position of each bank of boxes in both the x and y direction to best center the boxes over each lane. It is important to try and avoid as much of the low mobility smearing near each well as possible, lest that signal get confused for a broad peak. Once the boxes are positioned over each lane, the intensity is summed in the y-axis, and the intensity across the entire lane represented by a line.

Following the determination of the intensity in each lane, the *lane_analysis2* script is run:

```
%fits the intensity profile of a given lane, defined by "j", to the sum of
%four Gaussian functions (supercoiled, relaxed, linear and background).

%change the file name, j and center positions; it may sometimes be
%necessary to change the peaksize, fitsize and maxpeakwidth

clear all

%file name (wihout .tif)
name='nc20 yo frame average';

j=35;

center1=94;
center2=28;
center3=60;
center4=130;

peaksize=16; %defines the distance over wich data is "fit" on either side of
the center value
fitsize=4; %defines the maximum variation of the Gaussian center from its
initial value
maxpeakwidth=18; %defines the upper limit on the Gaussian function width

%nothing should need to be modified below this point

load(name)
fid=fopen([name ' fits.mat']);
if fid>-1
    load([name ' fits.mat'])
end

eval(['data=box' int2str(j) ';'])
offset=mean(data(1:10));
offsetEND=mean(data((end-10):end));
data2=colfilt(data,[1,30],'sliding',@median);
data2(end)=data2(end-1);
x0=[offset 0 (offsetEND-offset) 60 5];
lb=[-1 -1e4 0 0 0]; ub=[2*offset+1 1e4 1e5 180 10];
x=lsqcurvefit('baseline2',x0,pix,data2,lb,ub);
fit=baseline2(x,pix);
data3=data-fit;

cent=center1;
xx0=[(max(data3)-offsetEND) cent 4 offsetEND];
lbb=[0 cent-fitsize 1]; ubb=[2*max(data3) cent+fitsize maxpeakwidth];
xx=lsqcurvefit('gauss_off2',xx0,pix,data3,lbb,ubb,options);
data4=data((cent-peaksize):(cent+peaksize));
```

```
pix4=pix((cent-peaksize):(cent+peaksize));
xx0=[xx(1) cent 4 offset 0];
x1=lsqcurvefit('gauss_off_baseline2',xx0,pix4,data4,lbb,ubb,options);
fit1=gauss_off_baseline2(x1,pix);

cent=center2;
xx0=[(max(data3)-offsetEND) cent 4 offsetEND];
lbb=[0 cent-fitsize 1]; ubb=[2*max(data3) cent+fitsize maxpeakwidth];
xx=lsqcurvefit('gauss_off2',xx0,pix,data3,lbb,ubb,options);data4=data((cent-
peaksize):(cent+peaksize));
pix4=pix((cent-peaksize):(cent+peaksize));
xx0=[xx(1) cent 4 offset 0];
x2=lsqcurvefit('gauss_off_baseline2',xx0,pix4,data4,lbb,ubb,options);
fit2=gauss_off_baseline2(x2,pix);

cent=center3;
xx0=[(max(data3)-offsetEND) cent 4 offsetEND];
lbb=[0 cent-fitsize 1]; ubb=[2*max(data3) cent+fitsize maxpeakwidth];
xx=lsqcurvefit('gauss_off2',xx0,pix,data3,lbb,ubb,options);data4=data((cent-
peaksize):(cent+peaksize));
pix4=pix((cent-peaksize):(cent+peaksize));
xx0=[xx(1) cent 4 offset 0];
x3=lsqcurvefit('gauss_off_baseline2',xx0,pix4,data4,lbb,ubb,options);
cent3=cent;
fit3=gauss_off_baseline2(x3,pix);

cent=center4;
xx0=[(max(data3)-offsetEND) cent 4 offsetEND];
lbb=[0 cent-fitsize 1]; ubb=[2*max(data3) cent+fitsize maxpeakwidth];
xx=lsqcurvefit('gauss_off2',xx0,pix,data3,lbb,ubb,options);data4=data((cent-
peaksize):(cent+peaksize));
pix4=pix((cent-peaksize):(cent+peaksize));
xx0=[xx(1) cent 4 offset 0];
x4=lsqcurvefit('gauss_off_baseline2',xx0,pix4,data4,lbb,ubb,options);
fit4=gauss_off_baseline2(x4,pix);

figure(13)
plot(pix,data,pix,fit1,pix,fit2,pix,fit3,pix,fit4)
axis tight
ylim([min(data) max(data)])
title(num2str(j))

area1(j)=x1(1).*x1(3);
area2(j)=x2(1).*x2(3);
area3(j)=x3(1).*x3(3);
area4(j)=x4(1).*x4(3);
fitmat1(j,:)=fit1;
fitmat2(j,:)=fit2;
fitmat3(j,:)=fit3;
fitmat4(j,:)=fit4;
centervec1(j)=center1;
centervec2(j)=center2;
centervec3(j)=center3;
centervec4(j)=center4;
```

```
save([name '
fits'],'area1','area2','area3','area4','fitmat1','fitmat2','fitmat3','fitmat4
','centervec1','centervec2','centervec3','centervec4')
```

User input is required to provide estimated starting positions for the centers of the three band intensities- center1, center2, center3 and center4. Further, the *peaksize, fitsize,* and *maxpeakwidth* need to be altered by the user to obtain the best fit possible for each lane. These parameters often remain stable for each side of the gel. The quality of the Gaussian fitting is checked by eye, and the fit parameters adjusted until the peak-fits are optimized. This program is run for each lane that requires analysis.

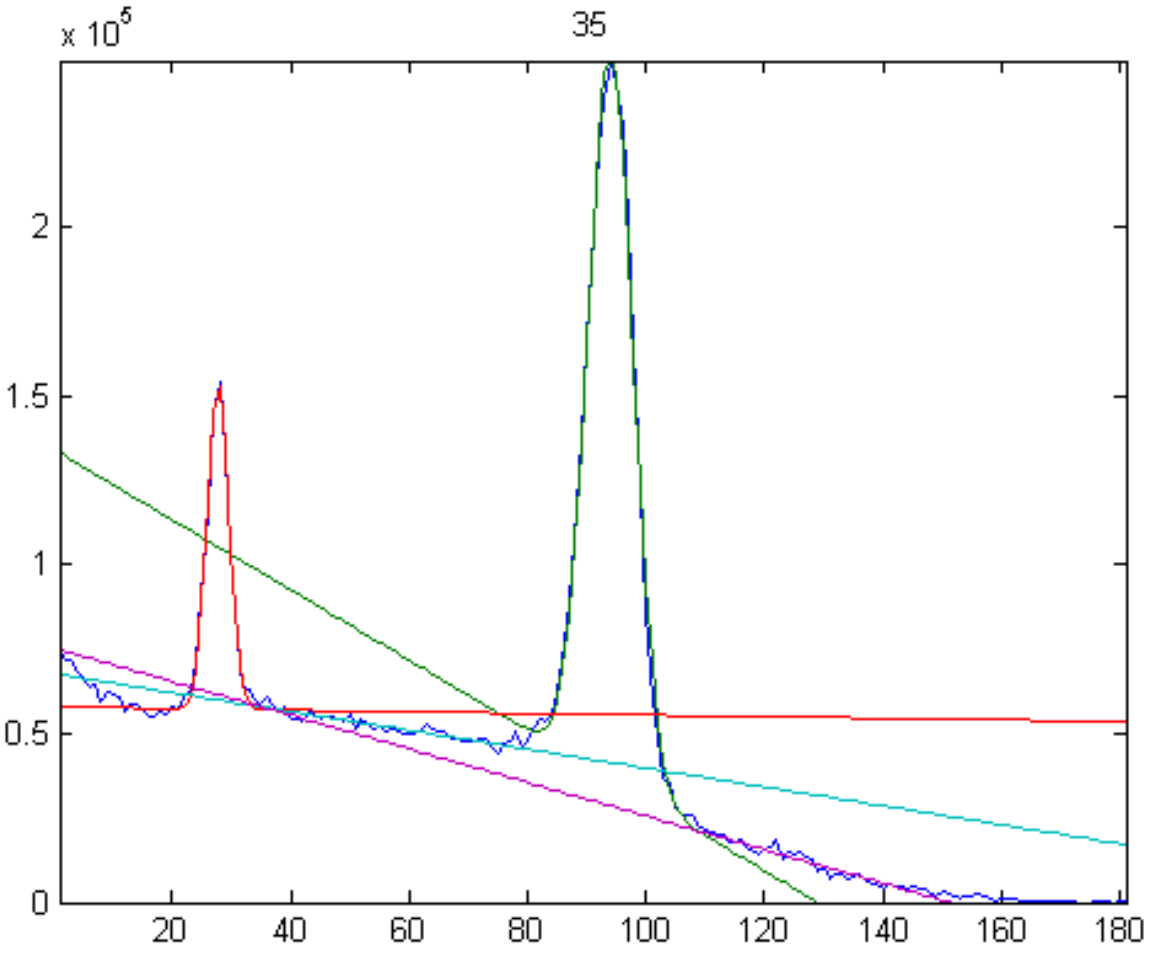

Figure A.2- Representative output of a single lane analysis. The lane intensity quantified from the gel is shown in blue. The green peak is a fit to the supercoiled species, the red peak a fit to the nicked species, and the magenta slope a fit to the baseline. As can be seen, the baseline of the quantification is not uniform across the entire lane. Further, only two peaks are present, therefore no linear band was detected. Accordingly, the cyan curve conforms to the baseline region were the linear species would be located. The fit parameters are adjusted until the fits are visually optimized; the procedure is then repeated for each lane.

Finally, the script *gel_analysis2b_MT* is used to compile the quantifications of the band intensities and perform the background subtractions. Typically, each gel had samples run in triplicate, thus three lanes are averaged together for the final result. If a lane failed or was very poorly quantified, the user must manually exclude it from the analysis. The correction factor for differential dye staining affinity is performed during the data fitting. No user input is required other than specifying the filename.

```
%processeses and plots all of the data from the lane_analysis programs (run
%this after running lane_analysis for all lanes).

%6/23/11 modified to include data without baseline background subtraction

%only need to change the name

clear all

%file name (wihout .tif)
name='nc20 yo frame average';

%nothing should need to be modified below this point

load(name)
load([name ' fits.mat'])

len=size(fitmat1,2);
fitmat1(41,:)=zeros(1,len);
fitmat2(41,:)=zeros(1,len);
fitmat3(41,:)=zeros(1,len);
fitmat4(41,:)=zeros(1,len);

figure(3), clf

for j=1:nolanes

    eval(['data=box' int2str(j) ';'])
    fit1=fitmat1(j,:);
    fit2=fitmat2(j,:);
    fit3=fitmat3(j,:);
    fit4=fitmat4(j,:);

    figure(3)
    subplot(5,8,j)
    plot(pix,data,pix,fit1,pix,fit2,pix,fit3,pix,fit4)
    axis tight
    ylim([min(data) max(data)])
```

```
    title(num2str(j))
end

%baseline background subtracted
p1=(area1-area4)./(area1+area2+area3-3*area4);
p2=(area2-area4)./(area1+area2+area3-3*area4);
p3=(area3-area4)./(area1+area2+area3-3*area4);
pmat=[p1;p2;p3];
%t0
d1=12;
d2=25;
d3=36;
av(1:3,1)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,1)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t0.5
d1=11;
d2=24;
d3=35;
av(1:3,2)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,2)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t1
d1=10;
d2=23;
d3=34;
av(1:3,3)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,3)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t2
d1=9;
d2=22;
d3=33;
av(1:3,4)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,4)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t3
d1=8;
d2=19;
d3=32;
av(1:3,5)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,5)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t4
d1=7;
d2=18;
d3=31;
av(1:3,6)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,6)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t5
d1=6;
d2=17;
d3=30;
av(1:3,7)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,7)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t10
d1=5;
d2=16;
d3=29;
av(1:3,8)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,8)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
```

```
%t15
d1=4;
d2=15;
d3=28;
av(1:3,9)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,9)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t20
d1=3;
d2=14;
d3=27;
av(1:3,10)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,10)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t30
d1=2;
d2=13;
d3=26;
av(1:3,11)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,11)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);

av1=av(1,:); av1b=av1;
av2=av(2,:); av2b=av2;
av3=av(3,:); av3b=av3;
sd1=sd(1,:); sd1b=sd1;
sd2=sd(2,:); sd2b=sd2;
sd3=sd(3,:); sd3b=sd3;

figure(4)
subplot(2,1,1)
time=[0 0.5 1 2 3 4 5 10 15 20 30];
plot(time,av1,time,av2,time,av3)
axis tight
ylim([-0.1 1.1])
title([name ' (background subtracted)'])

%baseline background NOT subtracted
p1=(area1)./(area1+area2+area3);
p2=(area2)./(area1+area2+area3);
p3=(area3)./(area1+area2+area3);
pmat=[p1;p2;p3];
%t0
d1=12;
d2=25;
d3=36;
av(1:3,1)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,1)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t0.5
d1=11;
d2=24;
d3=35;
av(1:3,2)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,2)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t1
d1=10;
d2=23;
d3=34;
av(1:3,3)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
```

```
sd(1:3,3)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t2
d1=9;
d2=22;
d3=33;
av(1:3,4)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,4)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t3
d1=8;
d2=19;
d3=32;
av(1:3,5)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,5)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t4
d1=7;
d2=18;
d3=31;
av(1:3,6)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,6)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t5
d1=6;
d2=17;
d3=30;
av(1:3,7)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,7)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t10
d1=5;
d2=16;
d3=29;
av(1:3,8)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,8)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t15
d1=4;
d2=15;
d3=28;
av(1:3,9)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,9)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t20
d1=3;
d2=14;
d3=27;
av(1:3,10)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,10)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);
%t30
d1=2;
d2=13;
d3=26;
av(1:3,11)=mean([pmat(:,d1) pmat(:,d2) pmat(:,d3)],2);
sd(1:3,11)=std([pmat(:,d1) pmat(:,d2) pmat(:,d3)],[],2);

av1=av(1,:);
av2=av(2,:);
av3=av(3,:);
sd1=sd(1,:);
sd2=sd(2,:);
sd3=sd(3,:);
```

```
figure(4)
subplot(2,1,2)
time=[0 0.5 1 2 3 4 5 10 15 20 30];
plot(time,av1,time,av2,time,av3)
axis tight
ylim([-0.1 1.1])
title([name ' (background not subtracted)'])
matout=zeros(11,15);
matout(:,1)=time;
matout(:,3)=av1;
matout(:,4)=sd1;
matout(:,5)=av2;
matout(:,6)=sd2;
matout(:,7)=av3;
matout(:,8)=sd3;
matout(:,10)=av1b;
matout(:,11)=sd1b;
matout(:,12)=av2b;
matout(:,13)=sd2b;
matout(:,14)=av3b;
matout(:,15)=sd3b;

save([name '.dat'], 'matout', '-ASCII')
```

The script outputs a visual confirmation of the fit for each lane, and well as the averaged quantification of each plasmid species.

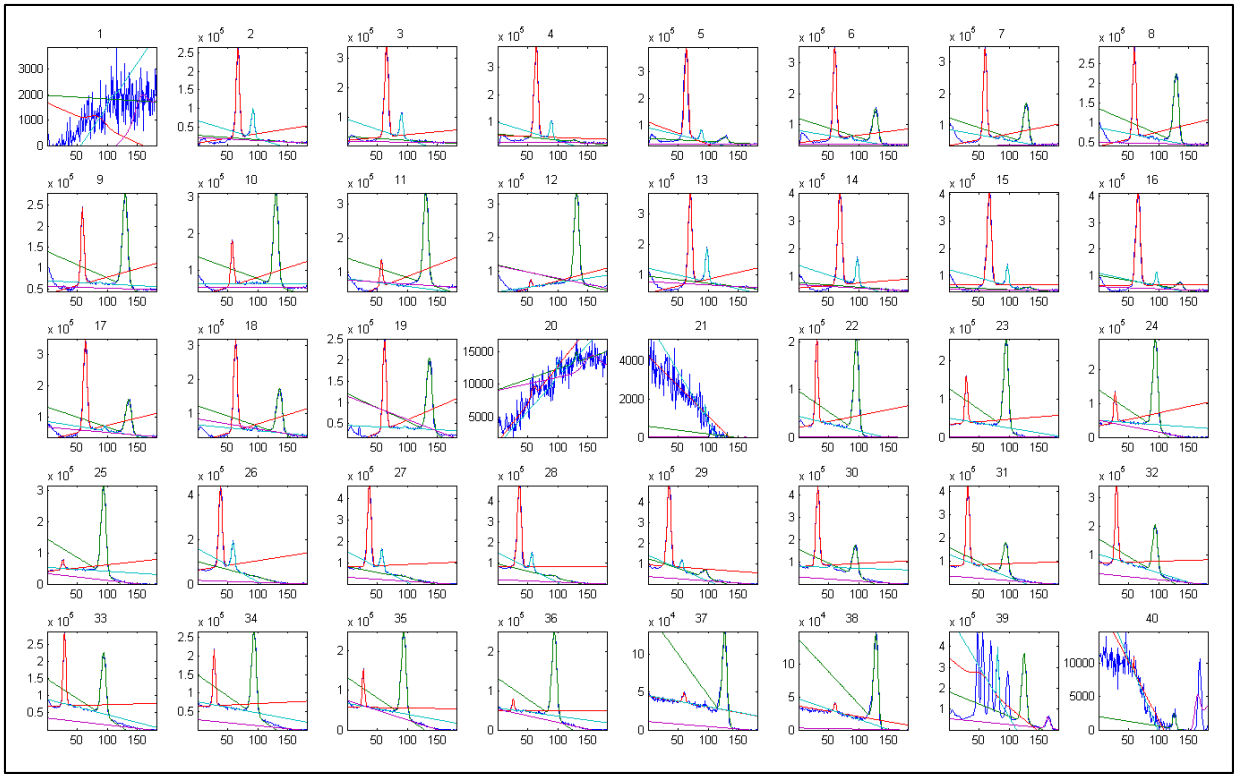

Figure A.3- Output of the Gaussian fits to each DNA band for every lane. This enables a rapid confirmation that each lane was properly fit. Aberrant fits or lanes that should be discarded are identified at this step and manually excluded from further analysis. Lanes 1,20,21, and 40, all correspond to empty lanes that are excluded from analysis.

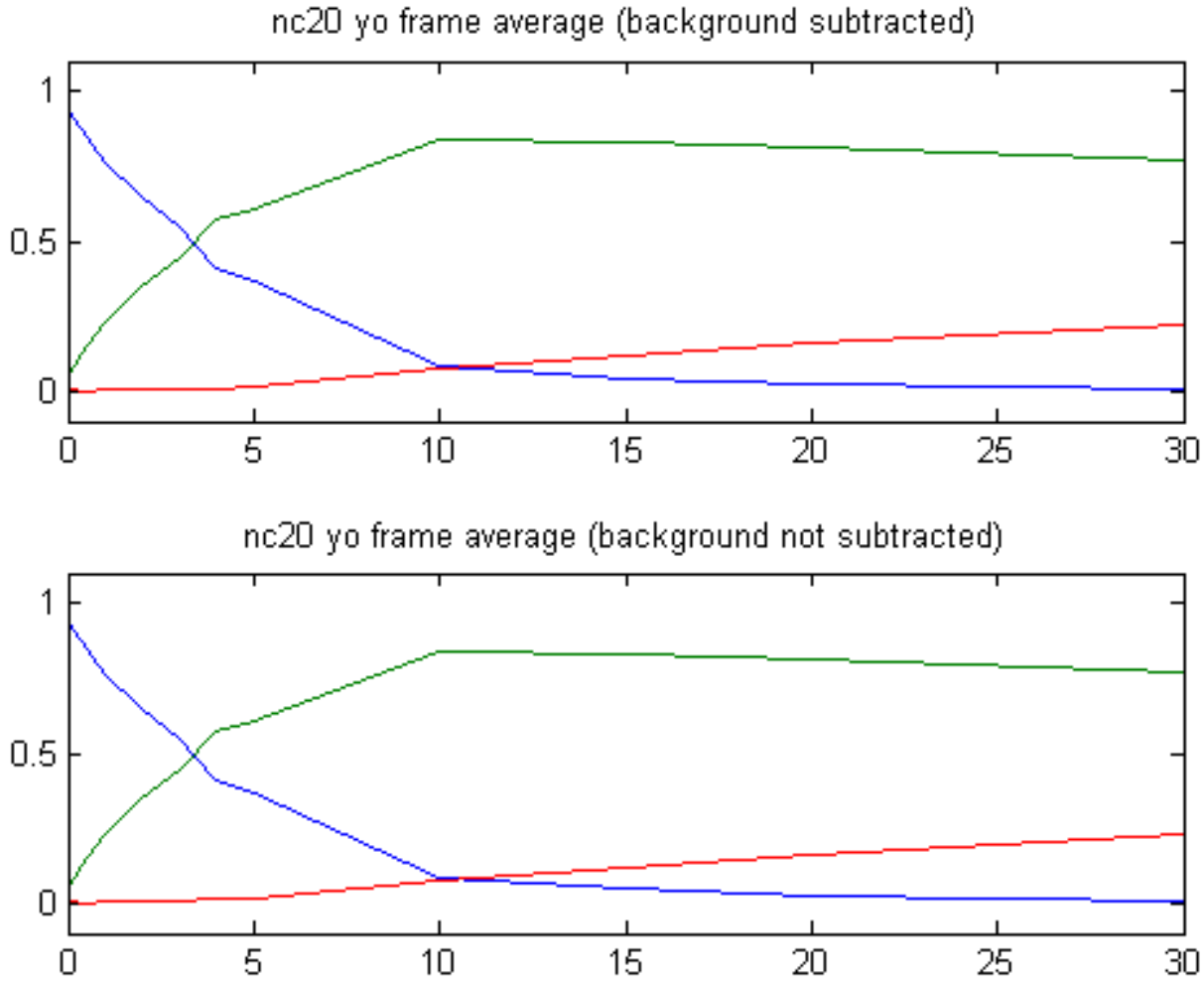

Figure A.4- Quantification of the three plasmid forms from the initial gel image. The blue curve is the supercoiled species, the green curve the nicked species, and the red curve the linear species. Each timepoint results from a triplicate analysis. The data is displayed with and without background correction. In many cases, the effect of the background subtraction is minimal.

APPENDIX B

AUTOMATED QUANTIFICATION OF DNA MOLECULE STRAND CLEAVAGE

The following details the data processing scripts used to analyze the SMI experiments presented in Chapter 3. In these experiments, DNA molecules were hydrodynamically elongated in a microfluidic flow cell and observed at different irradiation powers.To measure the cleavage rates of hydrodynamically elongated DNA molecules, time lapse videos were taken that recorded the state of the DNA strands in the microscope field of view.As the DNA strands accumulated single strand breakages, the DNA molecules would cleave and retract to the anchor points. This was visually manifested as a transition from a linear structure optically resolved against a dark background to a pair of bright points, located at the termini of the elongated molecule. A single field of view could contain tens or hundreds of elongated molecules, and videos would be collected for several hundred frames. It would be particularly onerous to attempt to quantify the intact molecules in each frame by hand, as well as prone to human error or bias towards the brightest strands. The challenge then became to automate the process of counting how many intact DNA molecules persisted at each time point (video frame).

In response, a two-step process was developed, using scripts termed *DNAid1* and *DNAid2stack*. The first script used the initial video frame, when all molecules were still intact, to generate a binary mask of the DNA molecules that qualified for further analysis. This script required the user to view the image and manually select which strands would be used in the analysis. In the second step, this mask was subsequently applied to every frame thus selecting

the regions in each video frame that corresponded to the location of the DNA molecules initially selected. At each frame in which a selected DNA molecules ruptured, the initial number of recorded DNA molecules was incremented down. Thus a running count of the remaining DNA molecules could be automatically generated.

Since the data consisted of a large number of sequential images, the video was saved as a .tif image stack using ImageJ. Early frames before irradiation was initiated were truncated at this point.

Using *DNAid1*, the first image frame was selected and entered by the user as the *name* variable:

```
%DNAid1
clear all

sigma=1;
blur=30;

%read in file
name='od 0-74';
im_raw=imread([name '.tif']);

%Create filter and apply to background and image
H=fspecial('gaussian',60,30);
bkg=imfilter(im_raw,H,'replicate','conv');

H2=fspecial('gaussian',10,1);
im_av=imfilter(im_raw,H2,'replicate','conv');
im_bkg=im_av-bkg;
stdev=std2(im_bkg);

im_blur=medfilt2(im_raw,[blur,3]);
im_blur_bkg=im_blur-bkg;

im_bw=(im_blur_bkg>(sigma.*stdev));
im_skel=bwmorph(im_bw,'skel',inf);

im_plot=uint16(zeros([size(im_av),3]));
im_plot(:,:,1)=uint16(2^16*im_skel);
im_plot(:,:,2)=im_av;
% figure(1)
% subplot(1,2,1), imshow(im_av)
% subplot(1,2,2), imshow(im_plot)
```

```
se=strel('rectangle',[5,5]);
im_mask=imdilate(im_skel,se);
im_plot2=uint16(zeros([size(im_av),3]));
im_plot2(:,:,1)=uint16(2^15*im_skel);
im_plot2(:,:,2)=4*im_bkg;
im_plot2(:,:,3)=uint16(2^15*im_mask);

figure(1), imshow(im_av)
figure(2), imshow(im_plot2)
figure(3), imshow(im_mask)
M=bwselect;

L=bwlabel(M);
objects=max(max(L));

figure(4)
im_plot3=uint16(zeros([size(im_av),3]));
im_plot3(:,:,1)=uint16(2^16*im_skel);
im_plot3(:,:,2)=4*im_bkg;
im_plot3(:,:,3)=uint16(2^15*M);
subplot(1,2,1), imshow(im_av)
subplot(1,2,2), imshow(im_plot3)

save([name ' mask'],'M', 'H', 'H2')
```

The image was smoothed with a coarse and fine median filter, the former to obtain the background intensity and the latter to remove noise. The image was then background subtracted to heighten the contrast and then thresholded to convert to a binary representation. Since the features of interest, the elongated molecules, were essentially wavy lines, the entire image was skeletonized. This is a morphological image processing function that reduces a feature to the minimum number of connected pixels, and had the effect of transforming the DNA molecules into single-pixel wide lines. The resulting image was then dilated significantly, expanding the width of the lines to several pixels wide. The MATLAB function *bwselect* was employed to enable the user to manually select which of the remaining features present would be retained in the final mask. A series of overlaid false-color images were presented at this step to assist in selecting the correct features, thus generating the final analysis mask.

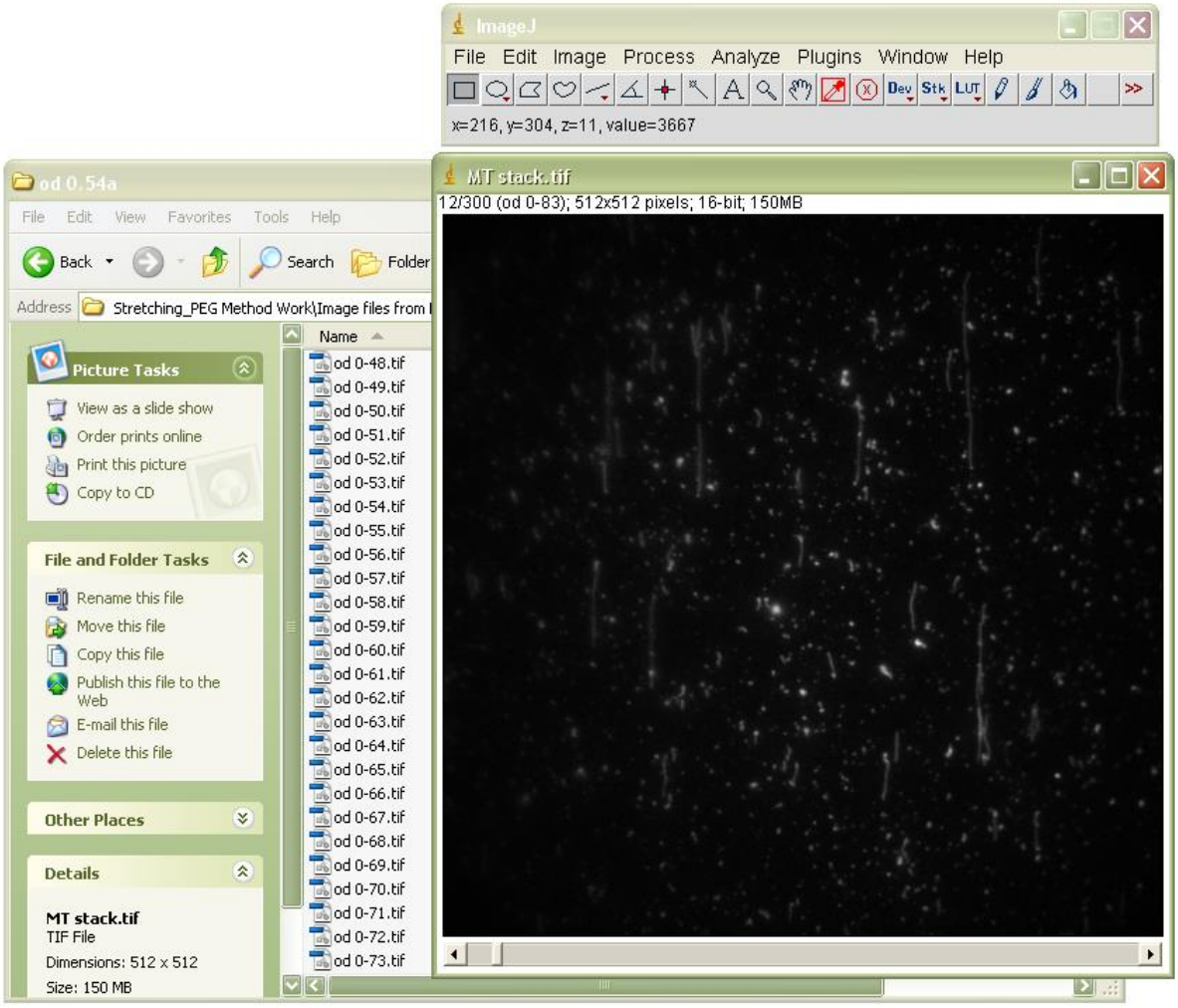

Figure B.1- Compiling datafiles into an image stack. The original data recording the time dependent cleavage of the DNA molecules is saved as a video file. Software included with the EMCCD camera system are used to export the video files as a collection of .tif images, with each video frame as a separate image. For data handling in MATLAB, ImageJ is used to compile the images into a stack. At this point, undesirable initial frames, before laser illumination begins, can be discarded.

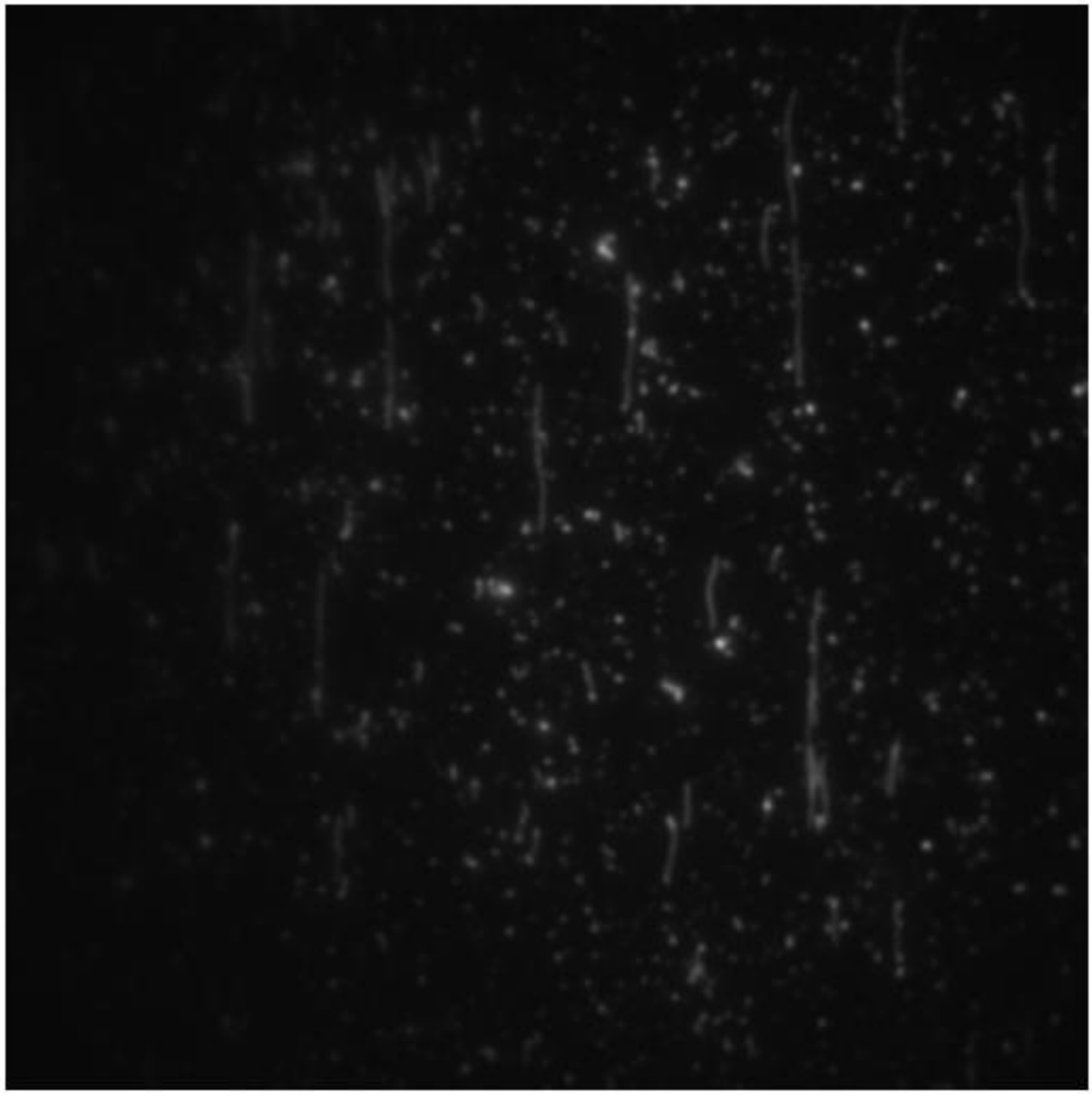

Figure B.2- Initial frame of a time-lapse movie recording the cleavage of elongated DNA molecules. The bright features against the dark background as dye-stained DNA molecules tethered to a passivated glass surface. Not all molecules of extended to the same extent or are poorly resolved from neighboring strands.

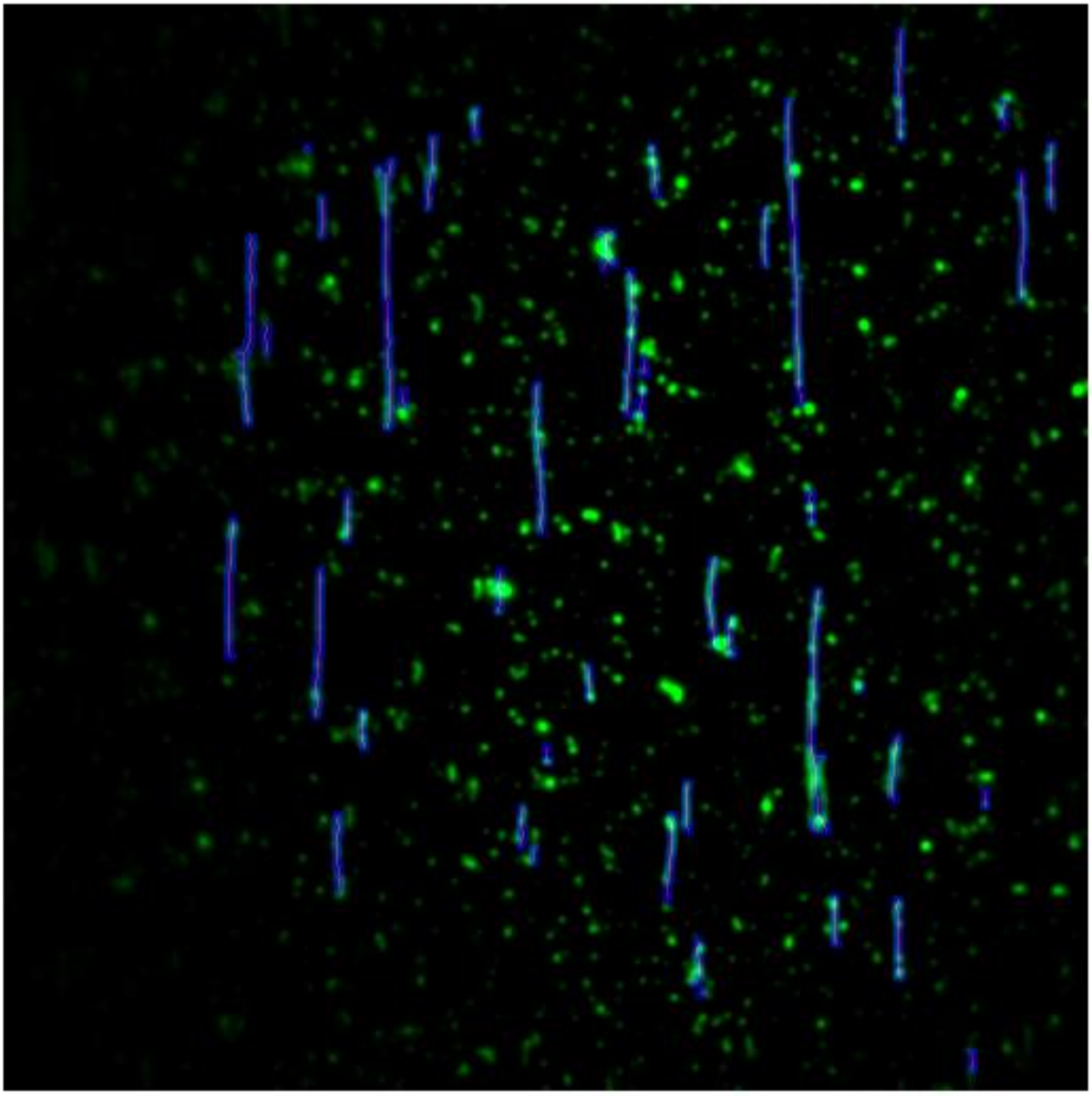

Figure B.3- False-color output used to guide user selection of intact DNA molecules. The initial recorded image, post filtering, is shown in green. The DNA skeletons are shown in magenta, while the dilated DNA skeletons are shown in blue. Overlaid images such as this are useful to visually associate each dilated skeleton with the original DNA feature, as some information, such as overlapping molecules or DNA fragments, can be obscured in the binary skeletonized image.

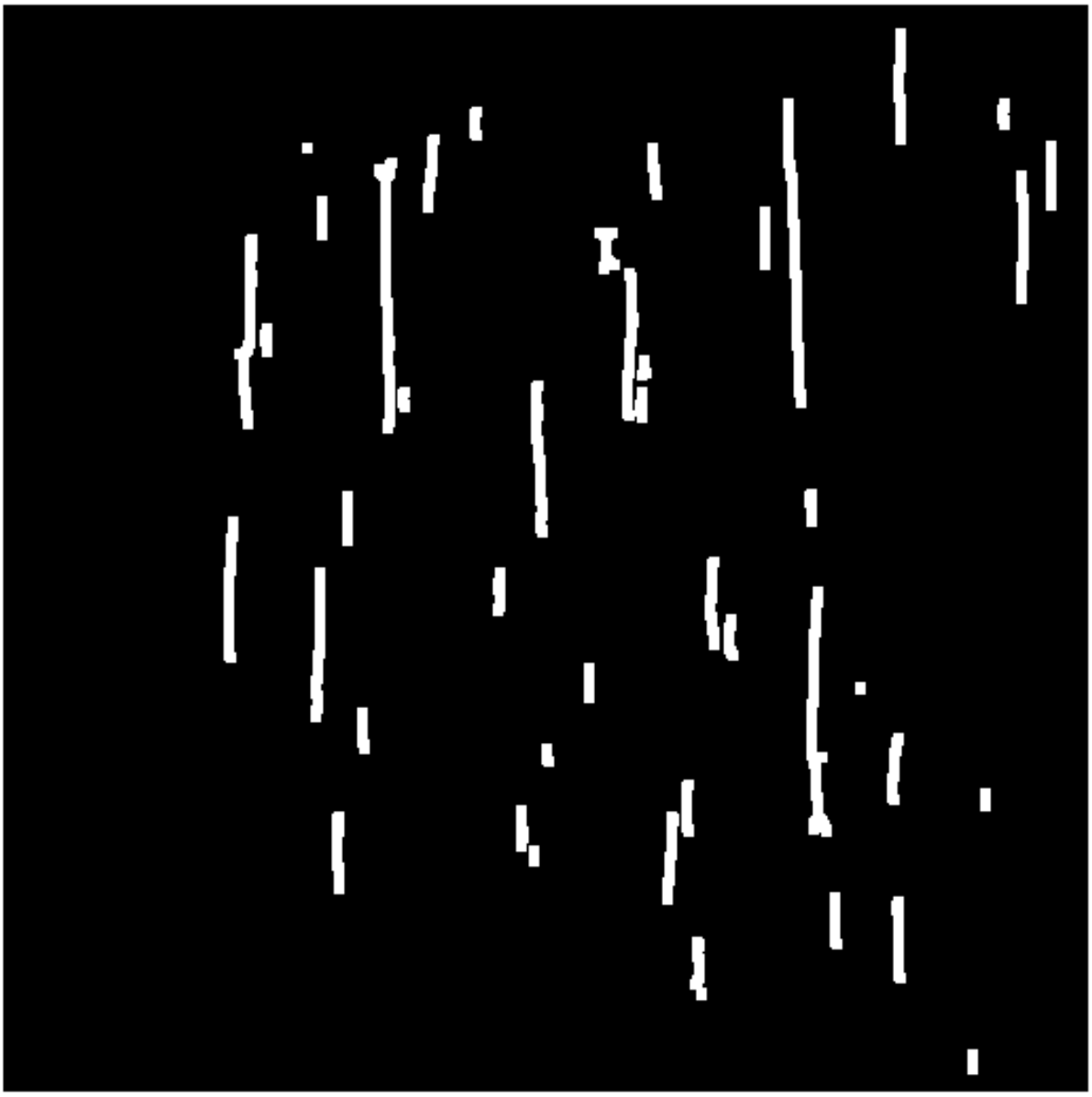

Figure B.4- DNAid1 output enabling user selection. This is the resulting output after the initial image was median filtered, background subtracted, converted to a binary image, and skeletonized. The white features on the black background correspond to the regions of the image containing DNA molecules. Using the MATLAB function bwselect, the user can select each binary feature with a single mouse click. All pixels corresponding to the binary feature are then saved, and later used to apply a mask to all frame of the cleavage video.

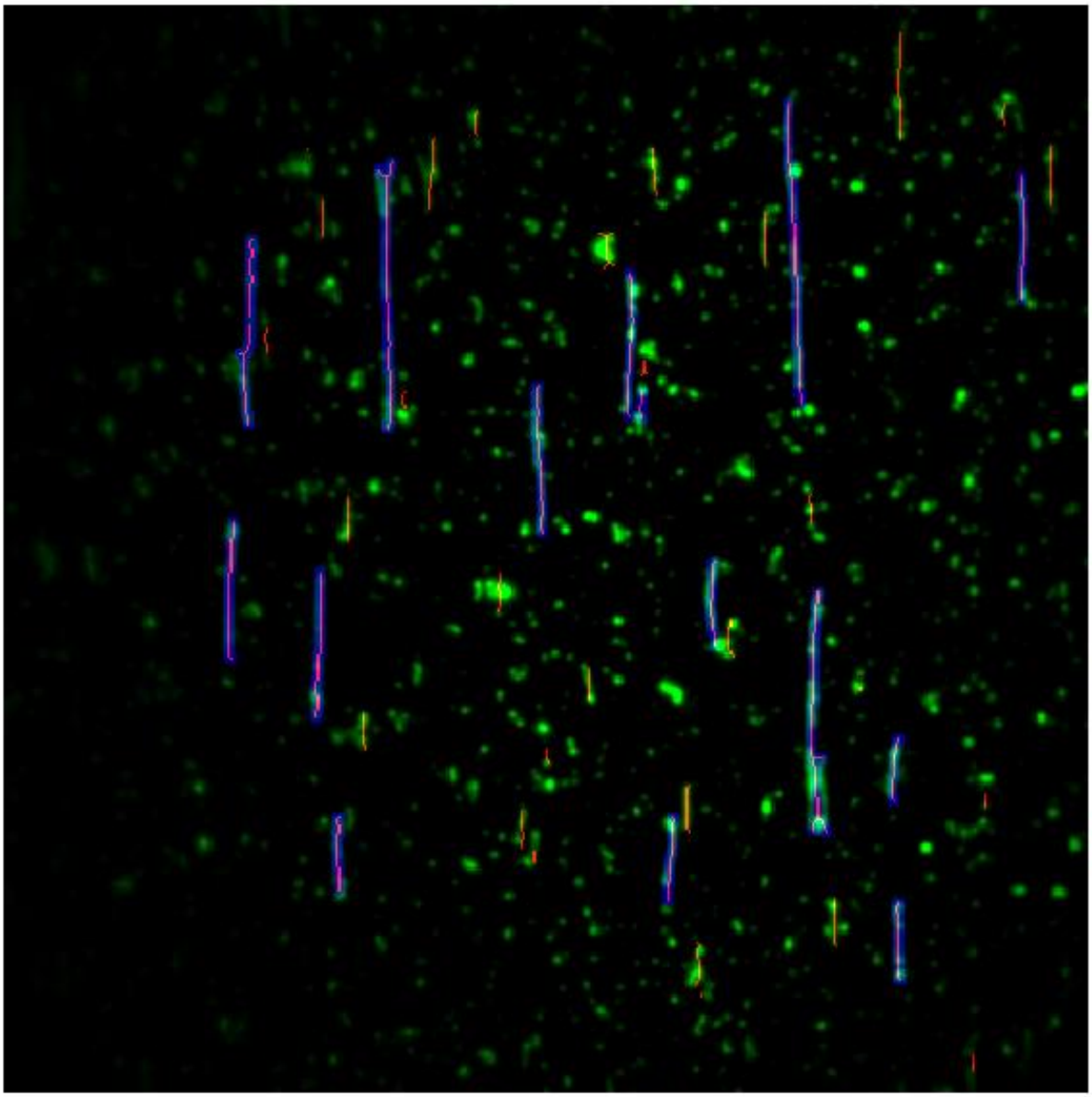

Figure B.5- Final output and resulting mask. This is the last output presented to the user, indicating the selected features of the final mask. The blue features are the dilated skeletons that will be retained in the subsequent analysis. The red features are the original skeletonized DNA molecules, notice that only a subset are selected for analysis. The original image is again presented in green, enabling the user to verify that overlapping or problematic molecules are excluded. While some intact molecules maybe missed by the user in the initial selection, since

the retained molecules are accurately tracked throughout, the effect is negated. If the mask is deemed defective, the script would be run again.

Following creation of the image mask, the script *DNAid2stack* is used. This program loads the image mask and the image stack. The mask is used to isolate the regions of interest in every frame of the stack. Each region of interest contains a skeletonized DNA image; the length of skeletons in subsequent frames are compared. If a skeleton length decreases by more than the value of the standard deviation of the intensity along the length of the skeleton, it is counted as a rupture. These rupture events are recorded for each object identified in the initial mask, thus the program is iterating over the number of objects in the mask and the number of frames:

```
%modified 5/2/11 to correct for when object projections into "lengths"
% are under the threshold.  This caused a condition of not initializing the
% lengths vector, cuasing an error when the max of that vector was asked
% for.
clear all

b=2;% for chopping off blank slides
delay=0.03053;%for converting slide nubmer to time
%scale factor for length threshold
scale=0.25;

%read in file
name='MT stack';
info=imfinfo([name '.tif']);
number=numel(info);
for j=1:number
    A(:,:,j)=imread([name '.tif'],j);
end
image=0:(number-1);

load('od 0-74 mask')
L=bwlabel(M);
objects=max(max(L));

for k=1:number

    im_raw=A(:,:,k);

    bkg=imfilter(im_raw,H,'replicate','conv');
    im_av=imfilter(im_raw,H2,'replicate','conv');
    im_bkg=im_av-bkg;
    stdev=std2(im_bkg);
```

```
for idx=1:objects
        mask=uint16(L==idx);
        obj_n=im_bkg.*mask;
        proj_m=sum(mask,2);
        len1=find(proj_m>0,1);
        len2=find(proj_m>0,1,'last');
        y=(0:(len2-len1))';
        proj_n=sum(obj_n,2)./sum(mask,2);
        projection=proj_n(len1:len2);

        projection_bw=(projection>(scale*stdev));
        l=bwlabel(projection_bw);
if max(l)>0
            lengths=zeros(1,max(l));
for j=1:max(l)
                obj=(l==j);
                lengths(j)=sum(obj);
end
            length(k,idx)=max(lengths);
else
            length(k,idx)=0;
end

end

end

%ADDED BY Michael Tycon to count lengths>30.
%coding by lengths chops off blank slides in length:
longs=sum(length(b:size(length,1),:)>30,2);
time=(b:size(length,1))-b;
%converts slide number into exposure duration
time=time*delay;
figure
plot(time,longs,'o');
xlabel('Time (sec)');ylabel('Number of Intact Strands');
save('breakagecurve','time', 'longs', 'length', 'idx', 'objects')
```

To use DNAid2stack, the user must enter five parameters. The name of the image stack (*name*) and mask (entered a string next to the load function) are required. Additionally, the number of early slides to be excluded (*b*), the video frame rate (*delay*), and a scaling factor used to distinguish the skeletons above the background (*scale*), need to be provided. The script saves the results and outputs a plot of the number of intact molecules versus time.

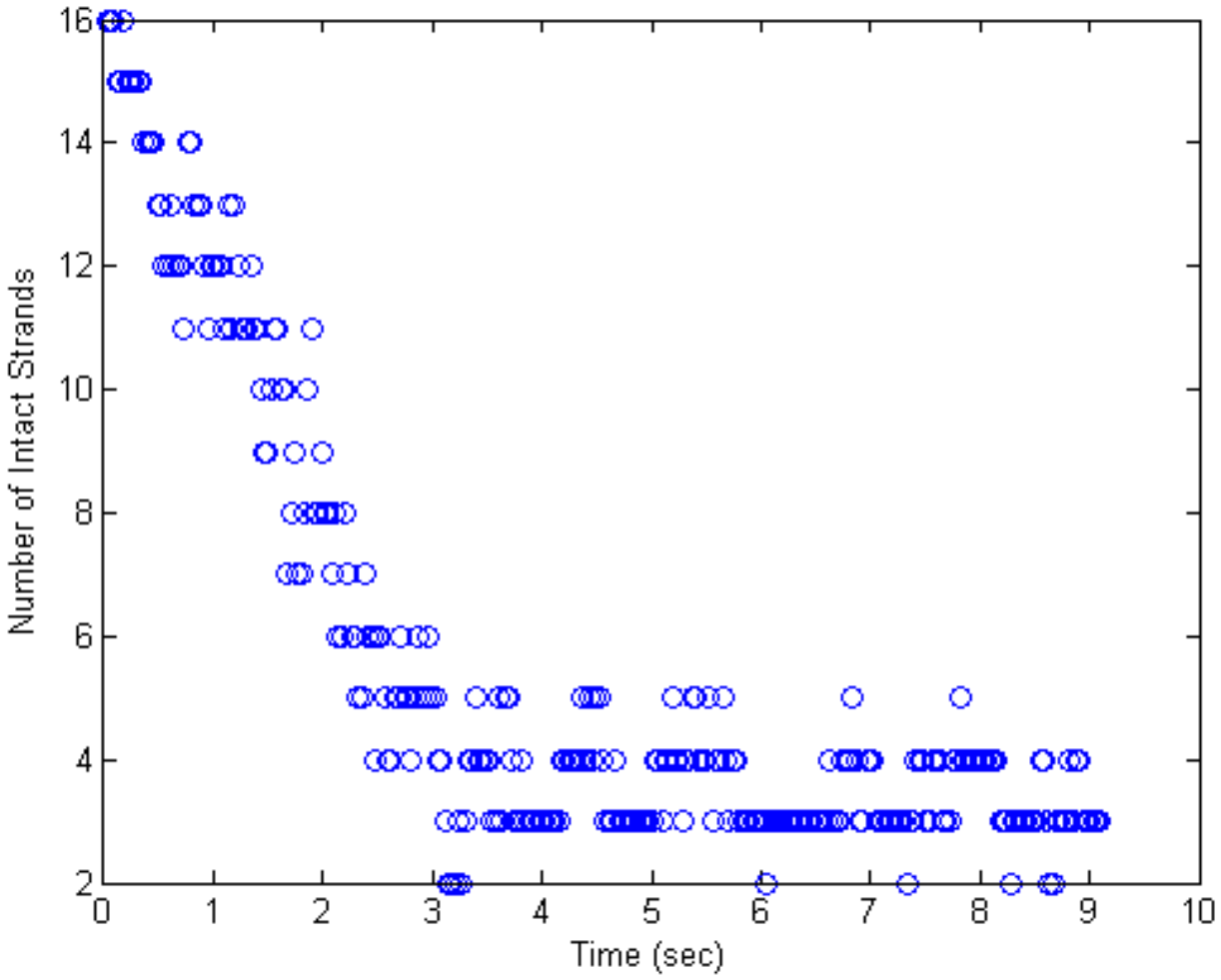

Figure B.6- Output of the DNAidstack2 program. Here, the initial user defined mask identified 16 unique DNA molecules to track over 300 image frames. Only 10 seconds of video were required to capture the breakage of nearly all of the selected molecules. Due to background noise, determining the cleavage of all the molecules is difficult, thus a slight offset sometimes remains. Additionally, frame to frame image brightness fluctuations can complicate the analysis. These complications result from the accumulation of un-intercalated dye on the flow cell surface, which increases the background brightness relative to the stained molecules. This can be seen as the oscillation in the remaining molecule count at long times. If the program

functioned flawlessly, the count would decrease monotonically. The slight “jitter” that results is minor and does not affect the downstream analysis.

APPENDIX C

AUTOMATED "SHOTGUN PTFRAP" IMAGING PROCESSING PROGRAMS

This section supports Chapter 4, and details the image processing program used to automate the selection of FRAP datapoints collected in lab with the desired physical region for data analysis. During data collection, for each cell investigated, FRAP datapoints were collected across the entire nucleus using a coarse thresholding criteria. After data collection, FRAP points corresponding to particular nuclear sub-regions, specifically interchromatin space devoid of chromatin signal, were selected and retained for further analysis. While detailed in Chapter 4, the general process was to use the images collected of both color channels as the basis of binary masks. These binary masks were then used to identify regions matching a strict criterion of distance away from interfering structures. The datapoint positions that met the criteria were then matched with the datapoint positions collected during the experiment, and only the FRAP recovery data from these position used in downstream analysis. Thus, by modifying the masking programs, all datapoints collected during the initial experiments can be segregated without the need for additional experiments.

Below is the masking program, called *ptFRAP_autoptselect_imageanalysis_v2*:

```
%Process each image set seperately, save an array of selected points per
%zone
clear all
filename = 'Rpb9_set2_interzone';
zone=26;
%Value to thicken and erode for the particle size thresholding:
number=8;
Cyel1=imread([filename num2str(zone) '_yel_z0_r0.tif']);
Cgre1=imread([filename num2str(zone) '_gre_z0_r0.tif']);
%Apply a median filter to image to reduce noise:
Cyel=medfilt2(Cyel1, [5 5]);
Cgre=medfilt2(Cgre1, [5 5]);
```

```
%Threshold image to get Binary Output:
[lv1,mask1yel] = thresh_tool(Cyel);
[lv2,mask1gre] = thresh_tool(Cgre);

%Blurr and process images to make masks:
%Yellow:shrink nucleus outline to avoid periphery
%Apply Particle Size Thresholding to Nucleus (Yellow Channel):
objs=bwconncomp(mask1yel);
numPixs = cellfun(@numel,objs.PixelIdxList);
maxnum=find(numPixs==max(numPixs));
tmask=logical(zeros(512,512));
tmask(objs.PixelIdxList{maxnum})=1;

%se=strel('disk',number1);
se=strel('diamond',number);
tmaskM=imerode(tmask,se);

tmaskM2=bwmorph(tmaskM,'thicken',number);
%imshow(tmaskM2)

objs2=bwconncomp(tmaskM2);
numPixs2 = cellfun(@numel,objs2.PixelIdxList);
maxnum2=find(numPixs2==max(numPixs2));
tmaskM2L=logical(zeros(512,512));
tmaskM2L(objs2.PixelIdxList{maxnum2})=1;

comp=uint8(zeros(512,512,3));
comp(:,:,1)=uint8(255*tmask);
%comp(:,:,2)=uint8(255*tmaskM);
comp(:,:,3)=uint8(255*tmaskM2L);

%This figure shows the selected largest region:
% figure(1)
% imshow(comp)

mask1yel3=bwmorph(tmaskM2L,'erode',5);

%Green: block out polytenes to avoid banded regions
mask1gre3=bwmorph(mask1gre,'dilate',3);

%Some of these plots are not useful:
% figure (2)
% subplot(2,3,1)
% imshow(Cyel)
% subplot(2,3,2)
% imshow(mask1yel3)
% subplot(2,3,3)
O=cat(3,100*mask1yel3,Cyel,zeros(512,512));
% imshow(O)
% subplot(2,3,4)
% imshow(Cgre)
% subplot(2,3,5)
% imshow(mask1gre3)
% subplot(2,3,6)
```

```
N=cat(3,100*mask1gre3,3*Cgre,zeros(512,512));
% imshow(N)
 %Find intersection of both masks:
mask1gre3=-(mask1gre3);
mask=mask1yel3+mask1gre3;

%Apply Grid to determine possible Bleach and Control Pt Locations
%Control region bleaches:
gridvect2=zeros(512,1);
for i=0:25
gridvect2(20*i+10)=1;
end
[A2,B2]=meshgrid(gridvect2,gridvect2);
grid2=A2.*B2;
%apply mask to grid and get corrdinates:  Pass to laser the bleach points
mask_grid2=mask.*grid2;
[cy2,cx2]=find(mask_grid2==1);
bleachpts2=length(cx2);%pass this value as the number of control pts taken

%Bleach region: must have same number of pts as control region
gridvect=zeros(512,1);
for i=0:25
gridvect(20*i+1)=1;
end
[A,B]=meshgrid(gridvect,gridvect);
grid=A.*B;
%apply mask to grid and get corrdinates:  Pass to laser to bleach points
mask_grid=mask.*grid;
[cy,cx]=find(mask_grid==1);
bleachpts=length(cx);%pass this value as the number of bleach pts taken

%Ensures equal number of control and bleach points
if bleachpts2==bleachpts
   cx=cx;
   cy=cy;
   cx2=cx2;
   cy2=cy2;
end

if bleachpts2<bleachpts
    cx=cx(1:bleachpts2);
    cy=cy(1:bleachpts2);
end
if bleachpts2>bleachpts
    cx2=cx2(1:bleachpts);
    cy2=cy2(1:bleachpts);
end



figure (3)
subplot(1,2,1)
imshow(O)
xlabel('Green Channel Mask')
subplot(1,2,2)
imshow(N)
```

```
xlabel('H2B-RFP Mask')

ptFRAP_postcollection_dataselection%contains figure 4
```

The program needs to be run for each z-section in which data was collected. The user must supply the information identifying the image corresponding to the two color channels, designated by the variables *filename* and *zone.* A core functionality of this script is the thresh_tool, a program freely available from http://www.mathworks.com/matlabcentral/. Additionally, the MATLAB function *bwconncomp*, is used to automatically select the nucleus from the surrounding signal, bypassing the need for the user to manually select the largest feature in the field of view.

APPENDIX D

SUPPORTING INFORMATION FOR THE CHAPTER 4

1. *High expression levels of fusion proteins are not responsible for the observed anomalous diffusion*

The Rpb3-GFP and Rpb9-GFP fusion proteins are exogenous insertions expressed under the control of the GAL4 driver system and believed to be functional due to recruitment to HSP promoter sites [1]. As a result they are highly over-expressed compared to the native, untagged RNAPII subunits. To test if the over-expression was creating a population of unincorporated subunit that was being manifest as apparent anomalous diffusion, we crossed our Rpb9-GFP with a GAL4 driver under the control of a heat shock induced promoter (Bloomington Stock Center #1799).(**d**) The expression level of this cross, Rpb9-GFPx1799, can be lowered by raising the fly larvae at 18°C (red bars) and was determined to reduce expression levels by up to 50% compared to the Rpb9-GFPxH2B-mRFP line raised at 22°C (black bars). The mean expression levels of these two populations were found to be statistically different ($p<0.001$). While this construct did not have the chromatin labeled by the H2B-mRFP histone protein, the Rpb9-GFP showed strong exclusion from chromatin regions (determined previously) still enabling us to restrict the FRAP analysis to the interchromatin space. (**a**) The FRAP recoveries and (**b**) normalized recoveries for the high (black) and low (red) Rpb9-GFP expression levels flies are shown. (**c**) Within experimental error, the effective diffusion coefficient and anomlity value of the reduced expression line matched the results found using the Rpb9-GFPxH2B-mRFP line. Thus we are confident that the over expression is not responsible for the anomalous diffusion.

This could not be repeated for the Rpb3-GFP construct since it is expressed by a GAL4 driver sequence previously bred into the fly line.

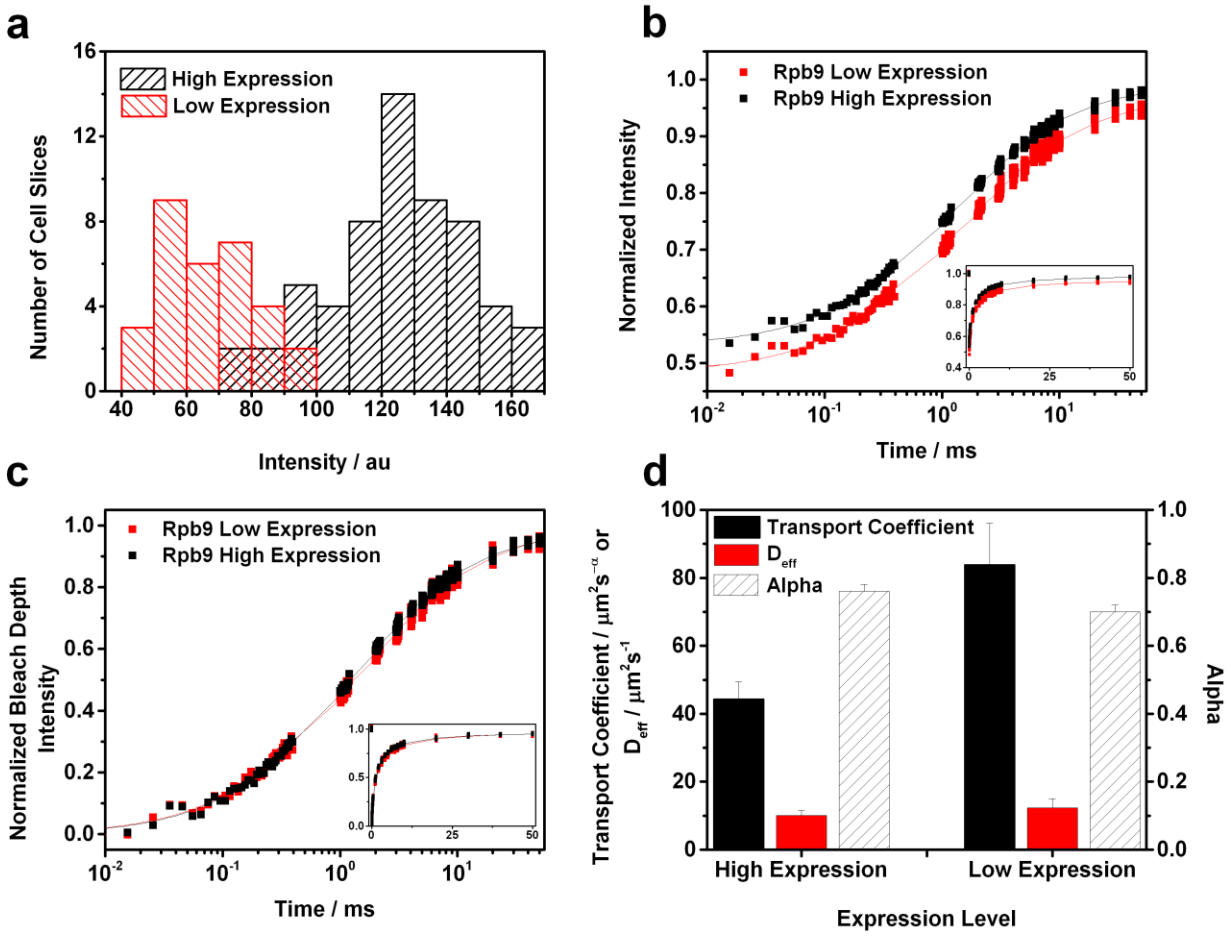

Figure D.1- High expression levels of fusion proteins are not responsible for the observed anomalous diffusion

2. Determining the resolution of the Point FRAP method

For slow moving species, determining the diffusion coefficient is difficult if the FRAP curve does not fully recovery to the pre-bleach level on the time course of the measurement. Despite the rapid time resolution of our data collection method, we are limited in how slow a diffusion component we can accurately measure by the 50 ms time duration of our recovery collection. If Brownian diffusion is assumed, our fitting algorithm estimates the final recovery extent based on the slope of the FRAP curve once it begins to level off. Further, the estimation of the recovery extent will strongly affect the estimated diffusion coefficient. For very slow moving species, the recovery will be very shallow and the algorithm is unable to accurately estimate the diffusion coefficient. This became a significant concern when applying the distribution model [2] as a threshold for reliable determination of diffusion coefficients needed to be established. We chose to empirically evaluate which diffusion coefficients were reliable by applying our fitting algorithm to simulated data and determining where the estimated diffusion coefficients began to deviate from the input value. (**a**) FRAP recovery curves were simulated that correspond to diffusion coefficients from 0.01 to 1000 $\mu m^2/s$. As can be seen, the majority of the curves exhibit a significant recovery, but the slow moving components are nearly flat on the 50 ms timescale of the simulation. (**b**) The fitting algorithm was applied to each curve and the estimated diffusion coefficient was plotted against the initial input value. We determined the diffusion coefficient estimation was accurate with as little as 10.3% recovery (**a**-horizontal black line), corresponding to a diffusion coefficient of 0.04 um^2/s (**b**-vertical black line). (**c**) Next, white noise was added to the FRAP curves resulting in simulated data with a signal to noise ratio (SNR) of 35 dB. This SNR corresponds well our experimental FRAP data. Again, we

applied the fitting algorithm to the noisy data and compared the estimated diffusion coefficients to the input values. At this SNR, the estimations begin to deviate once the recovery is less than 47.6% complete (**c**-horizontal black line), corresponding to a diffusion coefficient of 0.29 um^2/s (**d**-vertical black line). Thus we can see the accuracy of the fitting depends on the SNR of the data. Erring on the side of caution, we rejected any diffusion components that showed less than a 50% recovery. This method outlines a framework for evaluating the robustness of a FRAP fitting method as long as the SNR of the data can accurately be estimated.

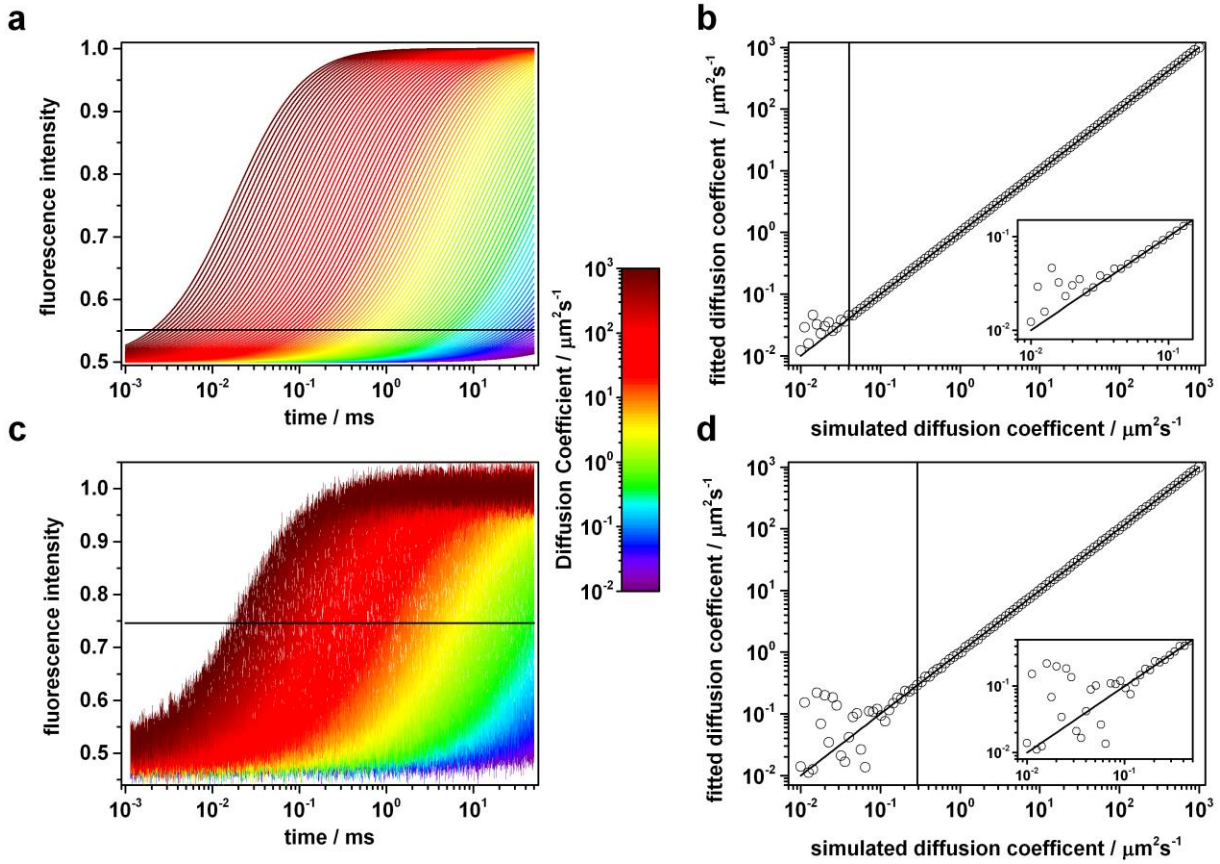

Figure D.2- Determining the Resolution of the Point FRAP Method

3. Establishing the robustness of the Distribution model on experimental data

As presented in the Results and Discussion, the Rpb3 datasets indicate a bimodal distribution. We wanted to ensure the robustness of the Distribution model to predict bimodal distributions without a bias predicated on the initial component amplitudes. To achieve this, we tested the output of the Distribution model in response to different initial amplitude profiles, as well as different fitting protocols. Four sets of initial conditions were tested: (1,2-Gaussian) shaped the initial amplitudes in a Gaussian envelope with 35 or 15 dB noise added, (3,4-Flat) provided 35 or 15 dB Gaussian white noise as the input. To test for reproducibility, each input condition was tested three times. In the first, unbiased implementation (panels **b,e,h,k,n**), the input profile amplitudes were floated to achieve a best-fit to the FRAP data. The output distribution was then smoothed with a median filter. This process was repeated five times until the fit residuals no longer improved. The last step omitted smoothing to prevent distorting the output. All the outputs are overlaid indicating the similarity regardless of input profile. Next, the effect of biasing the distribution to a single component by implementing a Gaussian smoothing step was tested. A five-step procedure was used, but in contrast to the previous method, between the third and fourth smoothing steps the output was fit to a Gaussian envelope. The final fit output was not forced to a Gaussian to reveal the most stable output. The fitting outputs from all twelve input distributions are shown (panels **c,f,l,n**); again the outputs are (1) very similar and (2) show the same structure as the un-biased fitting method. The results of the twelve outputs for both fitting methods were averaged and compared (panels **a,d,g,j,m**), indicating nearly identical distributions. This indicates that random noise on the input does not affect the output and the distribution fit find the most

stable output. This test was significant for the Rpb3 distribution results. If biasing the output to one component altered the final output away from a bimodal fit, then the distribution model algorithm could not be considered robust. However, since even when the fit was forced to conform to a single peak it still "stepped away" to a bimodal fit on the next iteration, the fitting method was considered stable.

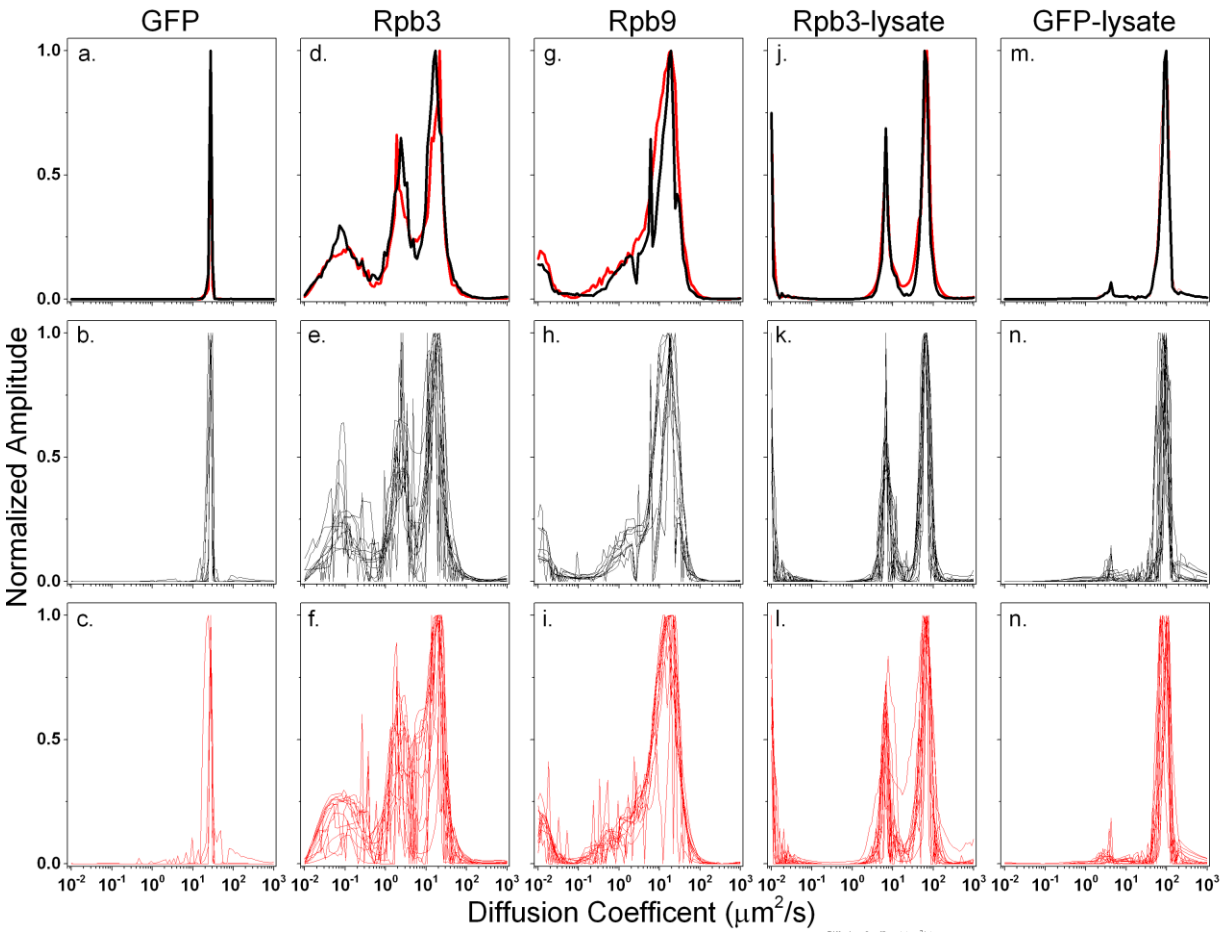

Figure D.3- Establishing the Robustness of the Distribution Model on Experimental Data

4. Slow Diffusion Components under the FRAP resolution method are not required for an accurate fit

After confirming that the Distribution modeling can robustly determine the number of components that comprise a FRAP curve and having established the FRAP resolution limit, we chose to investigate how accurately the retained components recapitulated the original data. The output distributions (panels **b,d,f,h,j**, black lines) were truncated at 0.30 $\mu m^2/s$ (red lines), and renormalized so the total distribution summed to unity. This slightly increased the amplitudes of the retained components. These truncated distributions were used to establish a fit to the data (panels **a,c,e,g,I**, fit to all components black line, fit to truncated distribution red line). For the Rpb3 *in vivo* data, the retained components do alter the recovery dynamics, shifting the curve to a faster recovery. For all other samples, the fits are unchanged.

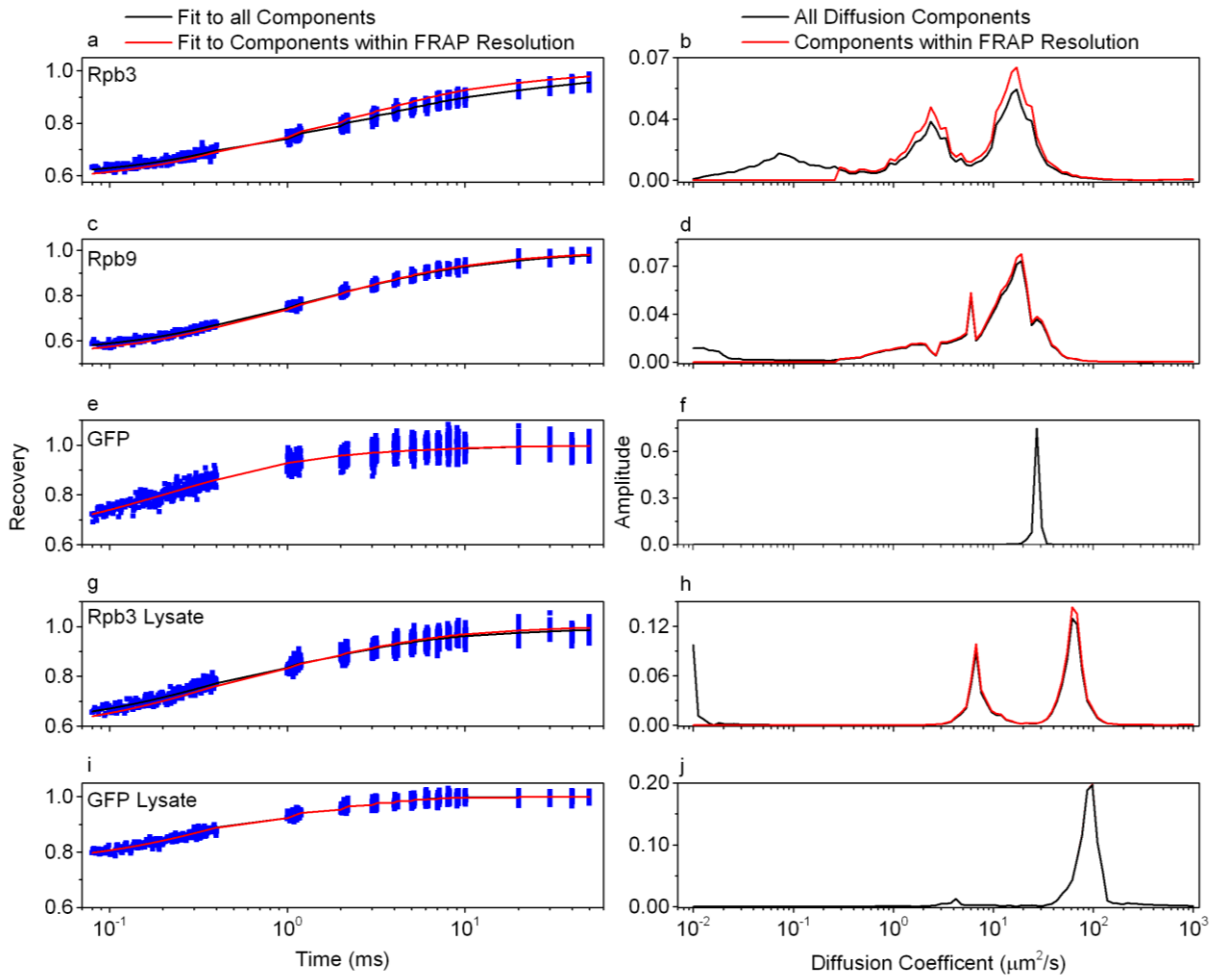

Figure D.4- Fit quality excluding diffusion components under FRAP resolution

5. FRAP fitting results for each dataset

For each experiment, several datasets were collected and the resulting raw data averaged together to yield finalized data with a high SNR. The finalized data was fit with the apparent anomalous diffusion and distribution models. To ensure that the averaging of several datasets did not distort the final results, each individual dataset was fit with the apparent anomalous diffusion model. The results are compiled below. Typically, the subset of the finalized data shows nearly the same anomlity and effective diffusion coefficient, but the 95% confidence error intervals are larger than if the datasets are compiled. As shown, averaging the fit outputs of the subsets is not identical to fitting the averaged data. This method is in line with how the data was analyzed in Daddysman et al. 2011.

Table D.1-FRAP diffusion fitting results for individuals datasets and ensemble averages

Conditions	Sample	Set	Gamma (um2/s^a)	D(um^2/s)	Alpha
		I	-	32.7±16.1	0.99±0.08
	GFP	II	-	36.2±20.1	1.00±0.09
		III	-	27.5±20.1	1.00±0.12
		Ensemble*	-	32.0±6.0	1.00
		I	70.8±11.7	21.0±4.5	0.78±0.06
		II	37.3±15.9	6.2±3.9	0.73±0.07
In vivo		III	54.1±33.3	4.4±5.0	0.64±0.10
(live polytenes)	Rpb3	IV	105.6±37.3	7.4±5.1	0.58±0.06
		V	271.9±130.0	9.2±13.6	0.43±0.08
		VI	90.3±23.5	5.0±2.5	0.57±0.05
		Ensemble*	69.1±10.5	5.5±1.4	0.62±0.03
		I	45.7±7.2	7.9±1.7	0.73±0.03
		II	38.9±14.3	9.70±4.8	0.78±0.06
	Rpb9	III	30.7±8.9	7.6±2.9	0.78±0.05
		IV	46.8±7.3	12.8±2.6	0.78±0.02
		Ensemble*	44.4±5.0	10.0±1.5	0.76±0.02
		I	98.0±50.0	79.8±43.0	0.96±0.07
	GFP	II	75.1±33.8	71.1±32.5	0.99±0.07
		Ensemble*	112.2±37.5	79.1±29.0	0.92±0.05
In vitro					
(cell lysate)		I	69.4±11.3	43.8±7.85	0.91±0.05
		II	246±136.7	41.2±40.1	0.65±0.08
	Rpb3	III	85.4±37.4	30.6±17.2	0.81±0.07
		IV	115.4±45.9	23.2±13.7	0.72±0.06
		Ensemble*	150±36.4	33.0±11.7	0.72±0.04
In vivo		I	83.9±12.2	12.3±2.65	0.70±0.02
Low Expression	Rpb9	II	118.3±21.9	10.7±3.3	0.65±0.03
Level					
		Ensemble*	97.3±12.1	11.7±2.2	0.67±0.02

*Parameters resulting from fitting the average of all the listed datasets. This procedure improves the fitting results by increasing the SNR of the data.

REFERENCES

(1) Yao, J.; Ardehali, M. B.; Fecko, C. J.; Webb, W. W.; Lis, J. T. *Molecular Cell* **2007***, 28*, 978-990.

(2) Periasamy,N., Verkman, A.S *Biophys. J.* **1998***, 75*, 557.