SINGLE MOLECULE INVESTIGATION OF TRANSCRIPTION ACTIVATOR-LIKE EFFECTOR SEARCH DYNAMICS

BY

LUKE WILLIAM CUCULIS

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Chemistry
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Associate Professor Charles M. Schroeder, Chair
Professor Huimin Zhao
Professor Paul R. Selvin
Professor Catherine J. Murphy

Abstract

Recent advances in genetic engineering hold great potential to profoundly change the treatment of human disease. Precise manipulation of genetic material allows for the creation of new disease models and the rapid translation of new therapies into the clinic. Several classes of programmable nucleases allowing for precise and targeted genomic edits are central to the rising popularity, flexibility, and accessibility of gene engineering. Transcription activator-like effectors (TALEs) are one such class that form a powerful gene-editing platform when fused to nuclease domains, thereby yielding TALEN systems. Despite pervasive use of TALENs for editing crops, small animals, eukaryotic stem cells, and human T-cells, remarkably little is known about the molecular mechanisms used to locate and bind their DNA target sites.

This work describes the application of single molecule fluorescence imaging to the study the TALE search process along specific and non-specific DNA templates. Our work provides a molecular-level picture of the dynamics of TALE-DNA interactions, and our results have revealed an apparently unique search mechanism for DNA binding proteins. We directly observe TALEs diffusing along non-specific DNA in one-dimension. Our results show that TALE diffusion occurs in a directionally unbiased and thermally driven manner along double surface tethered and extended DNA templates (Chapter 2). Interestingly, we observe significant intra-trajectory heterogeneity for diffusion of full-length TALE proteins. We further isolate and study the single molecule dynamics of TALE truncation mutants containing only the N-terminal region (NTR), and these results reveal the importance of the NTR for nucleating non-specific binding. We find that the TALE NTR alone is capable of short, rapid non-specific search.

Furthermore, we study the diffusion of a series of TALEs with variable size central repeat domains (CRDs). Taken together with insights from NTR dynamics and the heterogeneity of full-length TALE diffusion, we propose a two-state search mechanism for TALEs that is comprised of rapid search and interspersed periods of local sequence checking along DNA templates.

We further expand our characterization of TALE search by determining the impact of solution conditions, ionic strength, probe size, and the role of hydrodynamic flow on TALE dynamics (Chapter 3). Using this combination of single molecule experiments, we find that TALE diffusion does not fit the traditional definitions of binary classification of DNA-binding protein search, which have been characterized as protein hopping or sliding along DNA templates. Instead, our results suggest a mechanism wherein TALEs encircle DNA templates during search, but form only transient contacts with the DNA backbone. Furthermore, the non-specific search trajectory of TALEs is rotationally decoupled, in contrast to a broad class of other DNA binding proteins including DNA repair proteins and transcription factors. We further utilize a combination of bulk fluorescence anisotropy measurements and single molecule experiments to characterize the effects of divalent cations on TALE binding (Chapter 4). Our results show that TALE specificity is significantly enhanced in the presence of certain divalent cations, which can be attributed to a decrease in non-specific binding affinity. Finally, we generate long DNA templates with TALE target binding arrays at precise locations, and we directly visualize specific binding and localization of TALEs to their respective target sites following 1-D search. Taken together, our results have elucidated the fundamental search mechanism of TALE proteins along DNA templates.

Acknowledgements

I would first like to thank Professor Charles Schroeder for giving me the opportunity to work on this project, for being open to the prospect of starting a completely new, potentially risky area of work in the lab, and for providing guidance and advice. I greatly appreciate all of the freedom I was given to design experiments and pursue directions within the project that were most interesting to me. This freedom undoubtedly provided a rich environment for personal growth as a scientist. I would also like to thank several members of the Schroeder lab and Zhao lab, who helped with aspects of this work and provided support. Drs. Chris Brockman and Amanda Marciel aided in early attempts to generate and image DNA templates. Dr. Melikhan Tanyeri (along with Professor Emeritus J. Douglas McDonald) taught me how to properly handle and align optics, which proved a frequent requirement over the years. Songsong Li provided electron micrographs of quantum dot particles. None of this work would have been possible without the exceptional talents and hard work of Dr. Zhanar Abil, who designed and produced the aldehyde-tagged TALEs and DNA substrates, and provided helpful discussion during the completion of work in Chapter 2. Xiong Xiong provided the 8x binding array DNA templates, which allowed for the specific binding results in Chapter 4 and have opened up exciting new possibilities for continued work.

I would like to thank Professor Huimin Zhao for his continuous support and guidance, and for making such a collaborative project possible. I would like to thank Professor Cathy Murphy and Professor Paul Selvin for serving on my committee and providing helpful feedback during my preliminary exam.

I want to thank my parents, Bill and Kim, for their support over the past 4 years. Without the work ethic I learned from them, this thesis would have either been a lot shorter, taken much longer, or some combination of the two. I also want to thank everyone I have had the privilege to work with at Illinois Business Consulting, especially Andrew Allen, Whitney Smith, James Noonan, Rob Valli, and Josh Gajsiewicz. Being surrounded by so many motivated, talented individuals has been truly inspiring.

Importantly, I want to thank my girlfriend Taylor Plank. Being 686 miles apart for so long has been the most challenging part of completing this degree, but your understanding, compassion, and support made it possible.

I was funded partially by an FMC Corporation Fellowship.

Table of Contents

Chapter 1: Introduction

1.1 Overview

Genome engineering holds the potential to revolutionize the treatment of human disease[1] and the improvement of agricultural production for a rapidly growing global population with shrinking resources[2]. The ability to precisely regulate protein expression and to edit or delete genes across organisms including humans was merely a dream only a few decades ago. Recent advances in the field of genome engineering have drastically improved accessibility to this powerful technique, by lowering cost, improving efficiency, and reducing the technical skill needed[3]. These major advances have been primarily centered around the development of three key targeted nuclease systems[4]: zinc finger nucleases (ZFNs)[5–8], transcription activator-like effector nucleases (TALENs)[9–11], and the clustered regularly interspaced short palindromic repeats/CRISPR-associated protein 9 system (CRISPR/Cas9 system) [12,13].

Although the mechanistic details of how these systems carry out their natural and engineered gene targeting/editing mechanisms differ, these methods are unified by their remarkable ability to induce DNA double-strand breaks (DSBs) at nearly any DNA sequence of interest. Once a DSB has been induced, natural machinery within the cell works to repair the cut via one of two mechanisms for mammalian cells: non-homologous end joining (NHEJ) or homology-directed repair (HDR)[14] (**Figure 1.1**). When a DNA template has been introduced for the edited region along DNA, repair can proceed via HDR, thereby allowing for additions or replacements to a genome. In the absence of an appropriate DNA template, repair will proceed via NHEJ. In this case, some degree of

genetic information will be effective edited out of the genome, thereby allowing for targeted gene knockouts[4].

In recent years, the genome engineering community has capitalized on these two DNA repair pathways in conjunction with the specific editing ability of nucleases, thereby demonstrating an incredibly impressive range of applications of programmable nuclease-based systems. These include generation of disease-resistant rice[15], polled livestock for safer handling[16], new disease models[9], and correction of the gene defect responsible for sickle cell disease[17]. Concurrent with an explosion in the application of gene therapy and gene editing techniques, however, has been a rising public concern over the safety of these techniques, as well as ethical concerns within the genetic engineering community[18]. Although errors induced in genetically engineered crops or disease models can be considered an experimental error and potential source of difficulty, errors realized during gene editing therapies can be disastrous in their effects. The recent report of human embryos being edited using the CRISPR/Cas9 system[19] set off a firestorm of backlash and led to the announcement of a moratorium on germline editing that gained the support from the White House[18]. Combined with an overall increase in vocal critics of genetically modified food products, the current state of gene engineering is at a crossroads. On the one hand, the future of gene engineering seems exceptionally bright: costs have fallen, barriers for implementation have been substantially lowered, the specialized skills required have been drastically reduced, and labs throughout the world are finding new applications on a daily basis. On the other hand, however, major safety concerns have been raised. Thus, the path forward for commercialization of therapies derived from gene editing techniques is highly dependent on the development of

techniques with outstanding safety profiles, which includes molecular tools that can function in a precise manner with low off-target effects on a consistent basis.

Off-target effects in programmable DNA nucleases arise due to binding and editing at unwanted sites along a genome. Unfortunately, off-target effects are a major source of inherent risk in gene editing applications involving programmable nucleases. Although each of the three major programmable nuclease systems has seen marked improvements in target specificity since their introduction, there exist limits to what improvements can be made, due to incomplete understanding of the dynamic molecular-scale processes leading up to an off-target binding event and the associated factors that influence off-target binding. From this view, we need to develop a detailed picture of the search and target recognition processes for programmable nucleases, if rational design of proteins with high specificity is to be achieved. Achieving this improved understanding of the DNA search and bind mechanisms will reveal the molecular-level details and dynamics of these gene editing systems in real-time.

1.2 Single Molecule Fluorescence Microscopy (SMFM)

Single molecule imaging methods have skyrocketed in popularity over the past few years, and these techniques can be used to study molecular-level details of biological systems in real time. Conventional light microscopy suffers from two main drawbacks in the context of studying the dynamics of biological systems. First, conventional optical microscopy provides poor contrast amongst the diverse set biomolecules within a cell or *in vitro* system. Second, standard optical microscopy is unable to spatially resolve features smaller than the diffraction limit of light (~300 nm for visible light). Fluorescence microscopy addresses the first drawback, typically via conjugation of

spectrally distinct fluorescent probes to target biomolecules. SMFM can be used to address the second drawback, by allowing for the localization of individual fluorescence emitters with a greater precision than the diffraction limit of light[20], thereby achieving spatial resolutions less than ~300 nm.

Individual emitters can be modeled as point sources, and their emission point spread function (PSF) can be well approximated by a two-dimensional Gaussian function. In this way, a single fluorescent emitter can be localized to within a greater precision than the diffraction limit of light, given that sufficient photons are collected during image acquisition[21]. The error associated with the Gaussian fit approximation, and thus the theoretical error associated with single molecule localization are described by:

$$\sigma = \sqrt{\frac{s^2}{N} + \frac{\frac{a^2}{12}}{N} + \frac{8\pi s^4 b^2}{a^2 N^2}} \qquad (1.1)$$

where s is the standard deviation of the Gaussian function approximating the true point spread function of the emitter (a characteristic of each dye/fluorescent probe), N is the number of photons collected, a is the pixel size, and b is the standard deviation of the background[21]. For most fluorescence-based experiments with low background, the number of photons collected N dominates localization precision, which highlights the main limitation of localization precision[20]. By properly balancing the need to collect sufficient photons from individual fluorophores with the need to image biological systems at relevant time scales, high temporal and spatial resolutions can be extracted for the dynamics of labeled biomolecules. Key to realizing this objective is careful selection of fluorescent labels, with considerations including probe size, biocompatibility, and brightness[22]. In some cases, extremely bright probes such as quantum dots can be quite

large and potentially perturbative to the function of target biomolecules. On the other hand, the small probes such as single organic dyes can be challenging to specifically and stoichiometrically conjugate to biomolecules. In some cases, small molecular organix dyes may not provide sufficient photostability for extremely long imaging times (on the order of minutes). Genetically encoded fluorescent protein tags are (arguably) biocompatible, but they are generally large and not exceptionally bright. Nevertheless, fluorescent proteins can be specifically and stoichiometrically fused to target proteins for live cell imaging. Despite challenges in probe selection and optimization, advances in probe development across all three classes described above have opened up greater opportunities to apply SMFM to biological systems. These advances include development of self-healing dye molecules[23], reduction in the size of quantum dot probes[24], and importantly for this study, the development of genetically encoded, bioorthogonal labeling schemes enabling specific and stoichiometric conjugation of small molecule probes to proteins of interest[25,26].

1.3 Single Molecule Imaging of DNA-Binding Proteins

The high temporal and spatial resolution afforded by SMFM makes it an ideal technique with which to investigate programmable nucleases, such as TALEs and the CRISPR/Cas9 system, and to elucidate their sequence search mechanism and binding dynamics[27,28]. The study of DNA binding proteins via single molecule imaging approaches has resulted in a large number of new insights into the processes of various polymerases[29,30], nucleases[31], DNA packaging proteins such as histones[32], and transcription factors[33]. These studies began in the early 1990s with Kabata and coworkers studying the interaction of fluorescently labeled RNA polymerase with

bacteriophage lambda genomic DNA[34]. This seminal study reported the direct observation of the real-time dynamics of a DNA-binding protein (DBP) in an energy-independent and thermally driven fashion along DNA. In the intervening years, single molecule fluorescence microscopy has been utilized in conjunction with new protein labeling strategies to extract molecular-level details and dynamics inaccessible by bulk techniques.

The body of work encompassing SMFM investigations of protein-DNA interactions can be broadly classified into three major categories: *in vitro* single molecule tracking, single molecule fluorescence resonance energy transfer (smFRET), and *in vivo* single molecule tracking. *In vitro* single molecule tracking can be used to directly observe the long-range diffusion and sequence search of DBPs along extended DNA substrates. This method has permitted the determination of physical constants for DBP-protein interaction such as 1-D diffusion coefficients[28] and other molecular details such as the outcomes of molecular collisions[35]. Prior to the application of SMFM to study DBP dynamics, there had yet to be compelling, direct observation that DBPs were capable of facilitated search processes. Since the initial study of RNAP by Kabata[34], these *in vitro* experiments have revealed that many DBPs tightly track the DNA helix during their search, rotating around the major groove in a helically-constrained one-dimensional sliding search for their target site[36]. Two-color tracking experiments have provided a means to directly observe the collisions between DBPs, highlighting the ability of several processivity factors to eject bound, stationary DBPs such as nucleosomes and transcription factors[35,37]. Single molecule FRET provides significant improvements in spatial resolution and captures dynamics on the sub-nanometer scale, though long range

events are difficult if not impossible to measure[38]. The unique working range of smFRET (~1-10 nm) has afforded biophysicists the ability to study individual helicase molecules unwinding DNA at relevant length scales, offering real-time insights into their step-wise behavior[39]. Finally, *in vivo* single molecule tracking has provided views of the binding lifetimes and search dynamics of transcription factors within cell nuclei, highlighting the effects of chromatin structure on how readily these DBPs are able to search DNA and locate target sites[40–43].

1.4 Transcription Activator-Like Effectors

Transcription activator-like effector (TALE) proteins are a class of DNA binding proteins naturally secreted by the *Xanthomonas* bacteria to aid in their infection of plant cells[44] (**Figure 1.2**). Early genetics studies identified a large number of repetitive elements, determined to correspond 34 or 35 amino acids, in the *avrBs3* gene. At the time, this gene was only known as part of the bacterial genome of a deleterious infection in pepper plants[45]. Following identification of the *avrBs3* repeats, several homologs were isolated and identified from rice bacterial infections, followed quickly by those in citrus cankers and tomato blight[46,47]. The avrBs4 homolog found in tomato infection was particularly intriguing due to high sequence homology to the *avrBs3* gene, including a conserved N-terminal domain[47]. The only significant difference in sequence was identified as a rearrangement of the 34 or 35 amino acid repeat elements. *In vitro* binding studies demonstrated that avrBs3 strongly interacted with DNA, confirming that these proteins, with their nuclear localization signals, were modulators of transcription[48]. Nevertheless, these results were perplexing because researchers could not identify similar repeat structures amongst the broad class of site-specific DNA-binding proteins.

The breakthrough in understanding TALE-DNA interactions came as a result of the realization that the size of the UPA box (region in pepper genome up-regulated by *avrBs3*) was roughly the same size as the number of repeat domains in *avrBs3*, leading to the assignment of 'one repeat to one base pair' and subsequent determination of the hypervariable repeat diresidues[49]. Within each of the highly conserved 34 to 35 amino acid repeats, there were distinct differences in the amino acid identity at positions 12 and 13. Identification of these repeat variable diresidues (RVDs) led to elucidation of the binding code for TALEs. Once the binding code was determined and TALEs were established as programmable DNA binding proteins, performed independently by two different labs[50,51], the connection to zinc finger nucleases was made and TALEs were quickly fused to the dimeric nuclease FokI[52] (**Figure 1.3**). Successful demonstration of TALEN activity was shown in yeast in 2010[52], followed by an explosion of applications in 2011 and 2012. These included genomic edits to mice[53], zebrafish[11], rice[15], and human stem cells[54]. Commercial gene editing player Cellectis, along with Life Technologies, quickly picked up the commercial rights to TALENs around this time and began developing a gene therapy platform based on TALEN-engineered 'off the shelf' T-cells[8,55]. Recently, at a 2015 meeting of the American Society of Hematology, Cellectis announced clinical trial results for the first patient treated via their TALEN-engineered CAR-T therapy[56,57]. Their results showed a positive response for the patient, a young child suffering from leukaemia, which has helped to improve public sentiment for human-based gene therapy.

Despite an enormous body of work on gene editing and gene expression applications of TALE(N)s, reports of off-target binding and nuclease activity provide

evidence that improvements to TALE design and engineering are needed[58]. In order for improvements in TALE binding specificity to be made, however, a complete picture of TALE target search/recognition and the factors influencing specificity is critically needed.

1.5 Sequence Search of DNA-Binding Proteins

Searching for and stably binding a 10 - 30 base pair target sequence within millions of base pairs of non-target DNA is a major task. Remarkably, a broad class of DNA-binding proteins (DBPs) successfully navigates this task, despite what sound like insurmountable odds. Seminal studies of the *lac* operon by Berg, Winters, and von Hippel reached the conclusion that the rate at which the *lac* repressor reached its target was faster than the rate predicted by simple three-dimensional free diffusion within the cellular interior[59]. Their work studying the *lac* repressor led to their proposal of a method of facilitated diffusion in which the operator reduces the dimensionality of its search by binding non-specifically to and either sliding or 1-D hopping along the DNA, effectively using it as a track (**Figure 1.4**). Numerous follow-up studies supported this model, including experiments that focused on the ionic strength dependence of association rates and computational (molecular dynamics) studies that support a one-dimensional sliding motion for bound *lac* operator and other DBPs[60,61].

Perhaps the most compelling evidence supporting a facilitated diffusion mechanism for DBP target sequence search has emerged from single molecule experiments. Herein, single molecule studies of DBP search dynamics have provided the direct observation of 1-D diffusion of proteins along DNA substrates[28,33]. Such studies have focused on a broad class of DBPs, including helicases[62], transcription factors[33],

repair proteins[63], and a viral proteinase[64]. The overarching theme in these studies is been observation of 1-D sliding and/or hopping, such that experimental results are explained in the context of Berg and von Hippel's frameworks. Nevertheless, additional studies, including some incorporating single molecule imaging, have argued that in many cases these observations are made under contrived conditions and that DBPs could locate their targets primarily via 3-D diffusion *in vivo*[30]. Indeed, the long, extended, and bare DNA substrates typically used in single molecule experiments are quite different than packaged DNA encountered in mammalian cell nuclei. This finding is protein specific, however, as arguments against facilitated diffusion are only truly applicable to DBPs at very high copy number within a cell, or without the necessary structural characteristics to support a one-dimensional search. The contributions of three-dimensional diffusion and one-dimensional search thus lie on opposite ends of a spectrum for each protein; that protein's structure and copy number, along with the conformation of chromatin, determines its location on the spectrum.

A major obstacle for DBPs undergoing search is the innate similarity of non-specific DNA to their target sequence. Given that DNA consists of only 4 different nucleotides, it becomes difficult to imagine how a protein can rapidly scan the vast amount of non-specific sites while also exhibiting stable binding to a target site (timescales of minutes and longer). In order for rapid search along non-specific DNA to occur, the protein must experience a relatively smooth energy landscape with small barriers between each discrete step along its path. This smooth and homogenous energy landscape (on the order of 1-2 k_BT)[65], however, is in sharp contrast to the energy landscape the protein must experience for stable, sequence specific binding. Once bound

to a target site, the protein is effectively trapped, and thus requires an extremely rough landscape, with a variance greater than 5 k_BT[65]. These two sets of requirements, one for rapid search and one for stable target binding, are at odds with one another and thus require a model more complex than a rigid, unchanging protein structure to reconcile.

The search speed-stability paradox refers to the seemingly disparate requirements for DBPs and related energy landscapes during non-specific search versus specifically bound states[66]. Numerous models have been proposed to resolve this paradox, including those utilizing multiple conformations of the searching protein and/or multiple protein subunits[67]. Direct experimental evidence to support the search speed-stability paradox has been limited, as it requires both high temporal and spatial resolutions, as well as precise control over the DNA substrate. A single molecule study of tumor suppressor p53 by Tafizi and coworkers provided direct observation of a search process utilizing a two-state model for sequence search via selective delegation of tasks (search and sequence binding) to separate subunits of p53[68]. Additionally, a combined *in situ* NMR and coarse grained molecular dynamics study provided evidence for both a scanning/searching mode and a binding mode for the zinc finger protein Egr-1[69].

1.6 Project Overview

Despite progress in developing new applications for TALEN gene editing, we still do not fully understand the DNA search process of TALE proteins. In order to understand how to rationally design TALEs that exhibit decreased or entirely absent off-target binding events and increased target affinity, it is critical to understand the processes that lead to successful target binding or unwanted off-target binding. A deeper understanding of the TALE search process and how it is influenced by TALEs' unique structure can

potentially provide details about the behavior of the small, but important classes of structurally similar proteins, including tetratricopeptide (TPR) proteins[70] and mitochondrial transcription termination factor (mTERF)[71].

In this thesis, we combine single molecule methods with traditional biochemical binding measurements to study TALE target search and binding. In Chapter 2, we describe the development of a single molecule imaging assay for direct visualization of TALE search dynamics on dual-tethered DNA substrates, and we report the ability of TALEs to scan DNA in search of their target in a one-dimensional search. We also determine the relative contributions of the TALE N-terminal region (NTR) and central repeat domain (CRD), and propose a two-state model for TALE sequence search. In Chapter 3, we utilize single molecule techniques to further study the physical underpinnings of TALE search, and our results show that TALEs adopt a wrapped conformation around DNA during search. In this way, TALEs fully encircle DNA but do not remain tightly associated with the DNA major groove during non-specific search, such that TALEs follow a rotationally de-coupled trajectory as they scan DNA. Our results further suggest TALE search cannot be strictly classified as either sliding or hopping, despite the classic use of this binary classification scheme. In Chapter 4, we describe the effects of ionic strength and cation identity on TALE binding, and show that certain divalent cations significantly enhance TALE specificity. We utilize both bulk fluorescence anisotropy measurements and single molecule imaging to show that the presence of certain divalent cations significantly decreases the non-specific binding affinity of TALEs to a greater extent than monovalent cations, thereby increasing the apparent specificity of TALEs for their target site. Furthermore, we utilize our single

molecule imaging assay along with long DNA substrates containing arrays of specific TALE binding sites to directly visualize the binding and localization of TALEs to their targets following 1-D search. Finally, in Chapter 5, we propose how single molecule methods, including single molecule FRET and *in vivo* single molecule tracking, can further illuminate important yet poorly described aspects of TALE-DNA binding.

1.7 Figures

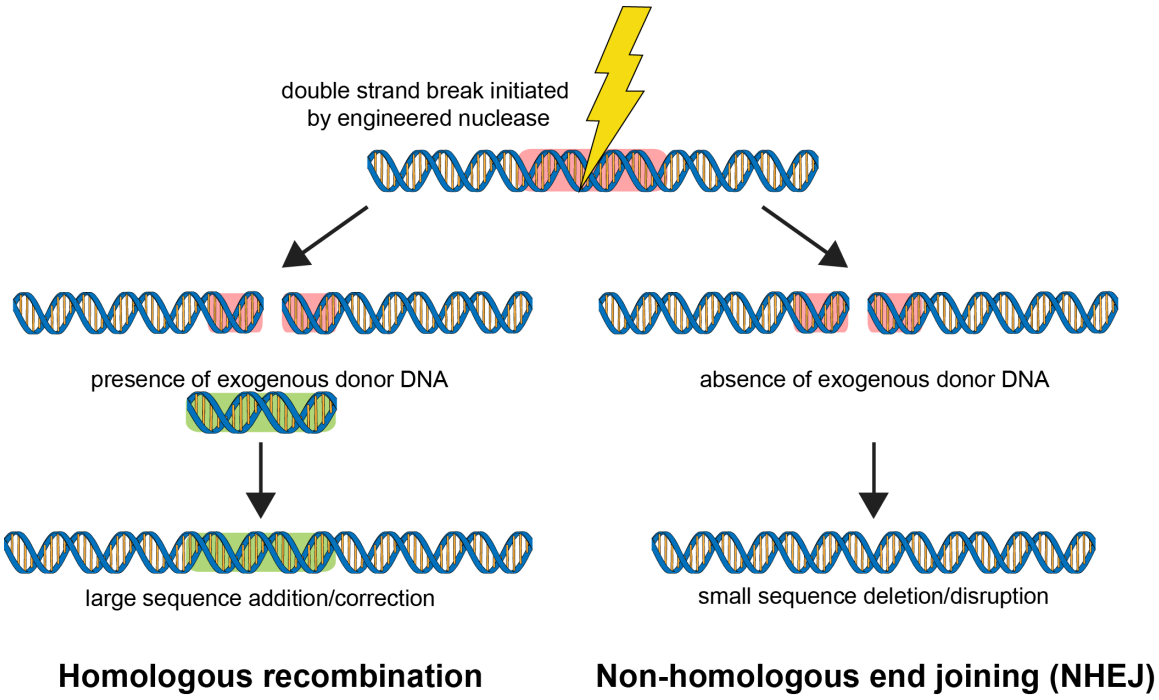

Figure 1.1 Overview of two pathways that lead to genomic edits mediated by programmable nucleases. Upon introduction of a double stranded break at a targeted site via a programmable nuclease, natural cellular machinery can follow one of the many processes grouped under homologous recombination, or the more straightforward non-homologous end joining (NHEJ) pathway. Briefly, in the case of NHEJ, the damaged ends are removed, re-extended, and then ligated. This process is performed in the absence of template DNA to serve as a replacement, and the resulting process is highly error prone. In contrast, homologous recombination proceeds by double stranded breakage, but here the presence of the donor DNA allows for strand invasion to take place and D-loop formation, which provides polymerases with a template strand[72].

Figure 1.2 Crystal structure of TALE protein PthXo1 bound to its DNA target. In this rendering, the TALE CRD is colored green, while the TALE NTR is colored yellow[73,74]. The TALE CTR is not shown.

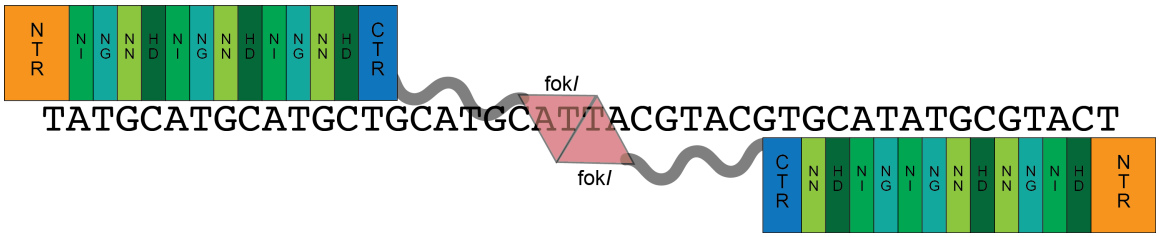

Figure 1.3 Schematic representation of hypothetical TALEN system. The fok*I* nuclease is fused to the TALE C-terminal region (CTR) via a flexible linker. Given that fok*I* is a dimeric nuclease, an additional level of specificity is imparted on the TALEN system.

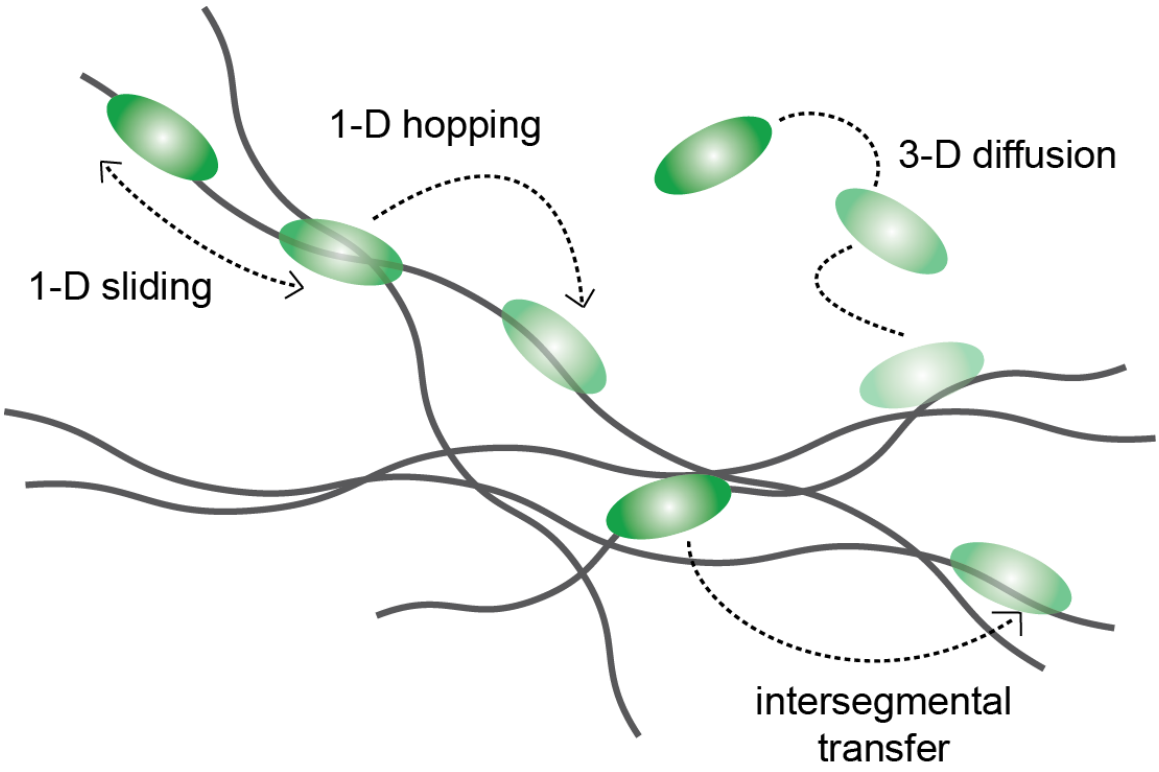

Figure 1.4 Possible search mechanisms for DNA binding proteins. These mechanisms are proposed by Winter *et al* in their seminal work studying the *lac* repressor[59].

1.8 References

1. Doudna, J. A. Genomic Engineering and the Future of Medicine. *JAMA* **313,** 791–2 (2015).

2. Voytas, D. F. & Gao, C. Precision genome engineering and agriculture: opportunities and regulatory challenges. *PLoS Biol.* **12,** e1001877 (2014).

3. Ramlee, M. K., Yan, T., Cheung, A. M. S., Chuah, C. T. H. & Li, S. High-throughput genotyping of CRISPR/Cas9-mediated mutants using fluorescent PCR-capillary gel electrophoresis. *Sci. Rep.* **5,** 15587 (2015).

4. Gaj, T., Gersbach, C. A. & Barbas, C. F. ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31,** 397–405 (2013).

5. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome

editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11,** 636–646 (2010).

6. Holt, N. *et al.* Human hematopoietic stem/progenitor cells modified by zinc-finger nucleases targeted to CCR5 control HIV-1 in vivo. *Nat. Biotechnol.* **28,** 839–847 (2010).

7. Urnov, F. D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435,** 646–651 (2005).

8. Torikai, H. *et al.* A foundation for universal T-cell based immunotherapy: T cells engineered to express a CD19-specific chimeric-antigen-receptor and eliminate expression of endogenous TCR. *Blood* **119,** 5697–5705 (2012).

9. Ding, Q. *et al.* A TALEN Genome-Editing System for Generating Human Stem Cell-Based Disease Models. *Cell Stem Cell* **12,** 238–251 (2013).

10. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **39,** e82 (2011).

11. Bedell, V. M. *et al.* In vivo genome editing using a high-efficiency TALEN system. *Nature* **491,** 114–8 (2012).

12. Jinek, M. *et al.* A Programmable Dual-RNA– Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337,** 816–821 (2012).

13. Jiang, W. & Marraffini, L. a. CRISPR-Cas: New Tools for Genetic Manipulations from Bacterial Immunity Systems. *Annu. Rev. Microbiol.* **69,** 150724172101001 (2015).

14. Kim, H. & Kim, J.-S. A guide to genome engineering with programmable nucleases. *Nat. Rev. Genet.* **15,** 321–34 (2014).

15. Li, T., Liu, B., Spalding, M. H., Weeks, D. P. & Yang, B. High-efficiency TALEN-based gene editing produces disease-resistant rice. *Nat. Biotechnol.* **30,** 390–392 (2012).

16. Tan, W. *et al.* Efficient nonmeiotic allele introgression in livestock using custom endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 16526–31 (2013).

17. Sun, N., Liang, J., Abil, Z. & Zhao, H. Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Mol. Biosyst.* **8,** 1255 (2012).

18. Bosley, K. S. *et al.* CRISPR germline engineering—the community speaks. *Nat. Biotechnol.* **33,** 478–486 (2015).

19. Liang, P. *et al.* CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes. *Protein Cell* **6,** 363–372 (2015).

20. Yildiz, A. *et al.* Myosin V walks hand-over-hand: single fluorophore imaging with 1.5-nm localization. *Science* **300,** 2061–5 (2003).

21. Thompson, R. E., Larson, D. R. & Webb, W. W. Precise Nanometer Localization Analysis for Individual Fluorescent Probes. *Biophys. J.* **82,** 2775–2783 (2002).

22. Agrawal, U., Reilly, D. T. & Schroeder, C. M. Zooming in on biological processes with fluorescence nanoscopy. *Curr. Opin. Biotechnol.* **24,** 646–653 (2013).

23. Reilly, D. T., Kim, S. H., Katzenellenbogen, J. A. & Schroeder, C. M. Fluorescent Nanoconjugate Derivatives with Enhanced Photostability for Single Molecule Imaging. *Anal. Chem.* **87,** 11048–11057 (2015).

24. Cai, E. *et al.* Stable Small Quantum Dots for Synaptic Receptor Tracking on Live Neurons. *Angew. Chemie Int. Ed.* **53,** n/a–n/a (2014).

25. Stagge, F., Mitronova, G. Y., Belov, V. N., Wurm, C. A. & Jakobs, S. Snap-, CLIP- and Halo-Tag Labelling of Budding Yeast Cells. *PLoS One* **8,** 1–9 (2013).

26. Shi, X. *et al.* Quantitative fluorescence labeling of aldehyde-tagged proteins for single-molecule imaging. *Nat. Methods* **9,** 499–503 (2012).

27. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67 (2014).

28. Cuculis, L., Abil, Z., Zhao, H. & Schroeder, C. M. Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nat. Commun.* **6,** 7277 (2015).

29. Harada, Y. *et al.* Single-molecule imaging of RNA polymerase-DNA interactions in real time. *Biophys. J.* **76,** 709–715 (1999).

30. Wang, F. *et al.* The promoter-search mechanism of Escherichia coli RNA polymerase is dominated by three-dimensional diffusion. *Nat. Struct. Mol. Biol.* **20,** 174–81 (2013).

31. Dikić, J. *et al.* The rotation-coupled sliding of EcoRV. *Nucleic Acids Res.* **40,** 4064–4070 (2012).

32. Visnapuu, M.-L. & Greene, E. C. Single-molecule imaging of DNA curtains reveals intrinsic energy landscapes for nucleosome deposition. *Nat. Struct. Mol. Biol.* **16,** 1056–62 (2009).

33. Wang, Y. M., Austin, R. H. & Cox, E. C. Single molecule measurements of repressor protein 1D diffusion on DNA. *Phys. Rev. Lett.* **97,** 1–4 (2006).

34. Kabata, H. *et al.* Visualization of single molecules of RNA polymerase sliding

along DNA. *Science* **262,** 1561–1563 (1993).

35. Finkelstein, I. J., Visnapuu, M.-L. & Greene, E. C. Single-molecule imaging reveals mechanisms of protein disruption by a DNA translocase. *Nature* **468,** 983–7 (2010).

36. Blainey, P. C. *et al.* Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.* **16,** 1224–1229 (2009).

37. Lee, J., Finkelstein, I. J., Arciszewska, L. K., Sherratt, D. J. & Greene, E. C. Single-Molecule Imaging of FtsK Translocation Reveals Mechanistic Features of Protein-Protein Collisions on DNA. *Mol. Cell* **54,** 832–843 (2014).

38. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat. Methods* **5,** 507–516 (2008).

39. Syed, S., Pandey, M., Patel, S. S. & Ha, T. Single-Molecule Fluorescence Reveals the Unwinding Stepping Mechanism of Replicative Helicase. *Cell Rep.* **6,** 1037–1045 (2014).

40. Elf, J., Li, G.-W. & Xie, X. S. Probing Transcription Factor Dynamics at the Single-Molecule Level in a Living Cell. *Science* **316,** 1191–1194 (2007).

41. Morisaki, T., Müller, W. G., Golob, N., Mazza, D. & McNally, J. G. Single-molecule analysis of transcription factor binding at transcription sites in live cells. *Nat. Commun.* **5,** 4456 (2014).

42. Gebhardt, J. C. M. *et al.* Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. *Nat. Methods* **10,** 421–6 (2013).

43. Knight, S. C. *et al.* Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350,** 823–6 (2015).

44. Mussolino, C. & Cathomen, T. TALE nucleases: Tailored genome engineering made easy. *Curr. Opin. Biotechnol.* **23,** 644–650 (2012).

45. Bonas, U., Stall, R. E. & Staskawicz, B. Genetic and structural characterization of the avirulence gene avrBs3 from Xanthomonas campestris pv. vesicatoria. *Mol. Gen. Genet.* **218,** 127–136 (1989).

46. Ishihara, H., Ponciano, G., Leach, J. E. & Tsuyumu, S. Functional analysis of the 3′ end of avrBs3/pthA genes from two Xanthomonas species. *Physiol. Mol. Plant Pathol.* **63,** 329–338 (2003).

47. Ballvora, a *et al.* Genetic mapping and functional analysis of the tomato Bs4 locus governing recognition of the Xanthomonas campestris pv. vesicatoria AvrBs4 protein. *Mol. Plant. Microbe. Interact.* **14,** 629–638 (2001).

48. Szurek, B., Marois, E., Bonas, U. & Van Ackerveken, G. Den. Eukaryotic features of the Xanthomonas type III effector AvrBs3: Protein domains involved in transcriptional activation and the interaction with nuclear import receptors from pepper. *Plant J.* **26,** 523–534 (2001).

49. Kay, S., Hahn, S., Marois, E., Hause, G. & Bonas, U. A Bacterial Effector Acts as a Plant Transcription Factor and Induces a Cell Size Regulator. *Science* **318,** 648–651 (2007).

50. Moscou, M. J. & Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326,** 1501 (2009).

51. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326,** 1509–1512 (2009).

52. Christian, M. *et al.* Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186,** 757–61 (2010).

53. Sung, Y. H. *et al.* Knockout mice created by TALEN-mediated gene targeting. *Nat. Biotechnol.* **31,** 23–24 (2013).

54. Ma, N. *et al.* Transcription activator-like effector nuclease (TALEN)-mediated Gene correction in integration-free β-Thalassemia induced pluripotent stem cells. *J. Biol. Chem.* **288,** 34671–34679 (2013).

55. Poirot, L. *et al.* T-Cell Engineering For 'off-The-shelf' Adoptive Immunotherapy. *Blood* **122,** 1661 (2013).

56. Qasim, W. *et al.* First clinical application of TALEN engineered universal CAR19 T cells in B-ALL. *Blood* **126,** 2046 (2015).

57. Frederickson, R. M. A New Era of Innovation for CAR T-cell Therapy. *Mol. Ther.* **23,** 1795–1796 (2015).

58. Stella, S. & Montoya, G. The genome editing revolution: A CRISPR-Cas TALE off-target story. *Insid. Cell* n/a–n/a (2015). doi:10.1002/icl3.1038

59. Berg, O. G., Winter, R. B. & von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* **20,** 6929–6948 (1981).

60. Givaty, O. & Levy, Y. Protein sliding along DNA: dynamics and structural characterization. *J. Mol. Biol.* **385,** 1087–97 (2009).

61. Winter, R. B., Berg, O. G. & von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor--operator interaction: kinetic measurements and conclusions. *Biochemistry* **20,** 6961–77 (1981).

62. Schwarz, F. W. *et al.* The Helicase-Like Domains of Type III Restriction Enzymes Trigger Long-Range Diffusion Along DNA. *Science* **340,** 353–356 (2013).

63. Gorman, J. *et al.* Dynamic Basis for One-Dimensional DNA Scanning by the Mismatch Repair Complex Msh2-Msh6. *Mol. Cell* **28,** 359–370 (2007).

64. Blainey, P. C. *et al.* Regulation of a viral proteinase by a peptide and DNA in one-dimensional space IV: Viral proteinase slides along dna to locate and process its substrates. *J. Biol. Chem.* **288,** 2092–2102 (2013).

65. Slutsky, M. & Mirny, L. a. Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. *Biophys. J.* **87,** 4021–4035 (2004).

66. Mirny, L. *et al.* How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. A Math. Theor.* **42,** 434013 (2009).

67. Zhou, H.-X. Rapid search for specific sites on DNA through conformational switch of nonspecifically bound proteins. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 8651–6 (2011).

68. Tafvizi, A., Huang, F., Fersht, A. R., Mirny, L. A. & van Oijen, A. M. A single-molecule characterization of p53 search on DNA. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 563–568 (2011).

69. Zandarashvili, L. *et al.* Asymmetrical roles of zinc fingers in dynamic DNA-scanning process by the inducible transcription factor Egr-1. *Proc. Natl. Acad. Sci.* **109,** E1724–E1732 (2012).

70. Filipovska, A. & Rackham, O. Modular recognition of nucleic acids by PUF, TALE and PPR proteins. *Mol. Biosyst.* **8,** 699 (2012).

71. Jiménez-Menéndez, N. *et al.* Human mitochondrial mTERF wraps around DNA through a left-handed superhelical tandem repeat. *Nat. Struct. Mol. Biol.* **17,** 891–893 (2010).

72. MAO, Z., BOZZELLA, M., SELUANOV, A. & GORBUNOVA, V. Comparison of nonhomologous end joining and homologous recombination in human cells. *DNA Repair (Amst).* **7,** 1765–1771 (2008).

73. Pettersen, E. F. *et al.* UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612 (2004).

74. Mak, A. N.-S., Bradley, P., Cernadas, R. A., Bogdanove, A. J. & Stoddard, B. L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335,** 716–9 (2012).

Chapter 2: Single Molecule Visualization of TALE 1-D Diffusion on DNA*

2.1 Introduction

Despite the pervasive use of TALE(N)s in engineered nuclease systems, the sequence search mechanism of TALE proteins is not well understood. Efficient sequence search and highly accurate and precise binding is essential for TALEN-based therapies. Moreover, a fundamental understanding of TALE search is needed to enable forward engineering and rational design of TALEN systems with improved function and safety profiles. For these reasons, we critically need a complete understanding of the molecular-level search process for TALEs.

The structure of TALEs was determined by two independent labs in 2012[1,2]. Structural data conformed initial speculation based on research published in 2009 suggesting that TALE structure was highly unique amongst the broad class of DNA binding proteins[3,4]. Indeed, the superhelical structure of TALEs has few close structural analogs; the mitochondrial transcription factors mTERF are perhaps the closest in terms of structure[5].

Due to the unique protein structure, the TALE sequence search mechanism could not be readily determined. Nevertheless, static crystal structures for TALEs provided some insights into the nature of TALEs proteins likely conformations. A crystal structure of DNA-free TALE showed an extended helical structure, out of phase with the pitch of natural B-form DNA[2]. Target DNA-bound TALE structures showed a compressed helical structure that tracked the DNA helix tightly, without exerting any apparent distortion on

*Portions of this chapter were previously published in: L. Cuculis, Z. Abil, H. Zhao, C. M. Schroeder, "Direct Observation of TALE Protein Dynamics Reveals a Two-State Search Mechanism", *Nature Communications*, 6, (2015).

the DNA structure[1,2]. Recent molecular dynamics studies showed that there is an inherent plasticity in the TALE structure, providing evidence that the DNA-free and target DNA-bound TALE structures most likely interconvert between an extended helical and compressed helical shape[6]. Despite these insights, however, no structure exists of TALE proteins bound to non-specific DNA. During non-specific search, it is likely that TALEs adopt rapidly changing geometries that interconvert between multiple conformations. Furthermore, it is likely that multiple protein conformations could exist during the overall process beginning with unbound TALEs in solution and ending with TALEs binding at their target site.

Previous studies and reviews of TALE structure and binding have suggested that the process might follow a 1-D sliding mechanism during search, closely tracking the DNA backbone, given its superhelical structure[7,8]. Given the requirements needed to satisfy the search speed-stability paradox (discussed in Chapter 1), however, a tight sliding mechanism for TALEs is difficult to reconcile[9]. If TALEs maintained a conformation bound to non-specific DNA that was very similar to the specifically bound conformation, then the difference in energy landscape they would experience going from random to target sequences would be too minimal. DNA, however, is too chemically homogeneous for this to be the case. A recent computational study on the energetics contributing to TALE specificity found that TALE specificity arises primarily due to negative discrimination by the RVDs, and these results support this reasoning[10]. These findings suggest that it is unlikely that TALEs follow a sliding mechanism in which proteins maintain their near-target bound conformation during search. TALEs are found to bind their target because it is the 'least bad' binding sequence, which means that a

static conformation homologous to the specific binding conformation would expose TALEs to a rough energy landscape that would not support an effective search process. Given that the vast majority of TALE binding energy is conferred by non-specific electrostatic interactions with DNA, a search process in which TALE conformation did not change between 'search' and 'bind' states would be prohibitively slow. Overall, this creates a conundrum. On the one hand, TALEs recognize their target sites via RVDs finding the 'least poor' match for binding; this requires the RVDs to sample the local DNA sequence in order to determine if a cognate match has been located. On the other hand, however, a dominant amount of binding energy being conferred via non-specific electrostatic interactions makes a tightly associated TALE an incredibly slow searching entity. For these reasons, molecular-level insights are critically needed to study the real-time process of non-specific TALEs search along DNA.

In this work, we utilize SMFM in order to directly observe interactions of individual, fluorescently labeled TALE proteins with long, dual-tethered DNA substrates[11]. We characterize the diffusive behavior of TALEs along DNA substrates by determining their apparent 1-D diffusion coefficients via a covariance-based estimator[12] and use these values to compare different TALE variants under an array of different conditions. Taken together, insights from these experiments suggest a model for TALE sequence search in which TALEs utilize two modes: an open, extended helical structure during rapid searching events, and a compressed helical structure when checking local sequences for the correct binding target.

2.2 Materials and Methods

2.2.1 Preparation of DNA constructs

Plasmids used for the dual-tethered DNA substrate were purified using the Plasmid Midi Kit (Qiagen) and linearized with SnaBI enzyme (NEB) for 3 hours, followed by further purification using phenol-chloroform extraction and ethanol precipitation. The purified plasmid was subjected to the 3'-5' exonuclease activity of T4 DNA polymerase (NEB) to create the 5' overhangs at both ends of the linearized DNA. 10 μg DNA was treated with 5 units of T4 DNA polymerase in NEB buffer 2 supplemented with BSA for 1 hour at 25 °C. The reaction was stopped with 1 μl 20 mM dCTP and the enzyme was heat deactivated at 75 °C for 20 minutes. The exposed 5′ overhangs were used to sequence-specifically anneal 3′-biotinylated oligonucleotides. First, the oligonucleotide with the sequence cagcagttcaacctgttgatagtac/3BioTEG/ (IDT) was annealed in 50x molar excess to the substrate DNA by heating the mixture at 90 °C for 5 minutes and gradually cooling to 4 °C. The second oligonucleotide with the sequence tacgtgaaacatgagagcttagtac/3BioTEG/ was annealed subsequently in the microfluidic flow chamber, as described below.

2.2.2 Preparation of TALE constructs

Cloning of tSCA21.5 and Naldt-tSCA21.5. The gene encoding for the untagged TALE protein (tSCA21.5) was assembled using the Golden Gate cloning method (Addgene TALEN Kit #1000000024, as described in prior work[13]) into a specifically engineered destination vector, pET28-GG-TALE. The destination vector contained an N-terminal His-tag and flanking N- (208 aa) and C- (63 aa) terminal regions of the TALE as well as the BsmBI sites corresponding to the kit BsmBI sites (**Figure 2.1**). For fluorescent

tagging of the TALE, we modified the original plasmid pET-tSCA21.5 with an oligonucleotide insert encoding an N-terminal LCPTSR hexapeptide (aldehyde tag[14,15]) upstream of the His-tag. To this end, the plasmid was amplified in fragments containing the insert, and assembled using the Gibson Assembly Kit (NEB).

Cloning Naldt-tSCA15.5 and Naldt-tSCA11.5. In order to construct TAL effectors shortened to the first 15.5 and 11.5 repeats, we first engineered a destination vector pET28-Naldt-GG-TALE, which contains the aldehyde tag upstream of the His-tag, using the Gibson Assembly Master Mix (NEB). Naldt-tSCA15.5 and Nald-tSCA11.5 were assembled using the Golden Gate cloning method[13] into the destination vector pET-NaldT-GG-TALE.

Cloning of Naldt-NTR. To create the NTR truncation mutant, we digested pET-Naldt-tSCA21.5 with SacII and HindIII, and inserted the sequence: CAAAGCGTGGTGGCGTGACCGCGGTGGAAGCGGTCCATGCCTGGCGTAATGC GTTGACGGGCGCCCCCCTGAACTAAGTCAGATAACCGGATACAGACAAGCTT GCGGCCGCACTCGAGCACCAC

synthesized as a gBlock Gene Fragment (IDT), using the Gibson Assembly Master Mix (NEB).

Cloning of Naldt-CRD-CTR. DNA oligonucleotide primers (IDT) containing BsmBI sites corresponding to BsmBI sites from previous destination vectors were used to PCR-amplify the backbone of the pET-Naldt-GG-TALE, excluding the NTR-region. The CRD containing 21.5 repeats was assembled into this PCR-amplified product using the Golden Gate method as previously.

2.2.3 Protein purification and labeling

Protein expression. BL21 (DE3) electrocompetent *E. coli* cells were co-transformed with plasmids encoding TAL effector constructs and the pBAD-FGE plasmid (generous gift of Dr. Taekjip Ha, University of Illinois at Urbana-Champaign). A single colony was grown in 5 ml LB supplemented with 100 μg/ml ampicillin and 25 μg/ml kanamycin as a seeding culture until saturation, and subsequently in 500 ml of Terrific Broth at 37 °C and 250 revolutions per minute (RPM) with the corresponding antibiotics until OD_{600} of 0.3-0.4. FGE expression was induced with 0.2 % L-arabinose (Sigma) and the culture was grown further at 37 °C 250 RPM until it reached OD_{600} 0.7-0.8. TAL effector expression was induced with 0.4 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) (Sigma). The proteins were expressed overnight at 16 °C and 250 RPM.

Protein purification. Cell cultures were centrifuged at 4,000 RPM at 4 °C for 15 minutes and the pellets re-suspended in 10-20 ml of lysis buffer (25 mM Tris-HCl (Fisher) pH 7.5, 300 mM NaCl (Fisher), 0.5 % Triton X-100 (Sigma), 5 % glycerol (Sigma), 4 U/mL DNase I (NEB), 0.3 mM phenylmethanesulfonylfluoride (Sigma), 1 mg/mL lysozyme (Sigma)). The cells were lysed by sonication for 20 minutes total with alternating 5 seconds of pulse and 5 seconds of rest, and cell debris were centrifuged at 13,000 RPM for 20 min at 4 °C. The His-tagged TAL effectors were purified using AKTA pure chromatography system (GE Healthcare) with a 1 mL HisTrap column (GE Healthcare). The cleared lysate was loaded on the column, washed with 20 mM Tris-HCl pH 7.5, 500 mM NaCl, and 20 mM imidazole (Sigma), and eluted with 20 mM Tris-HCl pH 7.5, 500 mM NaCl, and 250 mM imidazole. The purified protein was dialysed in 2 L 50 mM phosphate buffer pH 7-8.4, (depending on the predicted protein isoelectric point), 500

mM NaCl at 4 °C overnight. When necessary, the protein was further purified using 16/600 200 pg gel filtration column (GE Healthcare) using 50 mM phosphate buffer pH 7-8.4 (depending on the predicted protein isoelectric point), 500 mM NaCl.

Protein labeling with Cy3. The purified (>90 % purity by SDS-PAGE) TAL effector proteins were buffer exchanged using Amicon Ultra-0.5 ml centrifugal units (EMD Millipore) and concentrated to 100-300 μM in 30 μL of labeling buffer (250 mM potassium phosphate pH 6, 500 mM KCl (Fisher), 5 mM dithiothreitol (Roche)). The concentrated protein solutions were used to re-suspend 1 mg Cy3-hydrazide (GE healthcare) and labeled for 24 hours at room temperature in the dark. Labeled proteins were diluted with 400 μL fluorescence anisotropy buffer (20 mM Tris-HCl pH 7.5 100 mM Nacl 0.5 mM ethylenediaminetetraacetic acid (EDTA) (Fisher)) and purified from unreacted Cy3 by two consecutive passages through Micro Bio-spin 6 columns (Bio-Rad) following manufacturer's instructions.

2.2.4 Fluorescence anisotropy

A 29 bp oligonucleotide containing the 23 nt TALE binding site was labeled at the 5′ end with 6-FAM (fluorescein) (IDT) and annealed with its reverse-complementing oligonucleotide in the annealing buffer (10 mM Tris/Tris-HCl 1 mM EDTA pH 8, 50 mM NaCl) by heating up the mixture at 90 °C for 2 minutes and gradually cooling to 4 °C. A control double stranded oligonucleotide was prepared similarly, except the sequence was randomized to contain no binding site.

Mixtures of 1 nM ds oligonucleotide and various concentrations of proteins were prepared in the fluorescence anisotropy buffer, and 200 μL samples were assayed in black 96-well plates (Corning), in duplicates. Fluorescence polarization measurements

were taken on Infinite 200 Pro microplate reader (Tecan) using excitation and emission wavelengths of 485 nm and 535 nm, respectively. The fluorescence polarization values were converted to fluorescence anisotropy values using Equation 2.1, where A is anisotropy and P is polarization. The K_D value was calculated by curve fitting on Origin 8.5 using Equation 2.2, where A is observed anisotropy value, A_f is anisotropy of free DNA, A_b is anisotropy of bound DNA, L_T is total ligand (DNA) concentration, and R_T is total receptor (protein) concentration.

$$A = \frac{2P}{3-P} \tag{2.1}$$

$$A = A_f + (A_b - A_f) * \frac{(L_T + K_D + R_T) - \sqrt{(L_T + K_D + R_T)^2 - 4L_T R_T}}{2L_T} \tag{2.2}$$

2.2.5 Flow cell preparation

Mili-fluidic flow cells were constructed by sandwiching two pieces of double-sided tape between a pre-drilled quartz microscope slide and glass coverslip to form a flow channel (approximately 50 mm long by 4 mm wide by 0.05 mm high). Prior to assembly of the flow cell, coverslips were functionalized with PEG/PEG-biotin and Neutravidin for surface attachment of DNA and reduction of nonspecific protein adsorption[16]. Polyethylene tubing (PE60, Solomon Scientific) was affixed to the ports (0.048" OD) drilled in each end of the flow cell in order to facilitate rapid exchange of buffer solutions and allow for stretching of tethered DNA.

2.2.6 Surface attachment of double-tethered DNA templates

Assembled flow cells were first incubated with a blocking solution (50 mM MOPS, 10-110 mM KCl, 0.1 mM EDTA, 5% glycerol, 0.3 mg/mL BSA, pH 8.1) for 10 minutes and

then 5 pM biotin-functionalized DNA for 45 minutes. Unbound DNA was washed with blocking solution and double tethered DNA was subsequently formed by flowing 100 nM biotinylated primer, complementary to the single stranded overhangs previously generated on the long, single tethered DNA, at a rate of 100 uL/min in the presence of 100 μM chloroquine (Sigma Aldrich). The presence of chloroquine in the primer solution allows for the DNA to be extended to ~85% of its contour length immediately prior to formation of the second surface tether, reducing substrate fluctuations during single molecule imaging[17]. Chloroquine is subsequently removed by washing the flow cell with blocking solution containing 40 mM $MgCl_2$ and 500 mM NaCl for 5 minutes.

2.2.7 Single molecule imaging

Single molecule imaging experiments were carried out on an inverted microscope (Nikon IX70) equipped for total internal reflection fluorescence (TIRF) coupled to an EMCCD camera (Andor iXon Ultra 897). Cy3 labeled proteins were illuminated using a 532 nm diode-pumped solid state (DPSS) laser (CrystaLaser) and SYTOX Green and YOYO-1 labeled DNA was excited using a 488 nm DPSS laser (SpectraPhysics Excelsior) while Qdot705 labeled TALEs were excited using a 637 nm DPSS laser (Coherent). Image sequences were acquired at a rate of 30-50 Hz. Labeled TALE proteins were added at concentrations typically ranging from 25 to 100 pM, and in the case of the NTD, in the presence of 1 to 2 nM unlabeled constructs. For quantum dot-conjugated TALEs, 100-200 pM labeled TALEs were incubated in a prepared sample chamber, followed by a rinse with 150 mM KCl and introduction of 2 nM YOYO-1 (Invitrogen) in imaging buffer. 5 uM free biotin (Sigma) was also added during Qdot experiments to prevent nonspecific binding to the chamber surface. A reducing agent (7 mM β-

mercaptoethanol, Sigma Aldrich) and oxygen scavenging system (glucose oxidase and bovine liver catalse) along with 1% v/v glucose were added to the buffer, along with the protein.

2.2.8 Data analysis

Images are recorded as TIF stacks using the Andor Solis software. Regions of interest containing single diffusing TALE proteins are isolated using ImageJ, and the centroid locations of single proteins are determined using RapidStorm fitting software[18]. Single molecule trajectories were then further analyzed using custom MATLAB scripts.

2.2.9 Localizing single TALE proteins via fluorescence microscopy

Error in localizing single TALE proteins arises from two main sources: (1) uncertainty due to the finite number of photons collected within each frame during image acquisition, which amounts to localization precision, and (2) thermal fluctuations in the DNA template parallel to the direction of TALE diffusion, which amounts to localization accuracy. For these experiments, we use an acquisition time of 20-33 ms per frame, during which we collect ~70-100 photons. We note that the imaging assay also shows an extremely low fluorescence background. Based on these experimental conditions, we estimate the localization precision of single TALE proteins to be ~25 nm[19]. Regarding the second source of error, we estimate the magnitude of DNA fluctuations parallel to the direction of TALE diffusion to be ~40 nm, which is based on the equipartition theorem and a Taylor series expansion of the DNA stretching force under experimental conditions. This calculation relies on the average extension of DNA templates in our assay (~90% of full contour length) and the Marko-Siggia force-extension relation[20]. The summation of both sources of error is ~65 nm.

2.2.10 Covariance-based estimator (CVE) for 1-D diffusion coefficients

In order to minimize compounding localization errors of each frame in a single molecule trajectory, which is inherently present when using mean squared displacement (MSD) determination of diffusion coefficients, we utilized a covariance-based estimator to determine 1-D diffusion coefficients[12]. Using an average displacement Δx_n of a TALE protein during a single acquisition period (20 to 33 ms per frame) and the localization error $<\delta z^2>$ as described above, 1-D diffusion coefficients were determined according to:

$$D = \frac{\overline{(\Delta x_n)^2} - 2\langle \delta z^2 \rangle}{2(1-2R)\Delta t} \tag{2.3}$$

Where $R=1/6$ is a motion blur coefficient accounting for the finite exposure time of the camera, which corresponds to the camera shutter being open for the entire time Δt an image is recorded.

In this method for determining 1-D diffusion coefficients, the step size Δx_n is directly related to the camera frame rate. We image at frame rates of 20 to 33 ms, and a faster frame rate would decrease the magnitude of the apparent step sizes. This relationship is accounted for in the definition of the CVE, however, where the average step size is divided by the time Δt. If motions existed that occurred on timescales much faster than those accessed by the camera sampling rate, then this would affect the diffusion measurements. Given limitations of both probe photostability/brightness and camera technology, however, an increased frame rate must be balanced by the need to acquire sufficient photons for single molecule localization. We achieve this balance by ensuring that localization is successful for consecutive frames in a trajectory, essentially by

maximizing the frame rate until this is not the case. In this way, we acquire images at the fastest speed possible that allows for collection of sufficient photons.

2.2.11 Hydrodynamic models for TALE diffusion

In order to estimate the effects of hydrodynamic friction on TALEs with increasing numbers of central repeats, we utilized a model developed for a rigid helix diffusing along its long axis[21]. The mean squared displacement of such a helix is given as:

$$\left\langle X_r^{\,2}(t)\right\rangle = 2k_B T \Xi_\theta \left(\Xi_z \Xi_\theta - r^2 \Xi_I^2\right)^{-1} t \tag{2.4}$$

With coefficients of friction given as:

$$\Xi_z = \left[1 + (2\pi r/d)^2\right] \Psi_1 \tag{2.5}$$

$$\Xi_\theta = 2r^2 \Psi_1 \tag{2.6}$$

$$\Xi_I = -\left(2\pi r/d\right) \Psi_1 \tag{2.7}$$

where *d* is the pitch of the helix, *r* is the radius of the helix and

$$\Psi_1 = \frac{4\pi\eta\lambda d}{\left[2 + (2\pi r/d)^2 \, ln\left[(d/2b)\left\{1 + (2\pi r/d)^2\right\}\right]\right]^2} \tag{2.8}$$

where *b* is the radius of the helical element and λ is the number of pitches in the TALE protein. This model predicts a λ^{-1} scaling for the diffusion coefficient of TALEs as a function of the number of repeats in the entire helical structure.

We further considered a model for rotationally-coupled 1-D diffusion of DNA binding proteins proposed by Blainey and coworkers to examine the effects of TALE repeat length on 1-D diffusion[22]. In this model, a protein is modeled as a sphere that tracks the helical structure of DNA, wherein one portion of the protein (i.e. the 'reading' domain) maintains constant contact with the DNA such that this domain is able to continually

interact with the nucleobases for sequence checking. Here, we considered the TALE NTR as the 'reading' domain that maintains contact with the DNA, while the CRD+CTR is considered as a globular protein that is pulled along by the NTR during the 'search' mode. In essence, this model differs from the helical pitch model (above) in that during the non-specific search mode, the CRD+CTR domain is assumed to adopt a globular structure instead of a highly ordered helical structure. Upon transitioning to the 'check' mode, the protein would undergo a major conformation change for local sequence checking.

We estimated the hydrodynamic radius of CRD+CTR domains using an a common scaling formula for the size of a globular protein[23]:

$$R = \left(\frac{3MWv}{4\pi}\right)^{\frac{1}{3}} \tag{2.9}$$

where MW is the molecular weight of the domain and *v* is the specific gravity.

Next, we obtained the scaling of 1-D diffusion with protein size as:

$$D_{slide}^{helix} \approx b^2 \frac{k_B T}{\left[8\pi\eta R^3 + 6\pi\eta R\left(R_{OC}\right)^2\right]} F\left(\varepsilon\right) \tag{2.10}$$

where *b* is the helical pitch of DNA and *R* and R_{OC} are the protein radius and minimum distance from the protein center to DNA center, respectively. Here, $F(\varepsilon)$ describes the energy landscape experienced along the DNA helix by the diffusing protein, which is assumed to be small (~1 k_bT), as discussed in the main text. Using estimates from the available crystal structures[1,2], we approximate $R_{OC} \approx R + 1$ nm. We obtain a scaling for 1-D diffusion based on variations in protein size given by (11):

$$D_{slide}^{helix} \sim \frac{1}{\left[R^3 + \frac{3}{4}R\left(R_{OC}\right)^2\right]} \tag{2.11}$$

2.3 Results and Discussion

2.3.1 Generation of DNA substrates

We generated long DNA templates that are tethered on both ends to a microscope coverslip surface via non-covalent biotin-Neutravidin linkages. The initial circular plasmids were linearized with SnabI, which generated two blunt ends (**Figure 2.1**). In order to generate 5' overhangs, we incubated the linearized plasmid with T4 DNA polymerase, which, in the absence of free nucleotides, has substantial 3'→5' exonuclease activity. When properly timed, short overhangs on the order of ~15 to 20 base pairs in length can generated before quickly quenching the reaction and heat inactivating the enzyme. A biotin-functionalized primer complimentary to one of the generated overhangs is added to the plasmid and ligated. This biotinylated DNA is then attached to the PEG/PEG-biotin-Neutravidin functionalized microscope coverslip (**Figure 2.2**). Under constant (laminar) fluid flow, a second biotinylated oligo complimentary to the free (untethered) overhang is added in a large molar excess (>10,000 fold excess in comparison to the approximate surface-level concentration of DNA). These oligos hybridize to the flow-extended DNA, thereby forming non-covalent linkages with free Neutravidin on the surface and generating dual tethered DNA substrates that remain extended in the absence of flow (**Figure 2.2**). Addition of chloroquine with the second oligo allows for additional extension, due to base stacking of chloroquine.

We characterized the dual-tethered DNA substrates using fluorescence microscopy, in order to determine the distribution of extensions relative to the crystallographic length of the 44.5 kbp plasmid (**Figure 2.3**). We find that large areas of

the functionalized coverslips contain dual-tethered DNA substrates. When imaged using Sytox Green intercalating dye and a 488 nm excitation source, these molecules are tightly distributed around an average length of 14.4 μm, or ~95% of the contour length of the plasmid (**Figure 2.3**). At this relative extension, the DNA substrates have relatively minimal transverse fluctuations and nearly absent lateral fluctuations. We calculated the apparent tension on the dual-tethered DNA substrates and then used this value to approximate the maximum magnitude of fluctuations. We calculated the longitudinal spring constant $k_{||}$ given by 2.12

$$k_{||}(l) = \frac{\partial F}{\partial z}|_l \tag{2.12}$$

where l is the length of the DNA, F is the force, and z is the extension of DNA . The force is estimated using the Marko-Siggia model for chain elasticity[20] (2.13)

$$\frac{fA}{k_B T} = \frac{z}{L} + \frac{1}{4(1-\frac{z}{L})^2} - \frac{1}{4} \tag{2.13}$$

where f is the force, A is the persistence length of dsDNA (50 nm), L is the contour length for the DNA substrate in question, and z is the extension of the DNA.

We calculate the force on the DNA given a 95% extension, and then use this value along with the corresponding longitudinal spring constant to calculate the maximum magnitude of longitudinal fluctuations (δz) using the equipartition theorem (2.14)

$$\frac{1}{2} k_B T = \frac{1}{2} k_{||} \langle \delta z^2 \rangle \tag{2.14}$$

These ± 40 nm fluctuations are not insignificant, rather, they are on the order of the localization uncertainty introduced due to the finite number of photons collected from a single dye molecule. The proportion of dual-tethered DNA molecules to single-tethered DNA molecules is adequate, typically in the range of 1:1 to 1:3.

2.3.2 Generation of fluorescently labeled TALE proteins

Traditionally, biological imaging depends on one of three key strategies for labeling a protein for fluorescence microscopy studies[24]. Immunolabeling involves fluorescent functionalization of antibody proteins (typically with several small molecule organic dyes) that then bind with high affinity directly or via a secondary antibody to the protein of interest. This approach introduces a large (often 100 kDa+) tag onto the protein, however. Depending on the experiment, this approach can be perturbative to visualization of desired biological function. A second approach is the use of genetically encoded fluorescent proteins. These proteins, such as the canonical green fluorescent protein (GFP) or associated variants (mCherry, mEos, YFP) provide a highly relevant approach for imaging proteins within living cells with the ability for endogenous expression. They are, however, significantly less bright than small-molecule fluorescent dyes, and like antibodies, relatively large in size (35 kDa+). The third main approach is the direct labeling of proteins with small molecule organic dyes, using either malemide or N-hydroxysuccinimide-functionalized dyes, which can be covalently attached to thiols and primary amines, respectively. These strategies, however, make precise, stoichiometric conjugation of dyes often difficult, unless only a single cysteine residue exists in the protein of interest. Engineering proteins such that they contain only a single cysteine, especially in proteins like TALEs with more than 10 cysteine residues, is challenging.

There is a growing fourth class of conjugation strategies that consist of genetically encoded elements that can be specifically conjugated to small molecule fluorescent dyes.

These strategies possess two main advantages: they are much smaller than antibodies and thus can be non-perturbative to the biological function of the protein when strategically placed, and they allow for conjugation of a variety of small molecule dyes allowing them to produce brighter, more photostable molecules than fluorescent proteins for single molecule tracking. The HaloTag, SNAP tag, and CLIP tag all involve genetically encoded elements ~20 kDa in size that bind irreversibly to a moiety that is easily conjugated to a single fluorescent dye[25]. A unique fourth member of this class is the genetically encoded aldehyde tag[14,15]. This labeling strategy is a two-part system. First, the six amino acid sequence LCTPSR is encoded to the protein of interest. Next, the protein with the LCTPSR tag is co-expressed with formylglycine generating enzyme (FGE), which recognizes this motif, and converts the cysteine in the tag to a formylglycine. This formylglycine residue bears an aldehyde moiety, which is not found in any other naturally occurring amino acid side chain. The result is a single location with which to covalently label the protein of interest via reaction with hydrazine-functionalized organic dyes.

Here, we utilized an aldehyde labeling strategy in order to achieve a one-dye-to-one protein stoichiometry for labeling TALEs[14,15]. We chose a 21.5 repeat-long TALE, previously developed for the editing and correction of the human β-globin gene containing a mutation associated with sickle cell disease[26], as a TALE construct to study (**Figure 2.4**). We selected the TALE binding site with a minimal number of guanines, since recognition of guanine by various reported RVDs comes as a compromise between specificity and strength of binding[27]. Hence, the designed TALE only has 2 unspecific but strong NN RVDs (**Figure 2.4**). The most C-terminal repeat (22nd repeat in this case),

which is typical of TALEs and is commonly referred to as the 'last half-repeat', was designed to be only 20 residues long[27,28]. A graduate student colleague Zhanar Abil produced several variants in Professor Huimin Zhao's lab, including TALEs with the aldehyde tag at the N and C termini of the protein. In the end, the N-terminal variant was successfully expressed, purified, and labeled with Cy3 hydrazine (**Figure 2.5**). We chose Cy3 hydrazine because it is sufficiently bright and photostable to promote long-lived imaging of single TALEs with good spatial resolution. In order to confirm that this labeling scheme was non-perturbative to the biological function of TALEs, we compared the binding activity of wild type (WT) TALEs to aldehyde-tagged Cy3-TALEs using a fluorescence anisotropy binding assay. No significant difference in binding affinity was observed (**Figure 2.6**). Additionally, the aldehyde moiety allows for conjugation to a variety of other molecules of interest, which became a critical component in future studies. This initial TALE construct targeted a 22 base pair binding site, placing it well within the range of naturally-occurring and engineered TALE lengths[26,29].

2.3.3 Direct observation of TALE binding and 1-D diffusion

With a stable, fluorescently labeled TALE variant and robust dual-tethered DNA molecules both in hand, we sought to directly visualize the interaction of TALEs with DNA substrates (**Figure 2.7**). Fluorescently-labeled TALEs were added to microfluidic flow cells containing the DNA substrates via a computer-controlled syringe pump, and illuminated via total internal reflection (TIR) using a diode-pumped solid state (DPSS) laser (532 nm excitation) (**Figures 2.7, 2.8, and 2.9**). The DNA-functionalized surface of the microfluidic cell, once illuminated, was imaged using an electron multiplying charged coupled device camera (EM-CCD) that allowed for minimal background and

collection of a sufficient number of photons. Upon introduction of low concentrations (~50 pM) of TALEs, binding to the DNA substrates was immediately observed, indicating the strong propensity of TALEs to bind non-specifically to DNA under moderate salt conditions (90 mM KCl) and near-neutral pH (pH = 8.1) (**Figure 2.10**). The TALEs bound to DNA and the vast majority began diffusing one-dimensionally along the DNA substrates. In this way, we directly observed the ability of TALEs to utilize a facilitated search mechanism to reduce the dimensionality of their target search process for the first time.

In order to quantify the behavior TALEs during 1-D search along DNA, we tracked their centroid positions using localization techniques borrowed from super-resolution imaging software (RapidStorm)[18]. With the position of TALEs determined to ~25-50 nm precision within individual frames, we linked together successive imaging frames collected via the CCD camera to create molecular trajectories of TALE diffusion. We maintained low concentrations of TALEs in the microfluidic chambers such that there were only 1 or 2 TALE molecules bound to each DNA strand at any given time, enabling facile tracking of individual TALEs. Examining distributions of the displacement of TALEs between successive frames (termed step sizes), we found them to be best described by Gaussian distributions centered on zero (**Figure 2.11b**). Furthermore, plotting the mean squared displacement (MSD) of TALEs against the time of trajectories yields a linear relationship (**Figure 2.11a**). This result confirms that the TALEs we observed were engaging in a directionally unbiased random walk. This served as a useful control, given that TALEs are not ATP-dependent in any reported mode of action. Individual TALE diffusion trajectories were then analyzed to extract the

one-dimensional diffusion coefficients $D_{1\text{-}D}$ of TALEs. Here we utilized a covariance-based estimator (CVE) to determine $D_{1\text{-}D}$ as opposed to the more traditional mean squared displacement method[12]. This approach was taken to minimize compounding of the localization error of each frame, determined from photons collected, background, and DNA fluctuations to be on the order of 25-65 nm. Calculated values of $D_{1\text{-}D}$ were distributed broadly, with an average value of $5.9 \pm 2.3 \times 10^6$ bp^2/s (**Figure 2.12**). This average 1-D diffusion speed places TALEs at the upper limit of previously recorded DNA binding proteins, especially given their large size relative to many other proteins[22,30–32].

Seeking to better understand the broad distribution of apparent 1-D diffusion coefficients, we examined the individual trajectories of TALEs more closely. While many of the shorter lifetime binding events involved TALEs diffusing rapidly with no apparent periodicity, many of the longer TALE trajectories displayed regions of slow diffusion followed by regions of rapid diffusion (**Figure 2.13**). TALEs bind, diffuse rapidly, then slow down as if they are confined to a smaller region of DNA, and then take off again, diffusing rapidly, all within one individual trajectory. This heterogeneous behavior is likely the main reason we observed such broad distributions of one-dimensional diffusion coefficients for individual TALE trajectories.

In addition to determining the diffusion coefficients for TALEs, we also measured their bound lifetimes (**Figure 2.14**). Here we found that TALE binding lifetime distributions were best described by a multi-exponential model, indicating the interaction of TALEs and DNA was not just a simple bind-diffuse-unbind process. We fit these distributions to double-exponential functions to extract characteristic lifetimes of bound

TALEs. As expected, the lifetimes decreased with increasing ionic strength in solution, due to the increased charge screening effects. Given that TALEs have a patch of positively charged amino acids in their DNA binding-region, in addition to the positively charged amino acids clustered near the NTR[8], it is not surprising that their propensity to bind and remain bound to DNA is influenced by salt concentration.

By modulating the salt concentration in the imaging buffer, we also gained a second major insight into TALE sequence search. The average one-dimensional diffusion coefficient increased as we increased the salt concentration (via KCl) of the imaging buffer, in addition to decreasing the average bound lifetime of TALEs (**Figure 2.12a** compared to **Figure 2.12b**). Taking the seminal work of Berg, Winter, and von Hippel into account, a strong dependence of $D_{1\text{-}D}$ on ionic strength is indicative of a DNA-binding protein that utilizes 'hopping' events in some portion of its non-specific sequence search as opposed to a pure sliding mechanism[33,34]. The rationale for this argument is based upon the idea that a protein sliding along DNA, closely tracking the phosphate backbone, displaces ions in one direction (the direction of protein movement) while counter ions re-condense on the regions immediately vacated by the protein. In this way, the net change of counter ions associated with the DNA is zero and thus an increase in ionic strength does not impact the energetics of search. If instead the protein engages in short dissociation-association events (so-called hopping events), a higher ionic strength translates into a larger number of counter-ions that must be displaced upon each binding event, thus shifting the energetics of the search such that the protein spends more time unbound from DNA and diffusing freely in solution. Because the three-dimensional diffusion coefficient $D_{3\text{-}D}$ is much greater than $D_{1\text{-}D}$, it holds that an increase in ionic

strength translates into a greater apparent one-dimensional search for hopping proteins. We concluded that TALEs may utilize some degree of apparent hopping behavior in their search, though reconciling hopping behavior with a wrapped superhelical structure remained challenging. Previous single molecule studies of DBPs have revealed that both behaviors persist for different proteins, with the majority of observed proteins utilizing a pure sliding mechanism for search[31,32,35,36].

2.3.4 Dynamics of TALE NTR provide evidence for a delegation of function amongst TALE subdomains

Our initial observations of TALE-DNA dynamics at the single molecule level provided us with three key findings. First, TALE one-dimensional diffusion is heterogeneous, with periods of rapid diffusion interspersed with periods of attenuated diffusion. Second, TALE binding events are best described by a multi-exponential decay function, clearly displaying multiple lifetimes. Third, TALE diffusion is salt-dependent, suggesting they do not utilize a pure sliding mechanism for sequence search. These three findings point towards a model for TALE search that involves more than one simple mode of bind-diffuse-unbind. With this idea in mind, we sought to determine how the TALE structure could reconcile an apparent multi-state model for TALE sequence search. We started by more closely examining the structure of TALEs. Previous single molecule studies of tumor suppressor p53 had revealed two distinct modes of search[30], but p53 is a tetrameric protein and its respective subunits were found to contribute to the apparent 'search' and 'bind' modes. While TALEs are monomeric proteins consisting of essentially a single helix, a 2012 study by Gao *et al* revealed that the N-termainal region (NTR) of TALEs contains four domains that resemble the canonical TALE repeats, albeit

without sequence specificity and they subsequently termed these repeats 0, -1, -2, and -3[8]. A higher concentration of positively charged amino acids within this domain, combined with its structure, led Gao *et al* to propose that the TALE NTR may function to mediate the non-specific search of TALEs. Furthermore, they showed that the NTR was critical for TALE binding and could bind non-specific DNA on its own, even in the absence of any central repeat domain (CRD) or C-terminal region (CTR). Our hypothesis, taking into account a possible multi-mode search and the findings of Gao *et al*, was that the TALE NTR was responsible for nucleating TALE binding events and facilitating rapid searching events.

In order to test this hypothesis, we generated a fluorescently labeled TALE NTR mutant (again via aldehyde tag incorporation followed by Cy3-hydrazide conjugation), which lacked the CRD+CTR subdomains, and directly observed its interactions with DNA using our single molecule imaging assay (**Figure 2.15**). Our first insight was that the TALE NTR displayed significantly attenuated binding affinity compared to full-length TALEs. In order to observe any appreciable binding of the fluorescently labeled NTR to the dual-tethered DNA substrates, we had to increase the concentration of protein in the imaging solution by more than 20 fold, from 50 pM to 1-1.5 nM. This result corroborates the calorimetric measurements of Gao *et al* who found a large decrease in the binding affinity of the NTR compared to the full-length TALE[8], and is further supported by our fluorescence anisotropy measurements showing a major decrease in NTR affinity compared to the 21.5 repeat TALE binding (**Figure 2.6**). We observed no binding of an NTR-deficient TALE mutant containing only the CRD+CTR, via direct observation using SMFM or via anisotropy measurements. Thus, our first insight is that

the TALE NTR is indeed critical for nucleating TALE binding to nonspecific DNA, but is not the only component of TALE structure responsible for binding.

Observation of NTR binding via SMFM also required sufficiently low ionic strengths (10-50 mM KCl) in order for binding events lasting at least 5 to 10 consecutive frames (100 to 600 milliseconds total) to be recorded. Given that full length TALE binding events often persisted for several seconds or longer even at KCl concentrations greater than 100 mM, it is clear that the NTR is one component of TALE binding to nonspecific DNA, but not the sole determinant of stable nonspecific binding. We next determined the one-dimensional diffusion coefficients $D_{1\text{-}D}$ for the NTR using the above-described localization and CVE strategies, and these measurements yielded three key insights into the role of NTR. First, the average values of $D_{1\text{-}D}$ were more than seven-fold higher than those for the full length TALE measured under equivalent ionic strength (**Figure 2.15**). The hydrodynamic drag removed by the deletion of the CRD+CTR is not sufficient to account for the additional speed gained by the NTR in its nonspecific 1-D search. Second, we observe no significant ionic strength dependence of NTR diffusion (**Figure 2.16**). The range we are able to test is small (10-50 mM KCl), however, but this nevertheless elicits a significant change in $D_{1\text{-}D}$ for full-length TALEs. In this way, in view of the seminal work of Winter, Berg, and von Hippel[34], the NTR appears to utilize a pure 1-D sliding mechanism for nonspecific search, in contrast to the full-length TALEs. Finally, we observe no periods of 'attenuated' diffusion, characteristic of full-length TALE diffusional trajectories in those of the NTR. The trajectories of the NTR are short, however, due to the lower binding affinity for DNA, but appear to be comprised of only rapid 1-D sliding events wherein the NTR tracks the DNA helix closely. These insights

into TALE NTR function begin to help reconcile the heterogeneity of full-length TALE behavior. The NTR mediates critical initial binding of the TALE to DNA via its concentrated regions of positively charged amino acids, particularly at the most N-terminal portion, which appears to be a 'cap' of positive charge. Once the NTR has nucleated binding, it is capable of mediating short one-dimensional sliding events. If the rest of the TALE (CRD+CTR) successfully engages the DNA, likely by wrapping around in an extended helical conformation, these short, rapid sliding events can be interspersed by local sequence 'checking' which gives rise to periods of 'attenuated' diffusion. Short sliding events can also potentially be interspersed by partial/full dissociation of the TALE from the DNA, followed by subsequent rebinding. These events would explain the strong dependence of full-length TALEs' $D_{1\text{-}D}$ values on ionic strength.

2.3.5 Effects of TALE CRD size provide further evidence of a bind and check mechanism for sequence search

Given that naturally occurring and engineered TALEs range in length from 10 to 30 tandem repeats[29], and thus target DNA sequences of these lengths, we next sought to understand how the size of the TALE CRD influenced their search dynamics. To this end, we constructed fluorescently labeled TALE constructs containing 11.5 and 15.5 repeat CRDs and observed their search dynamics via our SMFM assay. Although slightly higher concentrations of shorter TALEs were required to observe appreciable binding, the difference between concentrations required for the 21.5 repeat TALE and these shorter TALEs was only two or three-fold at most (150-600 pM compared to the 25-100 pM required to observe TALE 21.5 repeat construct binding). This observation, combined with their longer binding lifetime, which were close to those of the 21.5 repeat

TALE, indicates that these TALE constructs behave more similarly to full-length TALEs compared to the NTR. We therefore hypothesize that the search behavior of TALEs is likely fundamentally similar over this range of CRD lengths (eleven and a half to twenty-one and a half repeats).

We quantified the diffusive behavior of the shorter TALEs using our CVE method and found that their average values tracked with the number of repeats. That is, the shorter, smaller TALEs diffused more quickly at equivalent ionic strengths when compared to the 21.5 repeat TALE (**Figure 2.17**). Next, we sought to understand if these differences in $D_{1\text{-}D}$ could be accounted for via a pure hydrodynamic drag argument. If the CRD repeats are disengaged from the DNA during search, for instance if the NTR simply dragged them along like a cargo, then the slowdown due to additional CRD repeats could be predicted by calculating the additional drag imposed by their presence. We modeled the additional drag imposed by larger CRDs two different ways. In the first, we assumed that the CRD was being drug along as the TALE spiraled along the DNA helix[22]. In the second, we assumed that the TALE takes on a true helical shape, and the entire helix spirals along the DNA[21], thus accounting for additional CRD repeats by increasing the length of our theoretical helix. Strikingly, we found that neither model accounted for the slowdown experienced by TALEs with longer CRDs. We propose a model for TALE sequence search that combines this insight with those gained through studies of the NTR-only mutant and TALE salt dependence. In this model, the TALE engages in short sliding events, interspersed by dissociation/re-association events and local sequence checking. It assumes a helical structure in some of these events (those that last for several seconds or more) wrapped around the DNA, allowing the TALE to

compress along its superhelical axis during checking events (**Figure 2.18**). These local checking events increase in either frequency, duration, or both frequency and duration as the length of the TALE CRD is increased, accounting for the apparent slowdown in $D_{1\text{-}D}$ for longer TALEs, given that this slowdown is not purely due to hydrodynamic drag.

2.4 Concluding Remarks

In this work, we directly observe the sequence search of TALE proteins using single molecule fluorescence microscopy. While some researchers had postulated the possibility of a facilitated diffusion mechanism for TALEs in the past, our work provides the first evidence that TALEs utilize a 1-D search to locate their target sequences. Crystal structures of TALEs reveal a tightly wrapped, superhelical conformation for specifically bound TALEs and an extended superhelical conformation for DNA-free TALEs[1,2]. Taking only the specifically bound TALE conformation into account, it could be difficult to picture TALEs diffusing rapidly along DNA; their tight association with DNA appears to present a large barrier to any translocation. Despite this, however, we find that TALEs readily diffuse over long stretches of DNA at rapid speeds, and we postulate that the TALE search conformation is likely closer to the DNA-free structure. This finding comes in contrast to a recent single molecule study of the CRISPR/Cas9 system wherein the authors observe no detectable 1-D diffusion of Cas9 and conclude that Cas9 locates its target purely by 3-D diffusion[37].

Molecular trajectories of TALE search along the dual-tethered DNA substrates revealed a surprising degree of heterogeneity, which explained the heterogeneous distributions of 1-D diffusion coefficients we determined. These trajectories contained periods of both rapid translocation in one direction as well as periods of reduced or

constrained motion. We postulate that these periods of fast and slow searching corresponded to two distinct modes of TALE search. Drawing on theoretical models for the search speed-stability paradox, we termed these search and check modes. In the search mode, we hypothesize that TALEs adopt an extended helical conformation that allows for rapid 1-D diffusion along DNA substrates wherein TALEs are out of phase with the DNA major groove. We propose that when in the check mode, TALEs compress along their superhelical axis and engage the local DNA more closely as they tightly associate with the DNA major groove. If the TALE correctly checks a target site, the energetic barrier to unbinding and resuming the search mode will be significantly high such that the TALE remains bound to its target. If instead the TALE has not located its target and checks an off-target site, it will generally be able to quickly return to the search mode, as it will not be trapped by in an energy well preventing it from assuming the search conformation. Failure to return to the search mode anywhere outside of the target site would result in an off-target event. This general mode for sequence search of DBPs has been experimentally observed for tumor suppressor p53, though it achieves such a multi state search mechanism by delegating function to its various distinct subunits[30].

Given that TALEs are monomeric proteins, consisting of only one helical domain, we hypothesized that there might be some delegation of function between its subdomains, and this division of function may help further reconcile a possible two-state model for sequence search. In late 2012, a crystal structure of the TALE N-terminal region (NTR) was published, and additional *in vitro* binding experiments carried out as part of this study demonstrated that the NTR was capable of binding DNA nonspecifically on its own[8]. Furthermore, the authors of the NTR crystal structure study and others have shown

that TALE binding is sensitive to excessive NTR truncation[38], highlighting the its importance in nucleating DNA binding. We generated a fluorescently labeled NTR without the accompanying CRD+CTR and observed that it readily diffused along DNA one dimensionally. NTR diffusion was rapid and lacked the slow/fast periods of diffusion characteristic of full-length TALE search. Additionally, it lacked an apparent dependence on ionic strength, suggesting it engages is short 1-D sliding events, remaining tightly associated with DNA during its brief excursions. Coupled with our observation of an apparent two state search mechanism, we postulate that the NTR serves as a primary driver of TALE 1-D diffusion, nucleating initial binding and facilitating 1-D sliding events. This finding further highlights the need for NTR-inclusive design rules for TALEs.

We extended our study of TALE search to include different CRD lengths (11.5 and 15.5 repeat constructs were added to compliment the NTR and 21.5 repeat construct) and we found that, as expected, shorter TALEs diffused more quickly than longer TALEs. Unexpectedly, however, a simple hydrodynamic drag model could not explain the relationship between diffusion speed and the size of TALEs. This finding leads us to posit that the TALE CRD is somehow actively engaging DNA during the search process as opposed to being pulled as cargo. We return to our two state model, and propose that the length of the CRD may directly impact the rate of checking events, their duration, or both. In this way, a longer CRD causes the TALE to spend more time overall in the check mode as compared to shorter TALEs that spend more time in the search mode. This hypothesis, if true, provides additional evidence for why off-target binding events occur more frequently for longer TALEs[39].

2.5 Figures

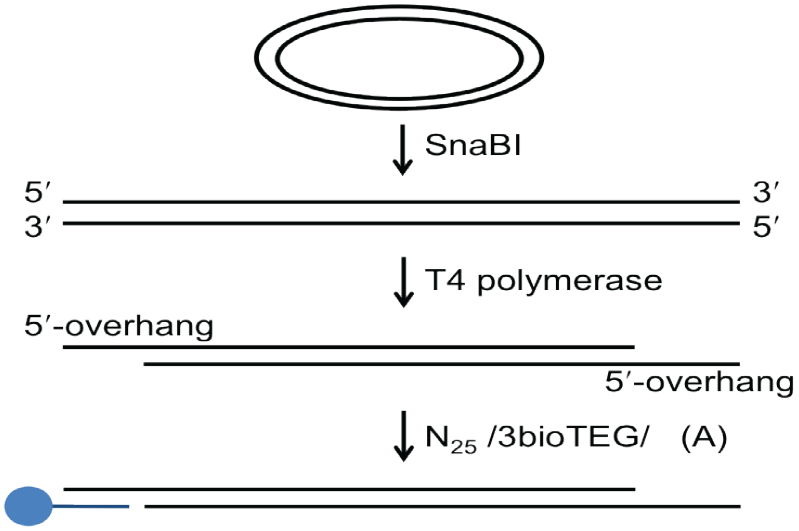

Figure 2.1 Process for generating biotin functionalized DNA templates. The circular plasmid is linearized and sticky ends are generated by the exonuclease activity of T4 polymerase. A biotinylated oligo complimentary to one sticky end is annealed to the now linear plasmid (biotin oligo in blue).

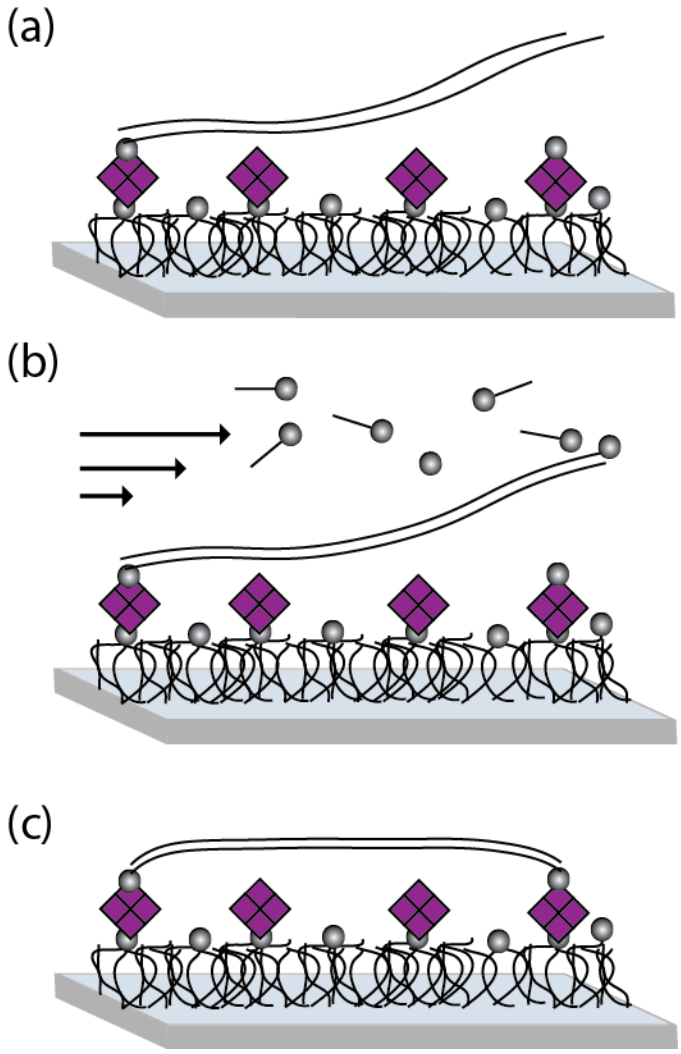

Figure 2.2 Generation of dual tethered DNA templates. (a) Biotinylated DNA templates are incubated in flow cells with Neutravidin (purple) functionalized surfaces. Excess DNA is washed away and the single tethered substrates are extended in flow. (b) A second biotinylated oligo and 0.1 mM chloroquine are added in the flow solution, and the oligo hybridizes with the free sticky end. (c) The now biotinylated free end attaches to a second Neutravidin molecule on the coverslip surface.

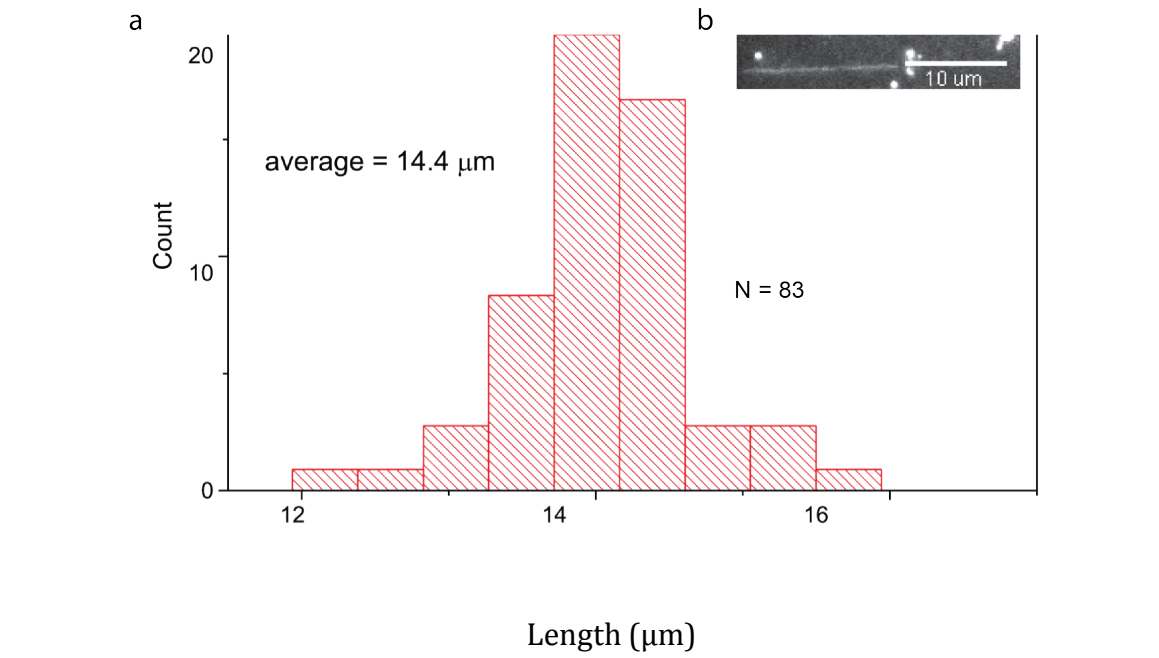

Figure 2.3 Distribution of end-to-end stretched length of dual-tethered DNA templates. (a) Distribution of sizes of stretched, dual-tethered DNA templates. The 44.5 kbp substrates are stretched to an average of 14.4 μm, approximately 95% of the crystallographic length of the DNA. (b) Representative image of SYTOX Green labeled, dual-tethered DNA molecules.

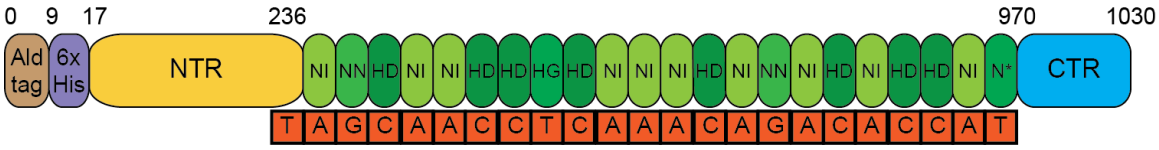

Figure 2.4 TALE21.5 construct used in this study. The TALE repeats are shown in green (representing the full CRD), bound to their target DNA substrate, with only the repeat variable diresidues (RVDs) specified. Numbers above represent the locations of each domain in the 1,030 residue total length of the construct.

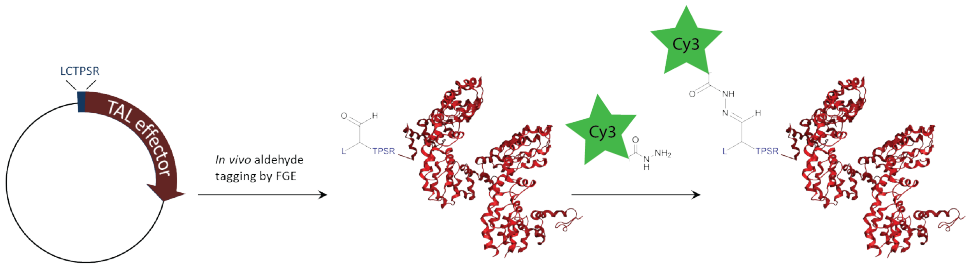

Figure 2.5 Aldehyde labeling scheme used to tag TALE proteins with Cy3 dye at the N-terminus in a 1:1 dye:protein stoichiometry. A six amino acid motif (LCTPSR) is cloned into the N-terminus of the TALE protein, and the construct is co-expressed with formylglycine generating enzyme (FGE), which then converts the cysteine to a formylglycine bearing an unnatural aldehyde. Aldehyde-specific conjugation via hydrazine-functionalized Cy3 organic dyes then allows for non-perturbative, site-specific labeling.

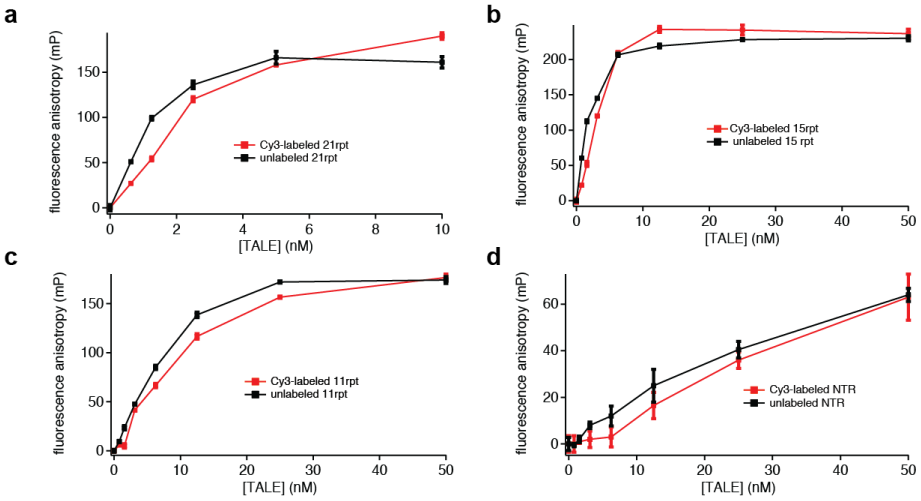

Figure 2.6 Fluorescence polarization data for TALE constructs used in single molecule experiments. (a) Binding of labeled and unlabeled 21.5 repeat TALE to target DNA (b) Binding of labeled and unlabeled 15.5 repeat TALE to target DNA (c) Binding of labeled and unlabeled 11.5 repeat TALE to target DNA (d) Binding of labeled and unlabeled TALE NTR to target DNA.

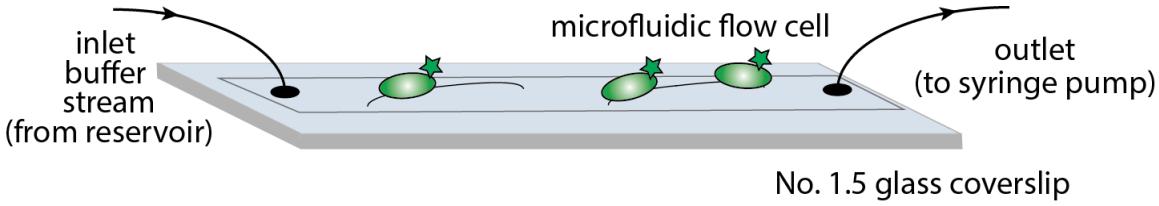

Figure 2.7 Microfluidic flow cell utilized in SMFM experiments. Solutions are introduced into the flow cell from a reservoir via a syringe pump connected opposite the buffer source. The reservoir allows for easy and rapid exchange of buffers during experiment, and the syringe pump allows for controlled flow that maintains the integrity of dual-tethered DNA substrates.

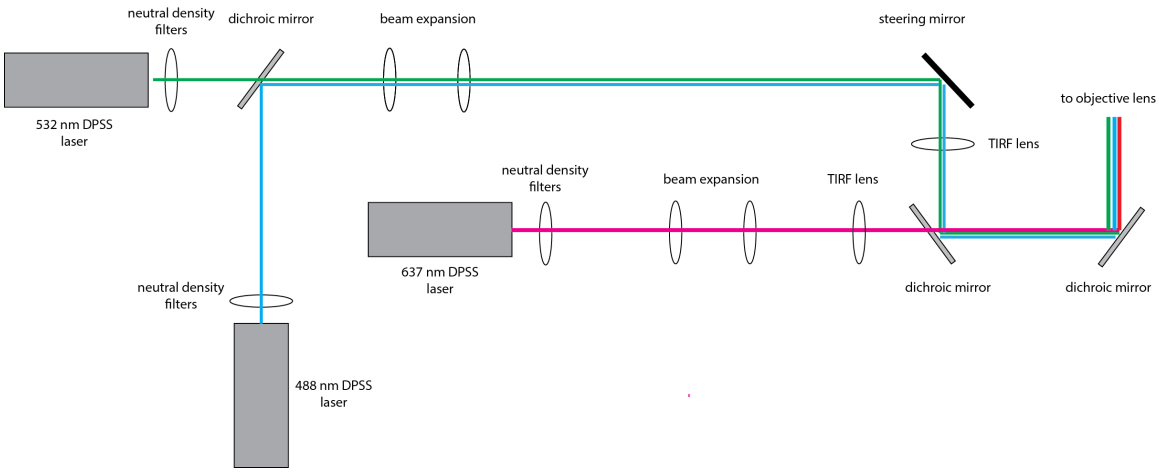

Figure 2.8 Schematic representation of the single molecule fluorescence microscopy setup used in this work. Illumination is provided by three spectrally distinct diode pumped solid state (DPSS) lasers that are all directed into an inverted fluorescence microscope via dichoric mirrors within a microscope filter cube,

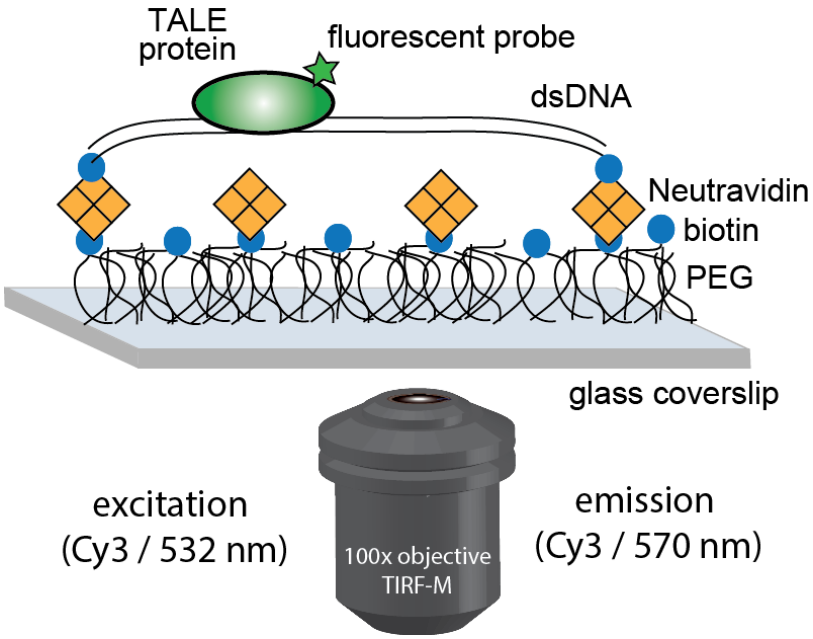

Figure 2.9 Dual tethered DNA teamplates showing chemical linkages and imaging setup. Fluorescently labeled TALEs are imaged via total internal reflection fluorescence microscopy (TIRF-M) via excitation at 532 nm and subsequent collection of 570 nm emission.

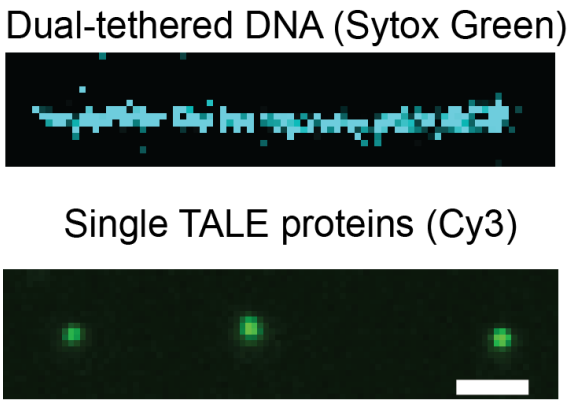

Figure 2.10 Representative images of Sytox Green stained dual tethered DNA and Cy3 labeled TALEs. Scale bar is equal to 2.5 μm.

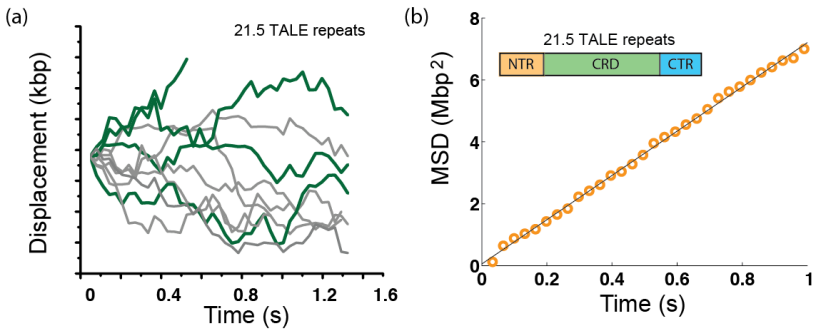

Figure 2.11 TALE diffusion can be described by a random Brownian walk. (a) Individual molecular trajectories of TALE proteins diffusing along DNA in one dimension, illustrating that TALE 1-D diffusion is not directionally biased. (b) Mean-squared displacement (MSD) versus time for an ensemble of TALE trajectories. The linear slope of the MSD versus time plot provides further evidence that TALE diffusion is a thermally driven process.

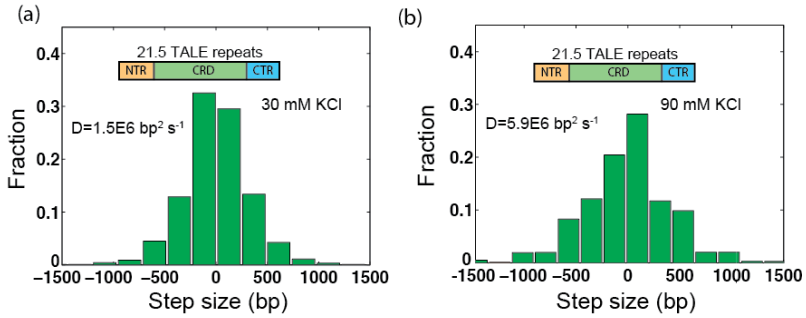

Figure 2.12 Step size histograms derived from 21.5 repeat construct TALE diffusion trajectories at different salt concentrations. Steps are defined as the distance an individual TALE diffuses between consecutive camera frames taken during an imaging experiment, but do not imply that TALEs are physically stepping as in the case of an ATP driven motor such as kinesin. 1-D diffusion coefficients are calculated from these step size distributions. Here we visualize the speed-up in TALE diffusion as the imaging solution is modulated from (a) 30 mM KCl to (b) 90 mM KCl.

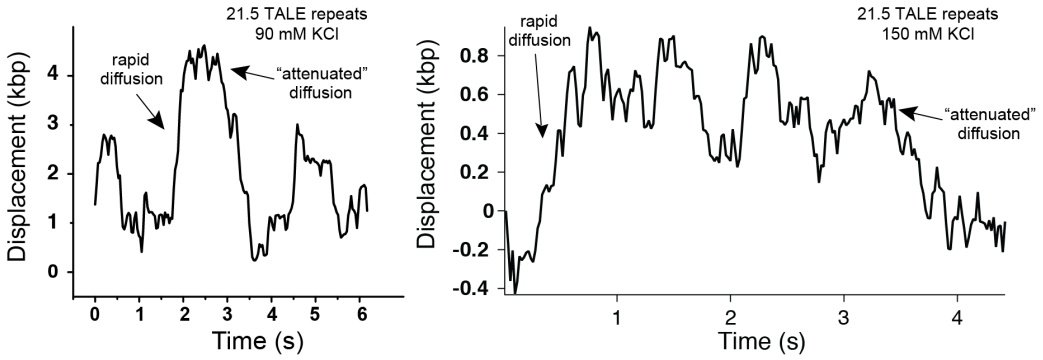

Figure 2.13 Sample trajectories of TALE21.5 searching DNA templates. In these trajectories, taken at 90 and 150 mM KCl, the proteins exhibit both rapid and slow (attenuated) diffusion, suggesting that that TALEs engage in a two-state search mechanism.

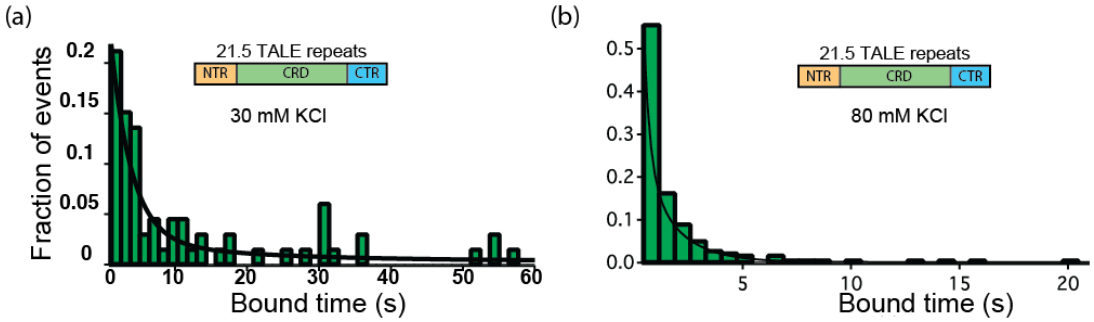

Figure 2.14 Distributions of TALE binding times. Distributions were acquired at (a) 30 mM KCl and (b) 80 mM KCl and fit to a double exponential decay function.

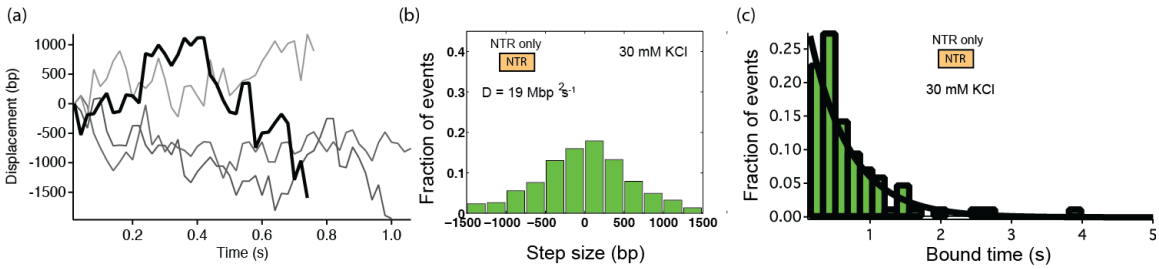

Figure 2.15 TALE NTR dynamic search behavior is qualitatively different compared to the full-length TALE21.5. NTR diffusion (a, b) is rapid and lacks the characteristic slow periods observed for full-length TALEs and (c) NTR binding times lack the characteristic longer binding modes found with the full-length TALEs. Here, the binding lifetimes are well described by a single exponential decay function.

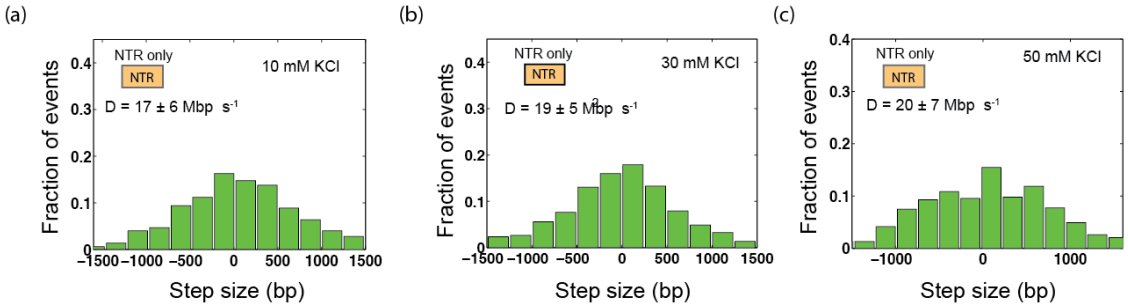

Figure 2.16 TALE NTR diffusion as a function of monovalent ionic strength. Measurements at (a) 10 mM, (b) 30 mM, and (c) 50 mM KCl show very little increase in 1-D diffusion coefficients. NTR diffusion events are very short in duration, but more than 12 times more rapid than full-length TALE diffusion at similar ionic strengths.

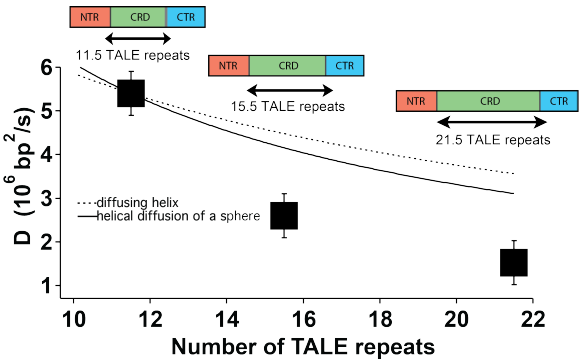

Figure 2.17 Relationship between TALE CRD length and 1-D diffusion speed. TALE diffusion was measured at 30 mM KCl for all constructs. The slowdown of diffusion with increasing CRD length was modeled both as a helical shape diffusing in a helical path, as well as the helical diffusion of a sphere, but neither hydrodynamic drag model could account for the extent of slowdown observed.

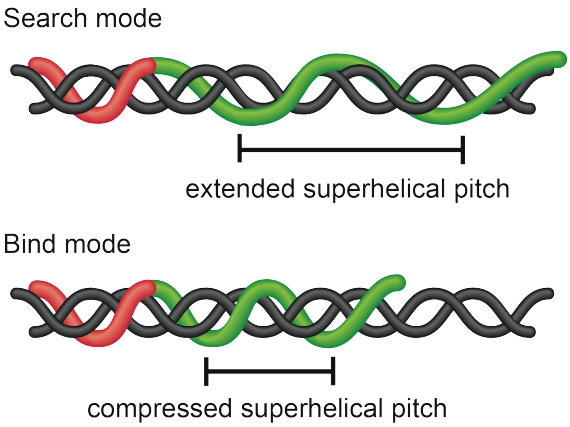

Figure 2.18 Proposed model for TALE two-state search model. In this model, TALEs are able to adopt an extend superhelical pitch during their search mode, placing them out of phase with the DNA major groove and reducing the effective frictional coefficient to diffusion. At an unknown frequency, they can then compress along their superhelical axis, aligning in-phase with the DNA major groove and effectively checking the local sequence (bind mode). The frequency and/or duration of checking events may scale with the length of the CRD.

2.6 References

1. Mak, A. N.-S., Bradley, P., Cernadas, R. A., Bogdanove, A. J. & Stoddard, B. L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335,** 716–9 (2012).

2. Deng, D. *et al.* Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **335,** 720–723 (2012).

3. Moscou, M. J. & Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326,** 1501 (2009).

4. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326,** 1509–1512 (2009).

5. Jiménez-Menéndez, N. *et al.* Human mitochondrial mTERF wraps around DNA through a left-handed superhelical tandem repeat. *Nat. Struct. Mol. Biol.* **17,** 891–893 (2010).

6. Lei, H., Sun, J., Baldwin, E. P., Segal, D. J. & Duan, Y. *Conformational elasticity can facilitate TALE-DNA recognition*. *Advances in Protein Chemistry and Structural Biology* **94,** (Elsevier Inc., 2014).

7. Mak, A. N.-S., Bradley, P., Bogdanove, A. J. & Stoddard, B. L. TAL effectors: function, structure, engineering and applications. *Curr. Opin. Struct. Biol.* **23,** 93–9 (2013).

8. Gao, H., Wu, X., Chai, J. & Han, Z. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.* **22,** 1716–20 (2012).

9. Slutsky, M. & Mirny, L. a. Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. *Biophys. J.* **87,** 4021–4035 (2004).

10. Wicky, B. I. M., Stenta, M. & Dal Peraro, M. TAL Effectors Specificity Stems from Negative Discrimination. *PLoS One* **8,** e80261 (2013).

11. Cuculis, L., Abil, Z., Zhao, H. & Schroeder, C. M. Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nat. Commun.* **6,** 7277 (2015).

12. Vestergaard, C. L., Blainey, P. C. & Flyvbjerg, H. Optimal estimation of diffusion coefficients from single-particle trajectories. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **89,** (2014).

13. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **39,** e82 (2011).

14. Carrico, I. S., Carlson, B. L. & Bertozzi, C. R. Introducing genetically encoded

aldehydes into proteins. *Nat. Chem. Biol.* **3,** 321–322 (2007).

15. Shi, X. *et al.* Quantitative fluorescence labeling of aldehyde-tagged proteins for single-molecule imaging. *Nat. Methods* **9,** 499–503 (2012).

16. Schroeder, C. M., Blainey, P. C., Kim, S. & Xie, X. S. Hydrodynamic flow-stretching assay for single-molecule studies of nucleic acid–protein interactions. *Single-Molecule Tech. A Lab. Man.* 461–492 (2008).

17. Yardimci, H., Loveland, A. B., van Oijen, A. M. & Walter, J. C. Single-molecule analysis of DNA replication in Xenopus egg extracts. *Methods* **57,** 179–86 (2012).

18. Wolter, S. *et al.* rapidSTORM: accurate, fast open-source software for localization microscopy. *Nat. Methods* **9,** 1040–1041 (2012).

19. Thompson, R. E., Larson, D. R. & Webb, W. W. Precise Nanometer Localization Analysis for Individual Fluorescent Probes. *Biophys. J.* **82,** 2775–2783 (2002).

20. Marko, J. F. & Siggia, E. D. Stretching DNA. *Macromolecules* **28,** 8759–8770 (1995).

21. Butenko, A. V, Mogilko, E., Amitai, L., Pokroy, B. & Sloutskin, E. Coiled to Diffuse: Brownian Motion of a Helical Bacterium. *Langmuir* **28,** 12941–12947 (2012).

22. Blainey, P. C. *et al.* Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.* **16,** 1224–1229 (2009).

23. Fournier, R. L. *Basic transport phenomena in biomedical engineering*. (CRC Press, 2011).

24. Agrawal, U., Reilly, D. T. & Schroeder, C. M. Zooming in on biological processes with fluorescence nanoscopy. *Curr. Opin. Biotechnol.* **24,** 646–653 (2013).

25. Stagge, F., Mitronova, G. Y., Belov, V. N., Wurm, C. A. & Jakobs, S. Snap-, CLIP- and Halo-Tag Labelling of Budding Yeast Cells. *PLoS One* **8,** 1–9 (2013).

26. Sun, N., Liang, J., Abil, Z. & Zhao, H. Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Mol. Biosyst.* **8,** 1255 (2012).

27. Christian, M. *et al.* Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **186,** 757–61 (2010).

28. Mussolino, C. & Cathomen, T. TALE nucleases: Tailored genome engineering made easy. *Curr. Opin. Biotechnol.* **23,** 644–650 (2012).

29. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29,** 143–148 (2011).

30. Tafvizi, A., Huang, F., Fersht, A. R., Mirny, L. A. & van Oijen, A. M. A single-molecule characterization of p53 search on DNA. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 563–568 (2011).

31. Kochaniak, A. B. *et al.* Proliferating cell nuclear antigen uses two distinct modes to move along DNA. *J. Biol. Chem.* **284,** 17700–17710 (2009).

32. Blainey, P. C. *et al.* Regulation of a viral proteinase by a peptide and DNA in one-dimensional space IV: Viral proteinase slides along dna to locate and process its substrates. *J. Biol. Chem.* **288,** 2092–2102 (2013).

33. Winter, R. B., Berg, O. G. & von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor--operator interaction: kinetic measurements and conclusions. *Biochemistry* **20,** 6961–77 (1981).

34. Berg, O. G., Winter, R. B. & von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* **20,** 6929–6948 (1981).

35. Etson, C. M., Hamdan, S. M., Richardson, C. C. & van Oijen, A. M. Thioredoxin suppresses microscopic hopping of T7 DNA polymerase on duplex DNA. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 1900–1905 (2010).

36. Komazin-Meredith, G., Mirchev, R., Golan, D. E., van Oijen, A. M. & Coen, D. M. Hopping of a processivity factor on DNA revealed by single-molecule assays of diffusion. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 10721–6 (2008).

37. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–7 (2014).

38. Schreiber, T. *et al.* Refined Requirements for Protein Regions Important for Activity of the TALE AvrBs3. *PLoS One* **10,** e0120214 (2015).

39. Meckler, J. F. *et al.* Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Res.* **41,** 4118–4128 (2013).

Chapter 3: TALEs Search DNA Using a Rotationally Decoupled Mechanism*

3.1 Introduction

DNA binding proteins responsible for locating and binding specific sequences face a tremendous task in their search of millions of non-specific sequences. As discussed in Chapters 1 and 2, this process has been presented as a search speed-stability paradox in which proteins must search quickly but also bind their target stably[1]. One mechanism by which the search speed of proteins can be enhanced is via facilitated diffusion[2]. Reducing the dimensionality of the search decreases the amount of time required to locate a target. Facilitated diffusion has been directly observed for numerous DBPs, via both *in vitro* and *in vivo* single molecule imaging[3,4]. These studies have revealed insights such as the ability of repair proteins to dislodge roadblock proteins bound to DNA[5], the propensity of the restriction enzyme EcoRV to engage in long-distance hopping events[6], and the utilization of facilitated diffusion by LacI in living bacteria[3]. All of these single molecule studies point towards the use of some form of facilitated diffusion by DBPs to increase the speed of and enhance the efficiency of their target search.

One-dimensional diffusion of DBPs along DNA is at the heart of the facilitated diffusion model for sequence search. Several decades ago, Schurr proposed a model for the one-dimensional diffusion of proteins along DNA wherein DBPs spiraled around the DNA helix as they translated along the DNA, maintaining a constant inward facing

---

*Portions of this chapter were previously published in: L. Cuculis, Z. Abil, H. Zhao, C.M. Schroeder, "TALE Proteins Search DNA Using a Rotationally Decoupled Mechanism", Nature Chemical Biology (2016) *in press*

structure that enabled them to read the DNA sequence[7]. In his model, the DBPs are represented as spheres, and one portion of the sphere remains constantly facing the DNA. The main factors influencing the diffusion coefficient of a protein translating along DNA in this model are the radius, *R,* of the protein and the viscosity, η, of solution. Due to the rotationally coupled translation of the proteins in this model, the dependence of the effective friction on the radius is an R^3 relationship as opposed to solely *R* as in pure translational motion. Bagchi and coworkers, in 2008, described this curvilinear model for protein motion along DNA with slight modifications[8]. They proposed that instead of modeling DBPs as spheres entirely offset from the DNA, that they should be modeled as spheres that partially envelop the DNA. Their model thus contains a value *R* that represents the radius of the protein from its center extending out into solution, and also a value R_{OC}, or the off-center radius. R_{OC}, which is a measure of the distance between the center of the DNA helix and the center of the protein, accounts for the fact that many proteins encompass some portion of the DNA with their DNA-binding site. There have been various studies proposing alternative means of facilitated diffusion for DBPs, including two-dimensional sliding[9]. A key argument against a pure sliding mechanism is the difficulty in bypassing obstacles bound to chromatin *in vivo*, such as nucleosomes[9].

A comprehensive study by Blainey *et al* examined a broad class of DBPs via SMFM with the aim of determining their diffusion trajectories, and whether rotationally coupled sliding could be taken as a generality for DBP motion[10]. In order to do this, they examined how one-dimensional diffusion speed scales with the size of the protein in question. They found that all of the proteins studied appeared to utilize a rotationally coupled sliding mechanism during 1-D search. By fitting both a $1/R$ and $1/R^3$ relation to

all of the measured $D_{1\text{-}D}$ values captured in their assay, they showed that across different proteins, which ranged from the Lac repressor to hOgg1, a $1/R^3$ fit was superior and thus rotationally coupled diffusion persisted. Furthermore, they showed that these proteins enveloped a portion of the DNA during their search, such that their previous model involving both R and R_{OC} provided further refinement of the fit. Experiments with a single protein conjugated to various probe sizes (increasing only R and not R_{OC} in this way) confirmed their finding. The proteins examined by Blainey *et al*, however, are relatively consistent in structure – they have a DNA-binding pocket but do not fully encircle the DNA as a TALE might.

In this work we seek to understand how TALEs reconcile their unique structure with the observed 1-D search behavior and the previously proposed two-state model for sequence search[11]. We modulate the solution conditions of TALEs diffusing along DNA across more than an order of magnitude of ionic strengths and find that TALEs are capable of incredibly fast one-dimensional diffusion, far surpassing the theoretical limit for rotationally coupled motion. We carry out this set of varied ionic strength experiments for a series of TALEs with different CRD lengths, ranging from the NTR-only mutant to a 21.5 repeat TALE. By quantifying the sensitivity of each TALE construct's one-dimensional diffusion to ionic strength, we are able to understand the extent to which their CRDs (which contain positively charge amino acids) are engaged in the search process. In order to elucidate the trajectory of TALEs along DNA, we conjugate a series of different sized fluorescent probes to the 21.5 repeat TALE and quantify the effects of protein-probe size on TALE one dimensional diffusion coefficients. Strikingly, we find that the protein-probe size dependence is best described

by a 1/R (where R is the protein-probe radius) relation as opposed to a $1/R^3$ relation. This finding suggests that TALEs do not utilize a rotationally coupled motion during apparent 1-D diffusion. Finally, we apply hydrodynamic flow to TALEs engaged in sequence search and quantify the flow bias as a product of ionic strength to determine if, and to what extent, TALEs are unbound from DNA during their search. We uncover a surprising result in our ability to trap TALEs against the chemical tethers of the DNA molecules when flow is applied under supraphysiological ionic strength conditions. Taken together, this chapter provides insights leading to an improved model for TALE search in which TALEs adopt a loosely wrapped conformation and rotationally decoupled trajectory during search, a model that is consistent with our previously proposed two-state model for TALE sequence search.

3.2 Materials and Methods

3.2.1 DNA and protein preparation

Dual tethered DNA substrates were prepared as previously described in Chapter 2, as were the TALE 21.5, 15.5, and 11.5 repeat constructs, and the TALE NTR. For experiments involving TALEs conjugated to Qdots and Cy5-labeled streptavidin for protein-probe size diffusion measurements, changes to Chapter 2 protocols are provided below.

3.2.2 Protein labeling with Cy5-streptavidin and Qdot705.

Biotin-functionalized TALEs were generated using the same protocol followed for Cy3 labeling (Chapter 2), albeit with the Cy3 hydrazide being replaced with (+)-Biotinamidohexanoic acid hydrazide (Sigma). Purified TALE-biotin was conjugated to the Cy5-labeled streptavidin by incubation with 2-fold molar excess Cy5-labeled

streptavidin for 15 minutes. Two separate methods were utilized for generation of Qdot705-TALE conjugates. In the first method, streptavidin-coated Qdot705 (Invitrogen) was linked to biotin-TALE (as generated above) by incubation with 5-fold molar excess Qdots for 15 minutes. In the second method, goat anti-mouse Qdot705 conjugates (Invitrogen) were coated with mouse anti-His antibodies (Genscript) in a 1:5 ratio. The resulting Qdot705 anti-His conjugates were incubated in a 1:1 ratio with TALEs for 15 minutes on ice before imaging.

3.2.3 Single molecule imaging assay

The assay developed in Chapter 2 was again utilized for imaging TALE search dynamics. The only deviations occurred when imaging biotin-functionalized TALEs conjugated with Qdot705 or Cy5-labeled streptavidin. In these experiments, 1 μM biotin was added to imaging conditions in order to block any free streptavidin binding sites both on the surface and to prevent multiple proteins from associating with one another. Furthermore, 1 mM methyl viologen (Sigma) and ascorbic acid (Sigma) were included in order to preserve photostability of these red shifted fluorescent probes.

3.2.4 Determination of theoretical limits for 1-D diffusion

The two prevailing models for 1-D diffusion of proteins along DNA consider the following effects: rotationally coupled and non-rotationally coupled (linear) motion[7,8]. In the case of non-rotationally coupled motion, the diffusion coefficient for a protein translating along DNA is given by:

$$D_{linear} = k_B T \left(\frac{1}{\xi_{trans}} \right) \tag{3.1}$$

where ξ_{trans} is the translational friction given as (3.2):

$$\xi_{trans} = 6R\pi\eta \tag{3.2}$$

Here, R is the radius of the protein+probe complex and η is the viscosity of solution.

In the case of rotationally coupled 1-D diffusion, we adopt a model developed by Bagchi and coworkers for a protein undergoing curvilinear motion along DNA, which for our system is identical for Schurr's original model[7,8]. Here, the diffusion coefficient is given by (3.3):

$$D_{linear} = k_B T \left(\frac{1}{\xi_{trans} + \xi_{rot}} \right) \tag{3.3}$$

where ξ_{rot} is the rotational friction given by (3.4):

$$\xi_{rot} = 6\pi\eta R + \left(\frac{2\pi}{10BP} \right)^2 \left[8\pi\eta^3 + 6\pi\eta \; R(R_{OC})^2 \right] \tag{3.4}$$

Here, BP is the distance between DNA base pairs, R is the radius of the protein-probe complex and R_{OC} is the distance between the center of the DNA axis and the center of the protein-probe complex. Due to the apparent wrapped helical conformation of the TALE complexes, we take R and R_{OC} to be equivalent.

We estimate the radii of the TALE proteins from previous structural characterizations carried out via NMR, SAXS and light scattering measurements[12]. The radii of the Qdot antibody conjugates were measured via scanning electron microscopy. The radii of streptavidin tetramers have been previously reported[13].

3.2.5 Hydrodynamic flow assay for TALE motion

In order to determine whether convective or diffusive behavior would dominate at the location of the DNA-bound TALEs during flow experiments, we utilized particle imaging velocimetry (PIV) to accurately measure the flow speed. Fluid flow is driven via a computer-controlled syringe pump operating at 150 uL/min. A solution of fluorescent

polystyrene microbeads (400 nm radius) was introduced into the microfluidic flow cell, and the particle velocities at a distance of ~50-100 nm from the flow surface were measured. Using the experimentally determined fluid velocity, the Peclet number *Pe* at the relevant length scale was determined according to (3.5).

$$Pe = \frac{Lu}{D} \tag{3.5}$$

where *L* is the characteristic length, *D* is the mass diffusion coefficient, and *u* is the local flow velocity. Given the estimate for *D*, determined via the approximate radius of the TALEs discussed above, *Pe* = 10 for the hydrodynamic flow experiments conducted in this work.

In order to quantify bias introduced by flow at lower salt (90-150 mM KCl) conditions, we calculated the sample skewness of the frame-by-frame displacements of TALEs (the step sizes as we define these displacements)[14]. We define skewness according to the sample skewness equation given by (3.6).

$$skew = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^{n}(X_i - X_{avg})^3}{\sigma^3} \tag{3.6}$$

Where *n* is the number of steps (frame-by-frame displacements recorded), X_i and X_{avg} are the individual and average step values, and σ is the standard deviation of the sample. In order to determine if the skewness present in sample sets is significant, we compare values of *skew* to the standard error of skewness (*SES*) given by (3.7).

$$SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \tag{3.7}$$

Where *n* again is the number of steps in the sample set.

3.3 Results and Discussion

3.3.1 Salt dependence of TALE search reveals details of structural conformation during search

In previous work (Chapter 2) we observed an ionic strength dependence for TALE one-dimensional diffusion wherein an increase in ionic strength yielded an apparent increase in $D_{1\text{-}D}$. This led to postulation that TALEs utilize, in part, micro-dissociation/association events (hopping events) during their search. The same dependence was not observed for the TALE NTR. We found that we could extend our previous results to much higher ionic strengths by first nucleating TALEs on dual tethered DNA templates at moderate protein concentrations and ionic strengths (~500 pM TALE, 50-90 mM KCl in imaging buffer) and then exchanging this standard imaging buffer for much a higher salt imaging buffer (200-1000 mM KCl). While a portion of TALEs, presumably those only transiently/loosely associated, were removed from the DNA substrates upon exchange into high salt concentrations, a subset of TALEs remained bound and began diffusing extremely rapidly along the DNA (**Figure 3.1**). We quantified the increase in diffusion speed due to increased salt concentration for our previously generated array of TALEs (11.5, 15.5, and 21.5 TALE repeats) across an order of magnitude (50-500 mM KCl) of ionic strengths (**Figures 3.2 and 3.3**) using the covariance based estimator approach[15] described in Chapter 2.

In contrast to many DNA-binding proteins, we found that TALE 1-D diffusion is extremely sensitive to ionic strength and thus charge screening effects, as demonstrated by diffusion trajectories for a TALE protein with 21.5 repeats in monovalent salt concentrations ranging from 0.05 to 0.5M KCl (**Figure 3.4**). We systematically

quantified TALE diffusion as a function of salt concentration for TALEs ranging in length from 0 to 21.5 repeats, where zero repeats corresponds to the TALE NTR-only construct (**Figure 3.4**)[16]. Well-accepted models for facilitated search of DNA-binding proteins classify protein search behavior as either hopping or sliding, depending on whether 1-D diffusion speeds are salt-sensitive or invariant[17]. The salt sensitivity of our data points toward a TALE search mechanism dominated by hopping behavior. The structure of target DNA-bound TALEs[18] however, clearly shows that TALEs track the major groove during specific binding. Based on these observations, we sought to reconcile how a DNA 'wrapped' TALE could follow a search mechanism involving hopping behavior, and we utilized single molecule techniques to elucidate the search process.

In order to further probe the search mechanism, we determined the change in apparent 1-D diffusion with respect to salt concentration dlog*D*/dlog[salt] for TALEs ranging in length from 21.5 to 0 repeats (**Figure 3.5**). The quantity dlog*D*/dlog[salt] provides a measure of the salt dependence of TALE 1-D diffusion. Previous single molecule studies of DNA-binding proteins have used this quantity to roughly estimate the number of charged residues contacting the DNA backbone during non-specific search[19]. Moreover, the dependence of protein binding affinity on salt concentration (dlog(K_a)/dlog[salt]) has been used in the context of counterion condensation theory to provide an estimate for the number of charged residues contacting the DNA backbone during specific binding of proteins to DNA[20,21]. In this way, the ionic strength dependence of TALE protein diffusion (dlog*D*/dlog[salt]) serves as a proxy to describe the number of charged residues engaged in the search process, thereby elucidating the role of TALE CRD repeats on non-

specific search. If the residues within the TALE repeat region (CRD) are not involved in the search process, then there would be a negligible change in dlog*D*/dlog[salt] with a concurrent change in the number of repeats within the TALE CRD. On the other hand, if TALE repeats are involved in non-specific search, then an increase in the number of charged residues via longer CRDs would induce more sensitive changes in ionic strength dependence, thereby resulting in larger values of dlog*D*/dlog[salt]. Our results clearly show that there is a nearly linear increase in the magnitude of dlog*D*/dlog[salt] upon increasing the CRD size (**Figure 3.5**). These results support a model in which TALEs are wrapped around or intimately associated with DNA during non-specific search.

Despite the insights provided by the ionic strength modulation, we were unable to reconcile our findings with the traditional binary classifications used for protein search on DNA. Traditionally, DNA-binding protein search mechanisms are classified as either sliding or hopping. As discussed in Chapter 2, search mechanisms are classified as pure sliding mechanisms if there is no significant change in their apparent one dimensional diffusion speed as the ionic strength of solution is increased[22]. Conversely, search mechanisms are classified as hopping mechanisms if there is a clear dependence of one-dimensional diffusion speed on salt concentration. The theoretical basis for this strict binary classification arises from the argument regarding energetics of ion displacement from the DNA backbone. By definition, a sliding protein tracks the DNA backbone tightly and remains closely associated with the DNA, such that ions displaced by the protein translocating in one direction are balanced by ions that re-associate with the DNA at the location immediately vacated by the diffusing protein[23]. This view has the protein-DNA contacts maintained with high integrity, assuming no ions can come between DNA

and protein (to screen the DNA backbone charges) during its search, and assuming that no part of the protein even temporarily dissociates from the DNA during search. A hopping protein, on the other hand, completely dissociates from the DNA template it is searching along, such that all interactions between protein and DNA are broken, and diffuses far enough away from the DNA that ions are able to re-condense around the DNA backbone. In this way, there is a higher energy cost associated with a protein re-binding at higher ionic strengths. This causes the protein to spend more time in the unbound state, and thus more time diffusing at a rate of diffusion given by $D_{3\text{-}D}$, which is more than an order of magnitude greater than $D_{1\text{-}D}$. As a result, a hopping protein will appear to speed up at higher ionic strengths. This strict binary classification of events does not allow for a model in which TALEs are wrapped around the DNA during search, a conformation that precludes hopping behavior, but still demonstrate a strong dependence of diffusion speed on salt concentration (classically defined as hopping). Thus, we designed further single molecule experiments in order to better understand how TALEs could contradict the principles of a strict sliding/hopping search classification.

3.3.2 Flow bias reports on TALE association with DNA during search

We next sought to reconcile how TALE diffusion could be described by hopping behavior. Hopping behavior is commonly associated with rapid dissociation/re-association events in which the protein fully dissociates from the DNA and then re-binds a short distance from the site of dissociation[22,24], however, this mechanism appears to be inconsistent with the superhelical structure of TALE proteins and our finding that the CRD repeats engage DNA during search. To study this further, we utilized hydrodynamic flow to determine whether TALE diffusion can be directionally biased during non-

specific search. We applied laminar fluid flow in microfluidic channels containing a field of dual-tethered DNA templates in the presence of fluorescently labeled TALEs (**Figures 3.6** and **3.7**). In the vicinity of the surface, TALEs bound to DNA experience a simple shear flow with a flow rate of ~25 μm/s (**Figure 3.6**). Interestingly, individual TALE proteins bound to DNA are 'pushed' to the distal end of DNA templates under the action of flow in high salt conditions (500 mM KCl) (**Figure 3.7**). Strikingly, TALEs remain bound to DNA and do not dissociate while being pushed along DNA templates in flow. Upon stopping the flow, TALEs bound to DNA immediately resume directionally unbiased 1-D diffusion, and this process can be repeated for multiple iterations for the same molecule (**Figure 3.8**). This behavior was observed for all TALEs except for NTR-only mutants that lacked a CRD (**Figures 3.8 and 3.9**). We found that the TALE NTR rapidly dissociated from the DNA template upon application of fluid flow and could not be 'pushed' along DNA using flow.

Using the hydrodynamic flow assay, we observed occasional unbinding and release of the TALE construct containing 11.5 repeats (**Figure 3.9**, 6 out of 18 observed events), however, our experiments showed no evidence of flow-induced release of TALEs containing 15.5 and 21.5 repeats (0 out of 17 and 0 out of 20 events, respectively). These observations are consistent with a model in which TALEs are wrapped around DNA in a helical conformation during non-specific search, albeit not tightly tracking the major groove during search. It is known that TALEs bind to DNA in a one-repeat-to-one-nucleotide manner, at least in the context of specific binding. From this perspective, the TALE construct containing 11.5 repeats is just long enough to completely encircle DNA within one helical turn of the protein. Therefore, the 11.5 repeat construct shows the

propensity to be pushed along DNA templates in the presence of flow, albeit with a finite probability for unbinding and release from DNA templates during the experiment, which was not observed for TALEs with larger CRDs.

We further investigated the impact of fluid flow on TALE dynamics by plotting the average displacement as a function of time for ensembles of single TALE trajectories at low ionic strength (90 and 150 mM KCl), in the presence and absence of flow (**Figure 3.10a**). Additionally, we quantified the presence of diffusional bias by determining the sample skewness of step size distributions (**Figures 3.10** and **3.11**). In the absence of flow, TALE motion is directionally unbiased (**Figure 3.10b**, skew of 0.015 ± 0.068). Similarly, at low ionic strength (90 mM KCl) under applied flow TALEs display no significant skew (**Figure 3.11**, skew of 0.010 ± 0.040). In the presence of flow at physiological ionic strength (150 mM KCl), however, TALE diffusion is clearly biased in the direction of flow (**Figure 3.10c**, skew of 0.64 ± 0.048). Under these conditions, convection dominates diffusion for unbound TALEs in solution near the surface, which can be quantified by the protein Peclet number (Pe = 10) for unbound TALEs in flow (**Figure 3.6**). For a model in which the protein spends some amount of time unbound from the DNA during its search, a convective flow would directionally bias the apparent 1-D diffusion in proportion to the amount of time spent unbound from the DNA, which is consistent with our experiments[25]. Conversely, if the protein remained tightly coupled to DNA during search, then there would not be a significant bias to its search process. Our findings from flow experiments support a search process in which TALEs remain wrapped around DNA, yet they are still able to effectively dissociate when the ionic contacts between positively charged residues and the DNA phosphate groups are

temporarily broken. In this way, the TALEs are spatially confined to DNA templates but are able to 'hop' along DNA during 1-D diffusion.

3.3.3 Probe size effects on $D_{1\text{-}D}$ provide direct evidence that TALEs utilize a rotationally uncoupled search mechanism

We next sought to use fluorescent probes of various sizes conjugated to the TALEs in order to determine the microscopic trajectories of TALEs as they are diffusing, in a wrapped conformation partially associated with DNA. More specifically, we sought to verify that TALEs were in fact not following a rotationally coupled trajectory, which is the mechanism used by most DBPs previously studied via SMFM. Modulating the size of the fluorescent probe attached to diffusing proteins is a strategy employed in previous works[4,10,19,26] in order to determine whether a protein tracks the DNA major groove or diffuses in a rotationally uncoupled mechanism. Our initial labeling strategy, using a genetically encoded aldehyde tag with a covalently attached small molecule fluorescent dye, was easily amended to include different probe sizes (**Figure 3.12**). By reacting biotin-hydrazine with aldehyde-tagged TALEs, we generated biotinylated TALEs that we then conjugated to both Cy5-labeled streptavidin and quantum dots (**Figure 3.13**) functionalized with streptavidin. In order to ensure that biotin labeleing schemes were not introducing artifacts into measurements, we also labeled TALEs via primary-secondary antibody conjugation. Here, we decorated primary antibody-coated quantum dots with secondary anti-His antibodies, which then enabled direct conjugation to TALEs that contain an N-terminal hexahistidine tag. These probes provided an additional ~6 and 12 nm of size, respectively, and thus added hydrodynamic drag to diffusing TALEs (**Figure 3.12**). We measured $D_{1\text{-}D}$ distributions for these newly labeled TALEs and

found that, indeed, increasing the size of the probe leads to a decrease in recorded one-dimensional diffusion speed (**Figure 3.14**). The decrease in diffusion speed, however, was more consistent with a rotationally uncoupled trajectory than one in which TALEs tightly track the DNA major groove. Plotting $D_{1\text{-}D}$ versus the inverse or inverse cube of the radii of the protein plus probe, the data were much better fit to a 1/r scaling than a than $1/r^3$ scaling (**Figure 3.15**). Thus, according to the models laid out by Schurr and later by Bagchi *et al*, TALEs utilize a primarily rotation-free trajectory as they diffuse along DNA. Furthermore, this trend persisted at higher ionic strengths (**Figure 3.16**), with a 6-fold decrease between TALE-Cy3 and TALE-Qdot diffusion at 90 mM KCl and only a 2-fold decrease at 500 mM KCl. In this way, TALEs appear to translocate more and more out of phase with the DNA helix as ionic strength is increased.

3.3.4 Comparison of TALEs to the broad class of DNA binding proteins suggests a unique search process

We further compared the 1-D diffusion behavior for TALE proteins to a variety of DNA-binding proteins that are known to undergo target site search along DNA. First, TALE diffusion at high ionic strength exhibits extremely rapid 1-D search with large diffusion coefficients (**Figure 3.16**). In fact, the Qdot705-labeled TALEs exceed the theoretical 1-D rotationally coupled diffusion coefficients by nearly 200-fold at high ionic strength. These results further show that at high ionic strength, TALEs readily exceed the largest reported 1-D diffusion coefficients for thermally-driven DNA-binding proteins[27,28], which underscores the unique nature of the TALE search process. We compared the diffusive behavior of TALEs across a range of salt concentrations (0.09 to 1.0 M KCl) to several other DNA binding proteins (under conditions ranging from 0.002

to 0.025 M monovalent salt) undergoing non-specific search (**Fig. 3.17**) previously investigated by single molecule fluorescence microscopy[10,27]. The vast majority of DNA-binding proteins previously studied diffuse along DNA with an approximately $1/r^3$ scaling, however, TALEs behave quite differently. Taken together, these data suggest that TALEs carry out their search loosely wrapped around DNA, without fully tracking the DNA helix (**Figure 3.18**), which is in stark contrast to a broad class of transcription factors and DNA-binding proteins that have been shown to rotate along the helical path of the DNA[29,30]. In this picture, TALEs adopt a loose helical conformation, wrapped around the DNA template but only briefly contacting the phosphate backbone (**Figure 3.18**). Searching in this manner, TALEs are then able to compress along their superhelical axis in order to check the local DNA sequence (**Figure 3.18**).

In recent years, researchers have probed the ionic strength dependence of 1-D diffusion rates to gain insight into the search mechanisms of DNA-binding proteins. In particular, the observation of an ionic strength-dependent diffusion rate is generally cited as evidence of a hopping mechanism for DNA search, whereas the absence of salt-dependent diffusion is generally considered as evidence supporting a sliding mechanism. From a mechanistic perspective, protein hopping involves continual cycles of protein binding and unbinding to DNA templates during the search process. From this view, hopping is thought to be salt dependent because the binding of the protein after a 'hop' is strongly impacted by the surrounding ion cloud and the number of ions that must be displaced in order for binding to occur. In the context of this model, an increase in ionic strength will decrease the affinity of the protein for the DNA template, thereby increasing the fraction of time the protein spends in free solution and increasing the apparent 1-D

diffusion speed. Conversely, sliding behavior is thought to be independent of ionic strength, because the displacement of ions during sliding along the DNA helix is offset by the rebinding of ions at the site previously occupied by the protein. From a mechanistic perspective, how can a single protein exhibit evidence of apparent hopping and sliding during non-specific search along DNA? Interestingly, a few recent studies have considered protein search behavior that can be described by both of these ideal mechanistic pictures[29,31–33]. Recent experimental[32] and molecular dynamics[33] studies have suggested that ionic strength dependence may support a quasi-sliding mechanism that is not well described by the strict classifications of sliding and hopping behavior. Here, we explore how the TALE search process cannot be described by the classical definitions of ideal hopping or sliding, rather, TALE search is described by a hybrid of the two models.

Our data clearly demonstrate a strong ionic strength dependence of TALE diffusion, however, flow-based single molecule experiments show that TALEs remain associated with DNA templates in the presence of strong convective flow during long-range search. These results cannot be reconciled within the classic hopping mechanism for non-specific search. Moreover, the probe-size scaling of 1-D diffusion coefficients for TALEs appear to argue against a sliding mechanism for TALE search, wherein TALEs would closely track DNA in a rotationally coupled trajectory. Taken together, our data for TALE search cannot be described by the traditional hopping or sliding models for non-specific search along DNA. Nevertheless, the results for TALE protein search should be considered in the context of a superhelical protein structure that effectively wraps the protein around the DNA double helix. From this view, our results on the salt sensitivity of TALE CRD size suggest that the residues within the CRD are directly impacted by charge screening

and electrostatic interactions in the system. Our data are consistent with a superhelical TALE protein loosely wrapped around a DNA template, albeit without the tight threading associated with major groove binding. In this model, TALEs are effectively wrapped around DNA templates, and electrostatic interactions between individual residues in CRD repeats and DNA phosphate groups can be broken and reformed depending on the ionic strength of solution. In the limit of high ionic strength, several of these interactions may be temporarily disrupted, thereby resulting in an effectively loosely bound TALE protein nevertheless associated with a DNA template due to the superhelical protein conformation. From a broad view, the solution ionic strength in part determines the strength of TALE residue interactions with the DNA backbone, and thus in effect acts to smooth the energetic barriers to 1-D diffusion. Finally, our results are consistent with prior studies suggesting commonalities between the sliding and hopping models of protein search[31–33]. In these studies, it is argued that as electrostatic interactions between DNA and protein are weakened by increased ionic strength, the energetic barrier to sliding decreases and thus apparent sliding speeds up. This argument is likely also valid for TALEs, which appear to be wrapped around DNA during search and mediate electrostatic interactions with DNA that are susceptible to screening.

Rotationally decoupled search has been previously reported for other DNA binding proteins undergoing non-specific search, albeit in the context of vastly different protein function[19,26,34]. The eukaryotic proliferating cell nuclear antigen (PCNA) forms a ring-like structure around DNA and serves as a processivity factor for ε and δ polymerases among other functions, none of which, however, requires sequence-specific binding. Single molecule investigation of PCNA dynamics revealed that this protein alternates

between rotationally coupled and decoupled 1-D diffusion, described by a rapid ($D_{1\text{-}D}$ values of 10 Mbp^2/s) rotationally decoupled sliding/1-D hopping mechanism[19]. The eukaryotic mismatch repair (MMR) protein MutSα undergoes a conformational switch upon mismatch recognition and subsequent ATP-mediated release from mismatch sites[34]. Interestingly, the conformational change transitions MutSα from a rotationally coupled 1-D trajectory along DNA to a rotationally decoupled trajectory. Similar behavior was also observed for Taq MutS[25]. Moreover, the type III restriction enzyme EcoP15I was also observed to diffuse rapidly in a rotationally decoupled trajectory ($D_{1\text{-}D}$ values of 8 Mbp^2/s) along extended DNA templates following target site binding and a subsequent ATP-driven conformational switch[26]. The common theme amongst the aforementioned DNA binding proteins is the absence of sequence-specific binding after transitioning to rotationally decoupled diffusive paths, unlike TALEs, which appear to utilize this mechanism to locate target sites for sequence-specific binding.

Tumor suppressor p53 was recently studied via molecular dynamics simulations and bears some similarities to TALEs in terms of search mechanism. Simulations of non-specific search for p53 on DNA under physiological ionic strength conditions revealed rotationally decoupled trajectories mediated by the protein C-terminal domain (CTD), with tethered hopping events initiated by the core domain[33]. Nevertheless, despite the apparent similarities to the TALE search process, the p53 core domain (and not the p53 CTD) is responsible for sequence-specific binding. In this way, TALEs appear to be unique in that the protein domains that are responsible for sequence-specific binding follow a rotationally decoupled trajectory during target search. However, although rotationally de-coupled diffusion (which can be described as 1-D hopping[34] and 2-D

sliding[9]) allows for a more rapid 1-D search compared to rotationally-coupled diffusion, this search mechanism intrinsically results in the protein being situated out-of-phase with respect to the helical pitch of DNA double helix, thereby hindering the ability of the protein to read the local sequence. In our previous work, we observed that TALEs utilize a two-state model for sequence search wherein we conjectured that TALEs compress along their helical axis during a specific binding event, which enables them to 'check' the local sequence[11]. This model is consistent with a rotationally decoupled search mechanism and allows TALEs to retain their sequence-specific binding activity while enabling rapid translocation interspersed between sequence 'checking' events, thereby satisfying the search speed-stability paradox[1].

Our results support a model wherein TALEs adopt a loosely wrapped conformation around DNA templates during non-specific sequence. Based on these findings, the design space of TALE fusion proteins could potentially be expanded given the robust binding and search abilities of TALEs. For example, it might be possible to generate chimeric proteins by fusing TALEs with larger 'payloads' or active domains. Finally, the rotationally decoupled model is consistent with a two-state model for the TALE search process[11]. During the search mode, TALEs are loosely wrapped around the DNA helix, which mediates rapid 1-D non-specific search. During the recognition mode, TALEs are readily able to compress their superhelical pitch, which has been previously proposed as a mechanism of target site binding[35,36]. Upon encountering a putative target site, TALEs can undergo a conformational change to compress and 'check' the local sequence, possibly resulting in stable binding. Additional molecular-based studies with higher

spatial resolution will permit a dynamic view of the detailed conformational changes occurring during transitions between these modes.

3.4 Concluding Remarks

Proteins bound nonspecifically to DNA are inherently challenging for crystallographers to capture due to the transient nature of these interactions, and the multitude of intermediates that persist between an unbound and specifically bound protein. Obtaining structural information of these complexes, therefore, requires alternative approaches such as circular dichroism or NMR. Limitations on the size of protein that can be studied (NMR) and the granularity of structural detail (circular dichrois) obtained from these techniques, however, leave both ill suited to probe a complex as large as a full-length TALE protein (more than a thousand residues in length) bound nonspecifically to DNA. Molecular dynamics simulations provide great structural detail, however fully atomistic simulations over appreciable time scales (greater than nanoseconds) remain computationally intensive and largely unattainable for proteins such as TALEs. Single molecule imaging techniques strike a balance between spatial and temporal resolutions, however, as they offer a view of molecular level dynamics over long time scales. In this work we utilize single molecule techniques to provide several insights regarding the conformation and dynamics of TALEs as they scan non-specific DNA in search of their target. This work builds on our previous single molecule studies of TALE dynamics (see Chapter 2) and lends further evidence to the search and check two-state mechanism that was proposed in our initial work.

We were intrigued by the sensitivity of TALE diffusion to ionic strength and probed this further. Since the work of Winter *et al* in 1981[37], diffusion that is salt-

sensitive has been classified as hopping, while invariance to salt has been classified as 1-D sliding. Hopping is defined as complete dissociation of the protein from the DNA and subsequent rebinding after a short diffusive (3-D) excursion. Hopping is said to be salt sensitive, as there is an energetic barrier to re-binding when counter ions associated with the DNA must be displaced, and thus higher ionic strengths will elicit less frequent rebinding events. Sliding is defined as the DBP strictly tracking the DNA helix, remaining completely bound during search. Sliding is said to be salt invariant as counter ions that must be displaced as the DBP tracks along the DNA are able to re-condense behind the protein, thus there is no net change in This strict binary classification of DNA binding protein (DBP) sequence search is perhaps easy to rationalize for DBPs that have a small contact area with DNA, when complete dissociation/re-association events are feasible on short timescales. In this work we show that TALE search cannot be classified using the strict sliding/hopping dichotomy. Instead, our results suggest a mechanism in which TALEs completely encircle DNA, precluding hopping behavior. We further support this hypothesis by demonstrating that TALEs can be trapped against DNA surface linkages by an applied flow. Quantification of flow bias experiments suggests TALEs, in this wrapped conformation, disengage from direct molecular contact with DNA at a frequency modulated by solution ionic strength, while remaining constrained to the DNA like a washer on a string. Finally, our findings from probe size scaling measurements suggest that TALEs assume a primarily rotation-free trajectory as they scan DNA. Taken together, the results of this chapter highlight the unique nature of TALE search, and show how their helical structure facilitates the search process. Indeed, a wrapped conformation during non-specific search allows for TALEs to compress along

their superhelical axis to check local sequences for their target, satisfying the search speed-stability paradox as laid out by our two state search model in Chapter 2.

3.5 Figures

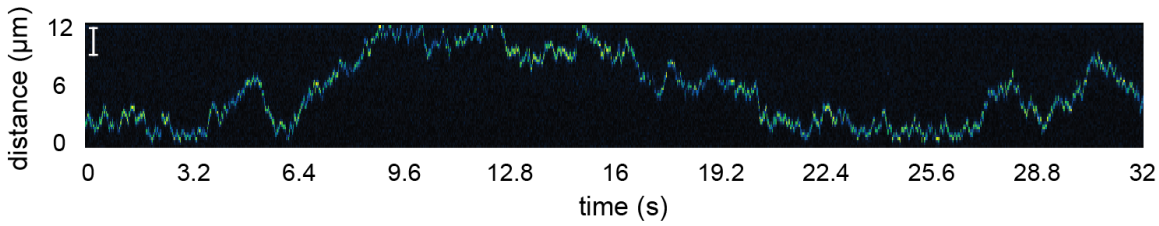

Figure 3.1 Kymograph of Cy3-labeled 21.5 repeat TALE construct diffusing one-dimensionally at supraphysiological ionic strength (500 mM KCl). In this searching event, the TALE scans over nearly the entire 44.5kbp substrate in search of its target, covering hundreds of thousands of total base pairs in only 30 seconds. The scale bar is equivalent to 10kbp.

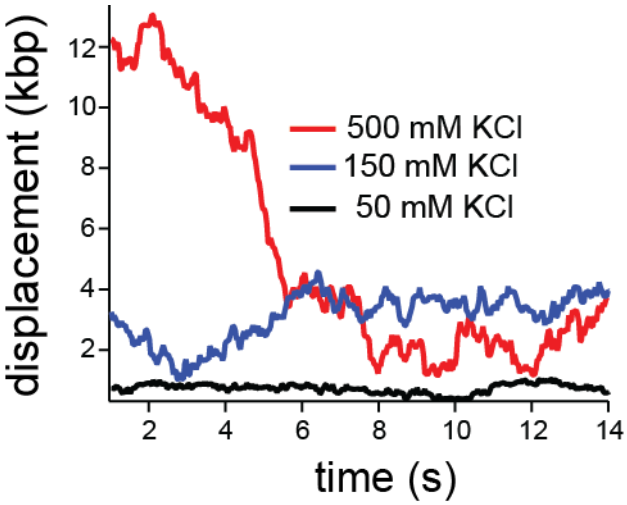

Figure 3.2 Diffusion of the Cy3-labeled 21.5 repeat TALE construct at a range of salt concentrations. Representative traces of individual diffusing TALEs at 50 (black), 150 (blue), and 500 (red) mM KCl highlight the strong salt dependence of TALE diffusion.

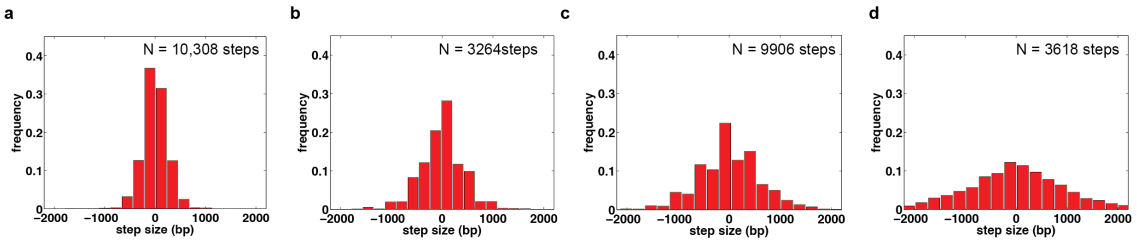

Figure 3.3 Step size distributions for the 21.5 repeat TALE construct at increasing ionic strength. Histograms at (a) 30, (b) 90, (c) 300, and (d) 500 mM illustrate the intense salt sensitivity of TALE diffusion.

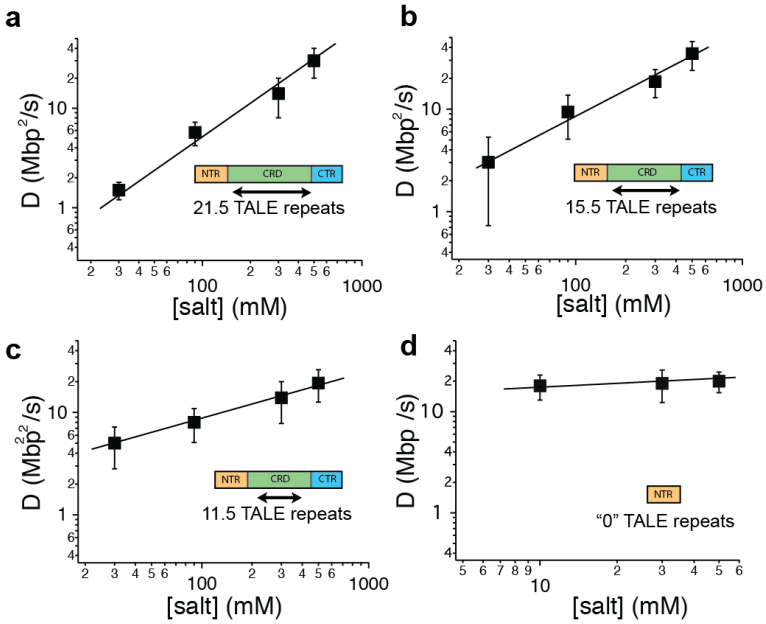

Figure 3.4 Salt dependence of TALE diffusion for the different TALE constructs investigated. The change in one dimensional diffusion coefficient (D) versus the salt concentration of the imaging solution ([salt]) is plotted for the (a) 21.5 repeat, (b) 15.5 repeat, (c) 11.5 repeat, and (d) NTR-only ("0" repeat) constructs. Each data point represents a diffusion constant derived from step size distributions of 12 to 40 individual molecular trajectories.

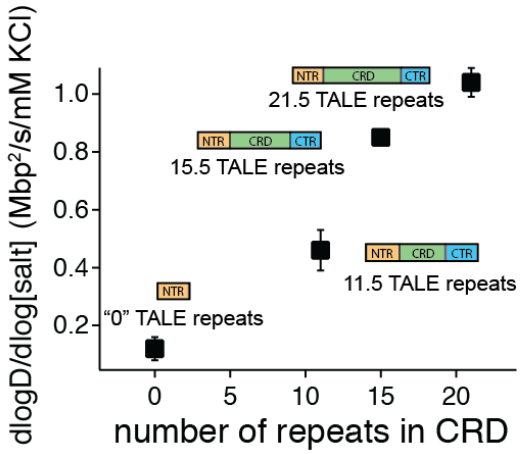

Figure 3.5 Relationship between TALE salt sensitivity and the length of the TALE CRD. Values of dlogD/dlog[salt] for each TALE construct are plotted versus the number of repeats in the CRD, showing a nearly linear dependence and suggesting that the repeats are in contact with the DNA phosphate backbone during search.

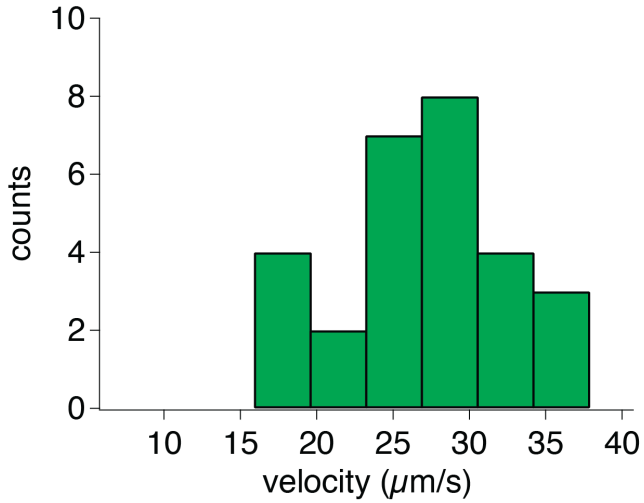

Figure 3.6 Characterization of hydrodynamic flow applied to samples via the syringe pump. Particle imaging velocimetry (PIV) was utilized to measure the flow velocity at the microfluidic sample cell surface, and yielded an average flow of 25.6 ± 4.8 μm/s.

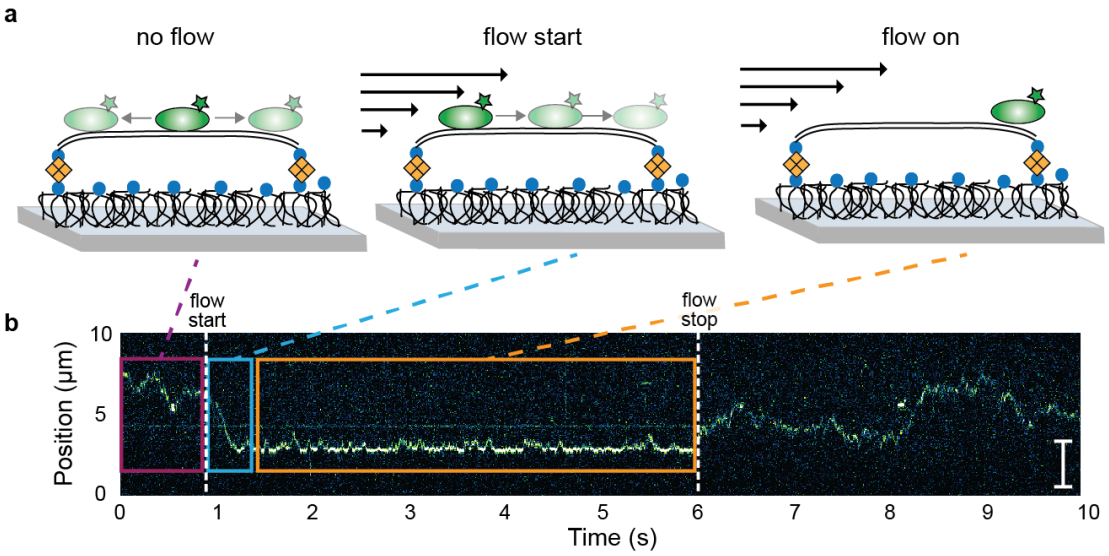

Figure 3.7 Overview of TALE flow bias experiments. (a) Cartoon view of events happening at flow cell surface during bias experiments correlated to (b) a kymograph tracking an individual 21.5 repeat Cy3-labeled TALE construct during the experiment. Initially, TALEs are allowed to diffuse in the absence of flow (purple box), carrying out their directionally unbiased Brownian search. Flow is then introduced via the attached syringe pump, and TALE 1-D diffusion is biased in the direction of flow (green box). At supraphysiological ionic strength (>300 mM KCl), TALEs are trapped against the chemical tethers of the DNA (orange box) by the applied flow. Upon cessation of flow, TALEs are then able to resume their Brownian search. The scale bar is equivalent to 10kbp.

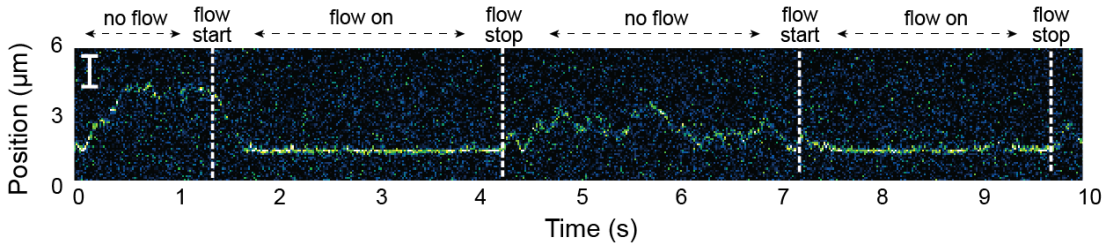

Figure 3.8 Kymograph depicting flow bias experiment with the Cy3-labeled 15.5 repeat TALE construct. In this experiment, which probed the behavior of the 15.5 repeat TALE construct at 500 mM KCl, the searching TALE is trapped against the DNA barrier by applied flow, released via cessation of flow, and then trapped once more before finally being released a second time. The scale bar is equivalent to 5 kbp.

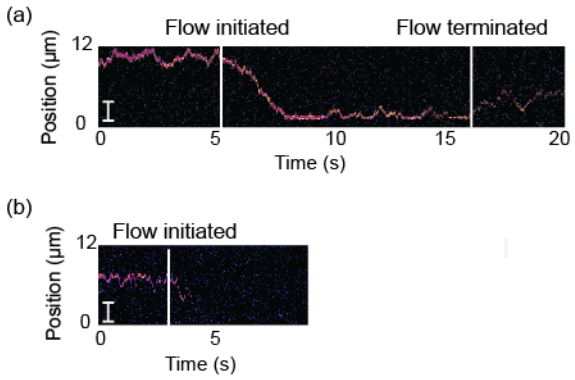

Figure 3.9 Kymograph depicting flow bias experiment with the 15.5 repeat TALE construct. In these experiments, conducted under 500 mM KCl, the searching TALEs are pushed against the chemical linkers of the DNA (a) or ejected from the DNA (b). While most molecules could be trapped as in (a), there were a small proportion of events observed in which flow ejected the TALE11.5. The scale bar is equivalent to 10kbp.

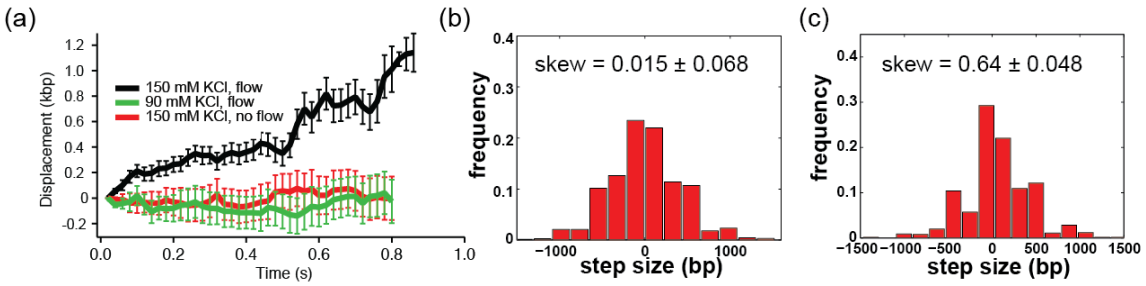

Figure 3.10 Quantification of flow bias introduced in TALE search at physiological ionic strength. (a) Average displacement of several trajectories for 21.5 repeat TALEs diffusing at 90 mM KCl in the presence of flow (green, N = 36 trajectories), 150 mM KCl in the absence of flow (red, N = 31 trajectories) and 150 mM KCl in the presence of flow (black, N = 29 trajectories). Error bars represent the standard error in the mean (SEM). (b) and (c) Distributions of TALE step sizes at 150 mM KCl in (a) the absence (N = 3218) and (b) the presence (N = 2734) of applied flow. Skew values are reported as the sample skew plus or minus the standard error of the skew (SES).

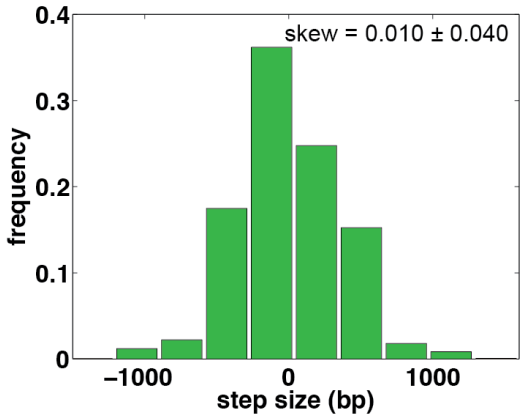

Figure 3.11 Quantification of flow bias introduced in 21.5 repeat TALE construct search at 90 mM KCl. Here N = 3687 steps, and the skew is reported as the sample skew plus or minus the standard error in the skew.

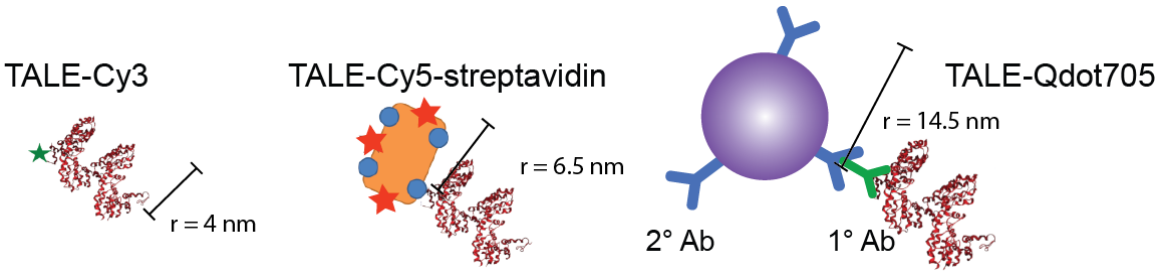

Figure 3.12 Estimates of the TALE-probe radii of various conjugates used during probe size scaling experiments. The Cy3 labeled 21.5 repeat TALE construct radius is conservatively estimated to be 4 nm, based upon values of 4.5 nm obtained via dynamic light scattering (DLS) and small angle x-ray scattering (SAXS) measurements from work by Murakami et al[12]. The well-characterized radius of streptavidin (2.5 nm)[13,38] is added to this value to obtain an estimate for the TALE-Cy5-streptavidin radius. The radii of Qdot705 conjugates were determined via scanning electron microscopy (**Figure 3.13**), which provided an average Qdot length of 12.5 nm.

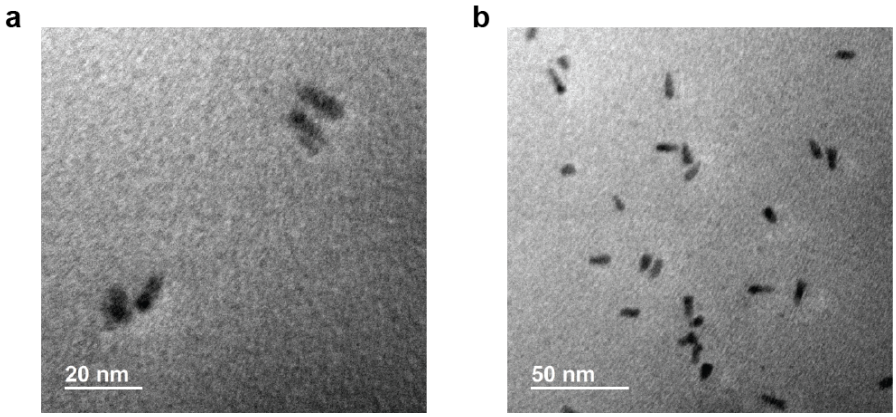

Figure 3.13 Transmission electron micrographs of Qdot705s used in probe size-scaling experiments. The Qdots are ovular in shape, with an aspect ratio of ~ 1.7. The average length is 12.5 ± 1.5 nm, and does not take into account the primary/secondary antibody length scales.

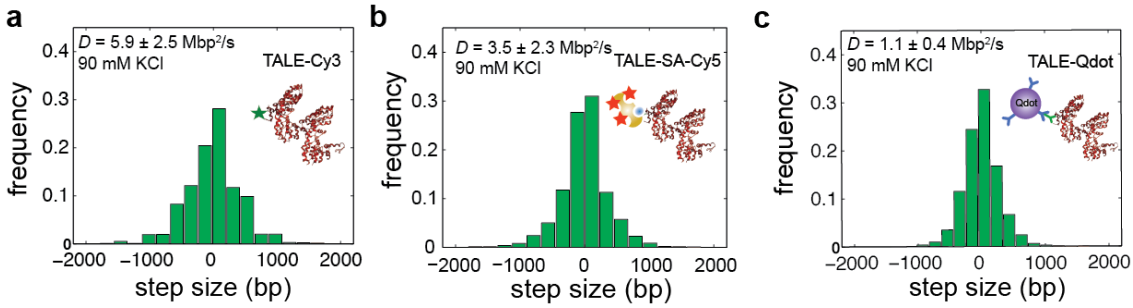

Figure 3.14 Distribution of step sizes for the 21.5 repeat TALE construct conjugated with different fluorescent probes. Distributions for (a) Cy3-labeled, (b) Cy5-streptavidin, and (c) Qdot705-labeled 21.5 repeat TALE construct are shown, taken at 90 mM KCl.

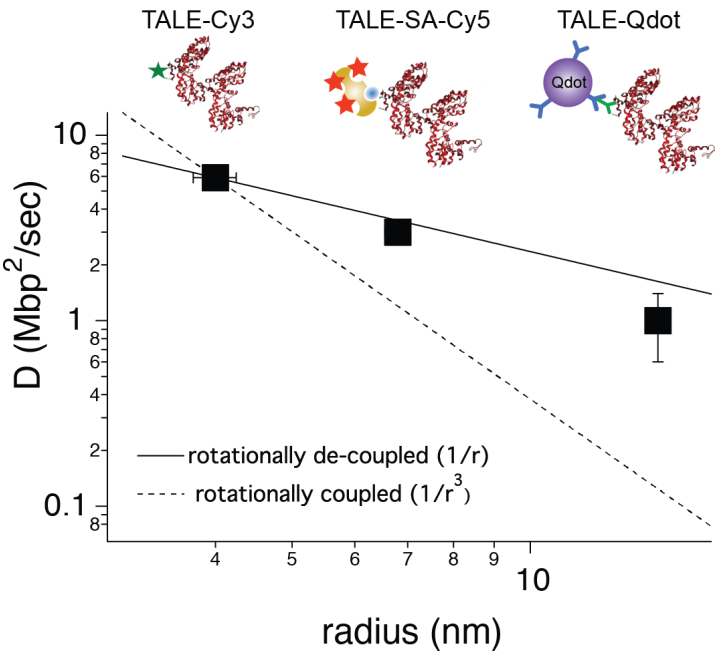

Figure 3.15 Probe-size scaling of 21.5 repeat TALE 1-D diffusion measured at 90 mM KCl. We systematically varied the size of the probe molecule (radius given by r) used to localize diffusing TALEs and measured $D_{1\text{-}D}$. While the vast majority of DBPs studied follow a roughly $1/r^3$ size scaling, we see that TALEs more closely follow a 1/r scaling, characteristic of a rotationally de-coupled 1-D diffusive trajectory.

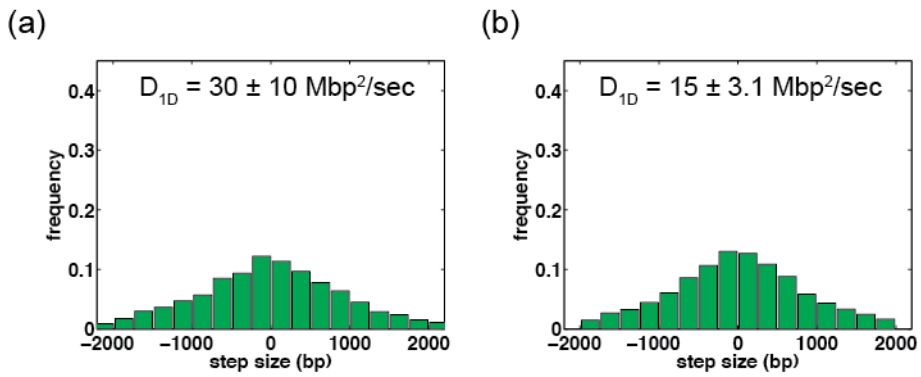

Figure 3.16 TALE diffusion at supraphysiologial ionic strength. (a) Distribution of step sizes of the 21.5 repeat TALE construct labeled with (a) Cy3 (N = 3618) and (b) Qdot705 (N = 3093), both at 500 mM KCl.

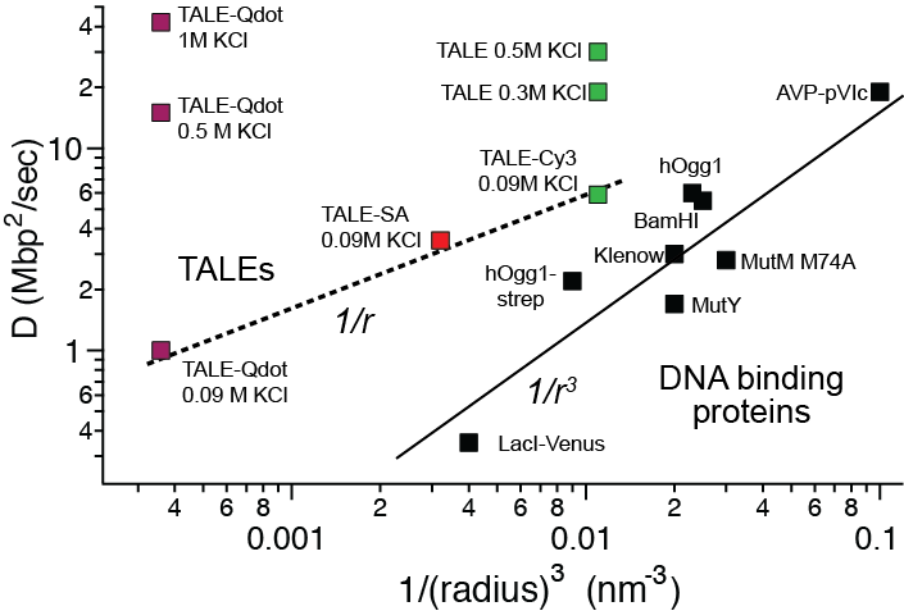

Figure 3.17 Comparison of 21.5 repeat TALE 1-D diffusion speeds with those recorded for other DNA binding proteins[10,27]. TALE diffusion was measured at a variety of salt concentrations (0.09-1M KCl) for TALEs conjugated to a variety of fluorescent probes. The clear separation of TALE diffusion from that of other DBPs highlights the unique nature of TALE sear

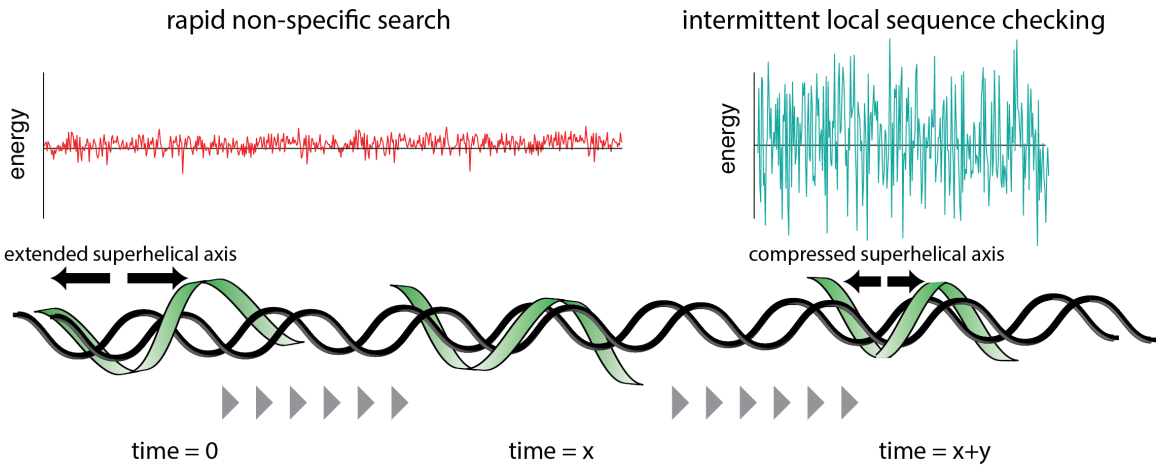

Figure 3.18 Proposed search mechanism for TALEs. In this model, TALEs assume an extended superhelical conformation during their rapid searching events (bottom left), allowing them to search large amounts of genetic code efficiently due to a low energetic barrier to diffusion (top left). They do not strictly rotate around the DNA helix during this scanning, and their extended conformation places them out of phase with the DNA major groove. Searching TALEs are able to compress along their helical axis to check the local sequence (bottom right) with a frequency that is likely a result of the length of the CRD. Checking events cause TALEs to experience a much rougher energy landscape in comparison to searching events (top right), as the number of interactions between TALE and DNA increase once the TALE is fully in phase with the DNA major groove.

3.6 References

1. Slutsky, M. & Mirny, L. a. Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. *Biophys. J.* **87,** 4021–4035 (2004).

2. Mirny, L. *et al.* How a protein searches for its site on DNA: the mechanism of facilitated diffusion. *J. Phys. A Math. Theor.* **42,** 434013 (2009).

3. Elf, J., Li, G.-W. & Xie, X. S. Probing transcription factor dynamics at the single-

molecule level in a living cell. *Science* **316,** 1191–4 (2007).

4. Dikić, J. *et al.* The rotation-coupled sliding of EcoRV. *Nucleic Acids Res.* **40,** 4064–4070 (2012).

5. Gorman, J., Plys, A. J., Visnapuu, M.-L., Alani, E. & Greene, E. C. Visualizing one-dimensional diffusion of eukaryotic DNA repair factors along a chromatin lattice. *Nat. Struct. Mol. Biol.* **17,** 932–938 (2010).

6. Bonnet, I. *et al.* Sliding and jumping of single EcoRV restriction enzymes on non-cognate DNA. *Nucleic Acids Res.* **36,** 4118–4127 (2008).

7. Schurr, J. M. The one-dimensional diffusion coefficient of proteins absorbed on DNA. Hydrodynamic considerations. *Biophys. Chem.* **9,** 413–414 (1975).

8. Bagchi, B., Blainey, P. C. & Sunney Xie, X. Diffusion constant of a nonspecifically bound protein undergoing curvilinear motion along DNA. *J. Phys. Chem. B* **112,** 6282–6284 (2008).

9. Kampmann, M. Obstacle bypass in protein motion along DNA by two-dimensional rather than one-dimensional sliding. *J. Biol. Chem.* **279,** 38715–38720 (2004).

10. Blainey, P. C. *et al.* Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.* **16,** 1224–1229 (2009).

11. Cuculis, L., Abil, Z., Zhao, H. & Schroeder, C. M. Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nat. Commun.* **6,** 7277 (2015).

12. Murakami, M. T. *et al.* The repeat domain of the type III effector protein PthA shows a TPR-like structure and undergoes conformational changes upon DNA interaction. *Proteins Struct. Funct. Bioinforma.* **78,** 3386–3395 (2010).

13. Desruisseaux, C., Long, D., Drouin, G. & Slater, G. W. Electrophoresis of Composite Molecular Objects. 1. Relation between Friction, Charge, and Ionic Strength in Free Solution. *Macromolecules* **34,** 44–52 (2001).

14. Hinton, P. *Statistics Explained*. (Routledge, 2014).

15. Vestergaard, C. L., Blainey, P. C. & Flyvbjerg, H. Optimal estimation of diffusion coefficients from single-particle trajectories. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **89,** (2014).

16. Gao, H., Wu, X., Chai, J. & Han, Z. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.* **22,** 1716–20 (2012).

17. Halford, S. E. & Marko, J. F. How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.* **32,** 3040–3052 (2004).

18. Mak, A. N.-S., Bradley, P., Cernadas, R. A., Bogdanove, A. J. & Stoddard, B. L. The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **335,** 716–9 (2012).

19. Kochaniak, A. B. *et al.* Proliferating cell nuclear antigen uses two distinct modes to move along DNA. *J. Biol. Chem.* **284,** 17700–17710 (2009).

20. Cravens, S. L., Hobson, M. & Stivers, J. T. Electrostatic properties of complexes along a DNA glycosylase damage search pathway. *Biochemistry* **53,** 7680–7692 (2014).

21. Etson, C. M., Hamdan, S. M., Richardson, C. C. & van Oijen, A. M. Thioredoxin suppresses microscopic hopping of T7 DNA polymerase on duplex DNA. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 1900–1905 (2010).

22. Berg, O. G., Winter, R. B. & von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* **20,** 6929–6948 (1981).

23. Berg, O. G., Winter, R. B. & von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry* **20,** 6929–6948 (1981).

24. Komazin-Meredith, G., Mirchev, R., Golan, D. E., van Oijen, A. M. & Coen, D. M. Hopping of a processivity factor on DNA revealed by single-molecule assays of diffusion. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 10721–6 (2008).

25. Cho, W. K. *et al.* ATP alters the diffusion mechanics of MutS on mismatched DNA. *Structure* **20,** 1264–1274 (2012).

26. Schwarz, F. W. *et al.* The Helicase-Like Domains of Type III Restriction Enzymes Trigger Long-Range Diffusion Along DNA. *Science* **340,** 353–356 (2013).

27. Blainey, P. C. *et al.* Regulation of a viral proteinase by a peptide and DNA in one-dimensional space IV: Viral proteinase slides along dna to locate and process its substrates. *J. Biol. Chem.* **288,** 2092–2102 (2013).

28. Xiong, K. & Blainey, P. C. Molecular sled sequences are common in mammalian proteins. *Nucleic Acids Res.* gkw035 (2016). doi:10.1093/nar/gkw035

29. Tafvizi, A., Huang, F., Fersht, A. R., Mirny, L. A. & van Oijen, A. M. A single-molecule characterization of p53 search on DNA. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 563–568 (2011).

30. Dunn, A. R., Kad, N. M., Nelson, S. R., Warshaw, D. M. & Wallace, S. S. Single Qdot-labeled glycosylase molecules use a wedge amino acid to probe for lesions while scanning along DNA. *Nucleic Acids Res.* **39,** 7487–7498 (2011).

31. Barbi, M. & Paillusson, F. in *Advances in Protein Chemistry and Structural Biology* **92,** 253–297 (Copyright © 2013 Elsevier Inc. All rights reserved., 2013).

32. Esadze, A., Kemme, C. A., Kolomeisky, A. B. & Iwahara, J. Positive and negative impacts of nonspecific sites during target location by a sequence-specific DNA-binding protein: Origin of the optimal search at physiological ionic strength. *Nucleic Acids Res.* **42,** 7039–7046 (2014).

33. Terakawa, T., Kenzaki, H. & Takada, S. P53 searches on DNA by rotation-uncoupled sliding at C-terminal tails and restricted hopping of core domains. *J. Am. Chem. Soc.* **134,** 14555–14562 (2012).

34. Gorman, J. *et al.* Single-molecule imaging reveals target-search mechanisms during DNA mismatch repair. *Pnas* **109,** 3074–3083 (2012).

35. Lei, H., Sun, J., Baldwin, E. P., Segal, D. J. & Duan, Y. *Conformational elasticity can facilitate TALE-DNA recognition*. *Advances in Protein Chemistry and Structural Biology* **94,** (Elsevier Inc., 2014).

36. Schreiber, T. *et al.* Refined Requirements for Protein Regions Important for Activity of the TALE AvrBs3. *PLoS One* **10,** e0120214 (2015).

37. Winter, R. B., Berg, O. G. & von Hippel, P. H. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli lac repressor--operator interaction: kinetic measurements and conclusions. *Biochemistry* **20,** 6961–77 (1981).

38. Kulkarni, G. S. & Zhong, Z. Detection beyond the Debye screening length in a high-frequency nanoelectronic biosensor. *Nano Lett.* **12,** 719–723 (2012).

Chapter 4: Direct Observation of TALE Specific Binding and Role of Divalent Cations on TALE Specificity

4.1 Introduction

Activity of site-specific DNA binding proteins requires that they achieve stable and specific binding at a sequence of interest, often for considerable lengths of time. In order for target site binding to be of use, the difference between binding affinity for random and target DNA must be considerable. Indeed, many transcription factors and restriction enzymes display binding discrimination over several orders of magnitude, when comparing binding constants for specific and non-specific interactions[1–3]. Binding affinity at equilibrium is typically presented as the dissociation constant, k_d. For a unimolecular reaction given by (4.1)

$$A_xB_y \rightarrow A_x + B_y \tag{4.1}$$

where A and B are reactants (with x and y being their respective stoichiometries) and AB is the complex (here the DNA-bound protein complex), k_d is given as (2)

$$k_d = \frac{[A]^x[B]^y}{[A_xB_y]} \tag{4.2}$$

In this way, k_d is given in units of concentration, and when $x = y = 1$, k_d represents the concentration at which there is equal likelihood to find the bound complex or the individual unbound reactants. In solution, 50% of *A* and 50% of *B* will be bound together/reacted and 50% will be unassociated. A very low k_d value therefore represents a strong association between *A* and *B*. DNA-binding proteins that are responsible for binding specific target sequences often have k_d values in the picomolar to nanomolar range for target DNA and k_d values for random, non-specific DNA in the micromolar to

millimolar range[1,4,5]. Dissociation constants, for a protein binding to specific and non-specific DNA, that are similar (i.e. on the same order of magnitude) are antithetical to high target specificity; proteins remaining 'trapped' on non-specific DNA are not at their target site and thus not accomplishing their function. In the case of a nuclease (such as a TALEN protein) a high propensity to remain at non-target sites could be highly deleterious, as a greater amount of time spent at non-specific sites increases the probability that off-target cleavage will occur.

The energetics that contribute to a protein binding to DNA can be broken down into two components: electrostatic and non-electrostatic interactions[6]. Typically, non-specific binding is mediated by electrostatic interactions, which vary little site to site across a genome. DNA is a negatively charged biopolymer due to its phosphate backbone[7], and DBPs thus typically have areas of basic amino acids which mediate initial binding. Since the negative charge of the DNA backbone is homogenous across sequences, this electrostatic binding contribution is non-specific in nature. Indeed, when engineering a peptide that is capable of binding and diffusing along DNA, a strip of positively charged amino acids (lysine or arginine) is the minimal requirement[8]. The electrostatic contribution to binding affinity is attributed entirely to entropic gain[6]. This increase in entropy arises from the release of ions upon binding, or the cratic entropy of mixing[9]. Specific binding is differentiated from non-specific binding via non-electrostatic interactions[6]. The non-electrostatic contribution to binding free energy can be both entropic (i.e. the removal of ordered water from the DNA and/or protein interface) and enthalpic (i.e. van der Waals interactions or hydrogen bonding between DNA and protein), in contrast to the purely entropic electrostatic contribution[6].

While studies of TALE proteins in gene editing applications have exploded over the past 6 years, there have been a comparatively small number of studies focused on comparing the binding affinity of TALEs to specific versus non-specific DNA. Furthermore, these studies have provided conflicting results, with some showing two or more orders of magnitude difference in specific versus non-specific binding[10–12], while others show less than an order of magnitude difference[13,14]. Only in those studies showing the greatest difference between specific and non-specific binding does the biological activity and gene engineering utility of TALEs appear feasible. A recent molecular dynamics study highlighted that TALE specificity appears to arise mostly from negative discrimination (i.e. the 'least bad' binding site)[15]. In this study, the major contributors to the binding of TALEs to their target sequence were in fact non-specific interactions between the glycine, lysine, and glutamine residues at positions 14, 16, and 17, respectively, within each repeat and the DNA backbone. The second largest contributor to binding energy was shown to be the NTR (an entirely nonspecific interaction). Specific interactions between the RVD (position 13 in the repeats, as 12 is only a modulator of repeat structure) were a distant third contributor to overall TALE binding affinity. The authors postulated that TAL specificity is thus largely a result of steric clashes between residue 13 of repeats (the second residue in the RVD), and therefore described TALE specificity as arising from negative discrimination[15]. Open questions still surround TALE specificity, however. With so much binding energy attributed to non-specific interactions, how do the ionic strength and the ionic species present in solution affect TALE specificity? How does the low contribution of specific binding to overall binding energy permit successful target search and localization? Given

a negative discrimination model for binding, how can experimental results showing similar binding activities for specific and non-specific binding be reconciled?

Studies of TALE(N) binding to date have mostly consisted of *in vivo* experiments using cell reporter assays or through measurement of the cut/uncut DNA of the cells/organisms that are subjected to TALENs[16–19]. These experiments make the determination of specific factors influencing TALE specificity difficult to obtain, due to the complex nature of the cellular environment, which in turn leads to an inability to assume a reductionist view of the systems. Surveying the *in vitro* studies of TALEs, gel shift assays (EMSA – electrophoretic mobility shift assays) have been the primary platform for studying TALE-DNA interactions[10–12,20–22]. These assays are inherently static in nature, and fail to capture any dynamics of binding due to the long times elapsing between incubation, binding, and the radiographic readout. A much smaller number of *in vitro* studies have utilized calorimetry[13] or fluorescence-based assays[14,23,24]. These *in vitro* studies have been useful in determining several components of TALE binding, including what portions of the protein are required for binding[13,24]. However, these studies have been far from exhaustive in the binding conditions they test – typically they have only examined small changes in the TALE structure or large-scale changes in the DNA binding substrates (i.e. full target DNA or fully random DNA).

There are three key factors contributing to TALE binding specificity that have been experimentally investigated up until this point. The first is the structure of the TALE itself. Work by Gao *et al* showed that the TALE NTR was critical for TALE binding (NTR-deficient mutants did not bind DNA in a calorimetric assay)[13], and that the NTR is capable of binding DNA on its own, in a non-specific manner. Recent work by

Schreiber *et al* further showed that the lengths of both the NTR and CTR are directly related to the binding affinity of TALEs[24], likely due to the addition of positively charged amino acids as each region is extended, which drives overall DNA affinity of TALEs. In this way, Schrieber and coworkers further established lower bounds for successful TALE binding. Several studies have shown a higher overall DNA-binding affinity of longer CRDs, which again is logical given the increase in number of positively charged amino acids each repeat provides, though conflicting conclusions regarding the effect of CRD length on target specificity have emerged[25]. On one hand, it has been reported that longer CRDs provide higher specificity due to their larger specific binding requirements. On the other hand, shorter CRDs have been reported to be more specific due to their overall lower DNA binding affinity and thus lower propensity to bind non-specific DNA. Third, and finally, the requirement of a thymidine at the 5' end of the binding site has been well characterized, and found to significantly impact the binding of TALEs to target sequences[25].

While the effects of solution conditions (i.e. ionic strength and ion types) on TALE specificity have not been studied, the effects of these factors on the specificity of other DBPs have been extensively investigated for several systems. One recurring theme in modulation of DBP specificity is the suppression/enhancement of electrostatic binding interactions. Several sequence specific DBPs, including the *lac* repressor[26,27] and tumor suppressor p53[28] demonstrate minimal discrimination of target binding under low ionic strength conditions, when electrostatic interactions are not significantly screened. These proteins display substantial target site specificity under elevated salt concentrations when non-specific binding affinity is supressed[27,28]. This behavior is rationalized when the

components of target binding (hydrogen bonding and van der Waals interactions) are considered.

Salt-titrations of DBP binding are regularly utilized to elucidate the contributions of both electrostatic and non-electrostatic binding energies[6]. These experiments measure the change in DBP binding constant (K_a) as the monovalent ion concentration of solution is increased. Counter-ion condensation theory attributes the electrostatic component of protein-DNA binding to the cratic entropy of mixing, and thus predicts at elevated ionic strength the cratic entropy becomes less favorable because the difference in concentration of ions surrounding DNA (which is high even at lower ionic strengths) and the concentration in bulk solution is reduced[9]. This understanding allows for a correlation between bulk salt concentrations and DBP binding affinity to be established. This relationship between salt concentration and binding affinity provides a simplistic, but insightful view of the electrostatic contributions to DBP binding affinity, including an estimate for the number of ions displaced by the protein during binding[6]. The relationship is typically presented as

$$\log(K_a) = \log(K_{a,nel}) - N\text{Log}[\text{Salt}] \qquad (4.3)$$

Where $K_{a,\ nel}$ is the binding affinity due to non-electrostatic interactions and N is the number of cations released during a binding event. According to counter-ion condensation theory, the number of cations N can be further described by:

$$N = Z\Psi \qquad (4.4)$$

where Z is the number of phosphate groups interacting with the protein and Ψ is the number of cations associated with a phosphate group.

Here, we present an in-depth study of the effects of ionic strength and different ions on the binding of TALEs. We utilize both single molecule fluorescence microscopy, as well as fluorescence anisotropy assays that permit us to easily modify the solution conditions for binding and the sequence of the probe molecule. Under purely monovalent salt conditions, we find that TALE specificity is prohibitively minimal, especially for longer TALEs (here, a 21.5 repeat TALE construct). At supraphysiological salt concentrations (greater than 150 mM KCl) the overall affinity for both target and non-target DNA is decreased, but the difference between the two is never appreciable. Strikingly, we find that a substantial degree of binding specificity is introduced upon inclusion of low concentrations of certain divalent cations. In particular, 5 to 10 mM Mg^{2+} or Ca^{2+} added to physiological monovalent salt solutions (100-150 mM KCl) introduces differences between target and non-target binding of several orders of magnitude, as non-specific binding is nearly abolished. For the 21.5 repeat TALE construct, we see a slight enhancement of target binding as well, leading to an even greater difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$. Seeking to better understand the origin of this divalent cation effect, we probe the binding of the N-terminal region (NTR), which is known to be critical for TALE binding to both specific and non-specific DNA[13], in response to divalent versus purely monovalent salt solutions and find that NTR binding is attenuated significantly more by divalent salts than monovalent salts. We also examine the effects of TALE CRD length on *in vitro* binding affinity and find that TALE affinity increases with the size of CRD, and that the effects of divalent salts are magnified in larger TALEs, here the 21.5 repeat TALE construct. We utilize the newly discovered binding conditions to provide evidence that TALEs are more sensitive to target site

mismatches at the 5' (NTR) end of the binding site, as previously proposed via *in vivo* and *in vitro* TALEN cleavage assays[22].

Taking the insights developed from our fluorescence anisotropy measurements, we utilized our previously described SMFM platform to directly observe TALEs binding to target site-containing DNA substrates. Consistent with the results from anisotropy measurements, we find that the presence of moderate concentrations of magnesium significantly reduces both the frequency and duration of TALE non-specific binding events. In both the presence and absence of magnesium, TALEs are able to locate their target site via either one-dimensional facilitated diffusion or apparent three-dimensional collisions. We find that under both pure monovalent salt and in mixed monovalent/divalent salt conditions, some TALEs display specific binding lifetimes of minutes, surpassing the maximal time to photobleaching for the Cy3 dyes under constant laser illumination. We also observe, for a small subset of events, TALEs 'missing' their target site, wherein they diffuse past the target location, suggesting that the target binding process is partially reaction-limited, and not solely dependent on the speed at which TALEs reach their target. Finally, we observe collisions between TALEs engaging in their search process, and specifically bound TALEs, which result in the reversal of the searching TALEs (i.e. the searching TALEs 'bounce off' the specifically bound TALEs).

4.2 Materials and Methods

4.2.1 Protein preparation

TALE 21.5, 15.5, and 11.5 repeat constructs, and the TALE NTR were prepared as described in previous chapters. In fluorescence anisotropy experiments, TALEs were not

labeled with a fluorescent dye, as the reporter in this assay was the fluorescently labeled DNA substrate.

4.2.2 DNA templates

Fluorescence anisotropy templates. Several DNA templates were used to study TALE binding using fluorescence anisotropy (**Figure 4.1**). DNA templates were prepared using the following method. Single stranded oligonucleotides (Integrated DNA Technologies) were first prepared in stock solutions in low salt Tris buffer (20 mM Tris-HCl, pH 7.5, 50 mM KCl). DNA concentrations were determined via UV/vis spectrophotometry (NanoDrop, Thermo Fisher). Equivalent concentrations of a 5,6-FAM (fluorescein derivative) labeled oligonucleotides and their corresponding complimentary oligonucleotides were combined into solution in low salt Tris buffer, placed in a covered heating block at 90°C for 2 minutes, followed by slow cooling to room temperature.

Single molecule imaging assay templates. The plasmid containing no TALE binding sites, 'Target receiver plasmid' (Tar-rec), is a 44,898-bp plasmid (unpublished) containing neoaureothin synthesis pathway from *Streptomyces orinoci*. The plasmid containing eight TALE binding sites, SM8x, was constructed from Tar-rec by introduction of a fragment containing the TALE target sequence flanked at its 3' end by a *LacZ* gene. To this end, Tar-rec was linearized with PspXI enzyme. The *LacZ* gene was amplified from pNEB193 plasmid (NEB), where a forward primer contained the TALE-binding 23-nucleotide sequence at it's 5' end. The amplified fragment was spliced with the linearized vector using the Gibson Assembly method. The resulting SM8x plasmid contained the eight TALE-binding sites approximately 11 kb away from the SnaBI restriction site used for double tethering of DNA on a glass slide.

4.2.3 Fluorescence anisotropy experiments

Mixtures of 1 nM double stranded (ds) oligonucleotide templates and TALE protein were prepared in a standard buffer (20 mM Tris-HCl, pH 7.5) with varying amounts of monovalent salt (KCl or NaCl) and divalent salt ($MgCl_2$, $MgSO_4$, $ZnCl_2$, $MnSO4_2$, $CaCl_2$) where specified). Mixtures of TALE protein and DNA template (200 μL) were incubated for 10 minutes at room temperature and then assayed in black 96-well plates (Corning) in duplicates. Fluorescence polarization measurements were performed on an Infinite 200 Pro microplate reader (Tecan) using excitation and emission wavelengths of 485 nm and 535 nm, respectively. Fluorescence polarization values were converted to fluorescence anisotropy values using the following relation (4.5):

$$A = \frac{2P}{3-P} \tag{4.5}$$

where A is anisotropy and P is polarization. The dissociation constant (K_D) was calculated using a non-linear least squares algorithm (Origin 8.5) using the following expression (4.6):

$$A = A_f + (A_b - A_f) * \frac{(L_T + K_D + R_T) - \sqrt{(L_T + K_D + R_T)^2 - 4L_T R_T}}{2L_T} \tag{4.6}$$

where A is experimental anisotropy value, A_f is anisotropy of DNA without protein, A_b is anisotropy of protein-bound DNA, L_T is total ligand (DNA) concentration, and R_T is total receptor (protein) concentration. For experiments with unlabeled competitor DNA, 100 nM of DNA templates lacking 5,6 FAM was added to solutions containing 1 nM labeled probe DNA.

4.3 Results and Discussion

4.3.1 TALEs show minimal target recognition in pure monovalent salt

We utilized a fluorescence anisotropy assay in order to probe the role of monovalent and divalent salts on the specificity of TALE binding. To this end, we used both target DNA sequences and a series of different random DNA sequences to probe the affinity and specificity of TALEs containing 11.5, 15.5, and 21.5 repeats in the CRD (**Figure 4.1**). We further compare these results to the binding data of TALE truncation mutants containing only the TALE NTR. In all cases, we characterize TALE binding affinities under a variety of solution conditions. In buffered solutions containing the monovalent salt KCl in the absence of divalent cations, TALE binding affinities to target and random DNA sequences appear to be similar (**Figure 4.2, Tables 4.1 and 4.2**). Strikingly, the apparent low degrees of specificity were observed for a series of different TALEs studied in this work (21.5, 15.5, and 11.5 repeats in TALE CRD) and across a wide range of different concentrations of KCl and NaCl (**Figures 4.2-4.6**). For the TALE with the largest number of repeats (21.5 repeats in CRD), we observed a significant lack of specificity between target and random DNA even under high concentrations of KCl (**Figure 4.5**). No difference in specificity was observed when the monovalent cation was exchanged for Na^+ in place of K^+ (**Figure 4.6**). In the presence of monovalent salt, TALEs with shorter length CRDs appear to show greater discrimination between random and target sequences compared to larger TALEs. Specificity for shorter TALEs increases with increasing concentration of KCl (**Figures 4.3** and **4.4**), albeit showing only a maximum ~3-fold preference for specific DNA (**Tables 4.3-4.6**) when comparing $k_{d,\ non\text{-}}$

$_{specific}$ and $k_{d,\ specific}$, which is not enough to account for the specificity required for TALE and TALEN function *in vivo*.

4.3.2 NTR exhibits electrostatic-mediated interactions between TALEs and DNA

We further sought to elucidate the role of electrostatics on TALE binding to specific and non-specific DNA. In particular, we aimed to understand how electrostatics plays a role in the binding specificity for TALE constructs containing all three domains (NTR+CRD+CTR) compared to TALE truncation mutants. Here, we utilized a truncation mutant containing only the TALE NTR (previously described in Chatpers 2 and 3), and we studied the binding of the TALE NTR and TALEs with 21.5 repeats in their CRD as a function of KCl concentration. As expected, overall affinity for both target and random sequences decreases with increasing monovalent ionic strength. We further analyzed the TALE binding data using counter-ion condensation theory[6]. Here, we plot the logarithm of the binding constant K_a versus the logarithm of the concentration of KCl (**Figure 4.7**). In this way, the effective number of monovalent ions displaced during a TALE binding event can be determined, thereby enabling an estimate of the number of phosphate groups on the DNA backbone contacted by TALEs during a binding event. In this way, we can determine the relative binding affinity using **Equations 4.3** and **4.4**.

Our data show that approximately 5 DNA phosphate groups are in contact with the NTR under monovalent salt conditions whereas 5 to 6 groups (on average) are in contact with TALEs containing 21.5 repeats in the CRD (**Figure 4.7** and **Table 4.7**). Overall, these results highlight the significant role of electrostatics in TALE NTR binding to DNA, which is consistent with the NTR crystal structure and previous models for TALE binding in which the CRD only makes partial contacts with the DNA backbone.[29]

Counter-ion condensation theory also provides a framework for determining the relative contributions of electrostatic and non-electrostatic binding energies to the overall binding energy, via extrapolation to high salt concentrations. Extrapolating the plots of log(K_a) versus log[KCl] to 1M salt, thus determining the y-intercept, allows for separation of electrostatic and non-electrostatic binding energy. The seemingly arbitrary selection of 1M KCl as the point at which electrostatic contributions go to zero can be considered a reference point for comparing across multiple proteins[30]. In addition, previous studies of DNA-binding proteins have examined the binding affinity of multiple protein truncation mutants in which regions of electrostatic and non-electrostatic affinity are clearly defined[30]. These studies also revealed that at or very near 1M salt the electrostatic contributions to binding free energy go to zero. From these measurements on TALE constructs, the non-electrostatic component of binding for the NTR is -4.51 kcal/mol. However, the non-electrostatic contribution for TALEs with 21.5 repeats is -6.24 kcal/mol and -7 kcal/mol for random and cognate DNA, respectively. Based on these data, it appears that the additional binding affinity for TALEs with intact CRD domains (known to mediate specificity in DNA binding) is primarily conferred by non-electrostatic interactions. Overall, these data are supported by previous findings that the TALE CRD repeats are engaged in non-specific search/binding, but only partially.[29]

4.3.3 Divalent cations enhance specificity in TALE binding

We further sought to understand the apparent lack of sequence specificity for TALE binding in monovalent salt by assessing the impact of ionic strength and solution conditions. Here, we studied TALE binding to target and random DNA in the presence of various divalent salts ($MgCl_2$, $MgSO_4$, $ZnCl_2$, $MnSO_4$, $CaCl_2$). Strikingly, the lack of specificity for TALE binding to target sites is restored in the presence of certain divalent

cations (**Figures 4.8** and **4.9, Table 4.8**). In particular, the divalent cations Mg^{2+} or Ca^{2+} substantially diminish non-target binding and in some cases enhance the binding affinity to target sites. We characterized TALE binding in the presence of a variety of divalent cations, and we observed the strongest effect with Mg^{2+} or Ca^{2+}, a moderate effect with Mn^{2+}, a slight effect with Sr^{2+}, and no effect with Zn^{2+}. No differences in TALE binding were observed upon changing the monovalent anions or divalent anions (**Figure 4.10**). Interestingly, these effects appear to be governed by the overall concentration of divalent cations, suggesting that this phenomenon is not entirely driven by ionic radius or any particular periodic trend. Rather, this effect appears to be sensitive to the chemical identity of divalent cations, with an apparent correlation with cations with higher cellular abundance. To our knowledge, similar effects are uncommon in non-catalytic DNA binding proteins. A analogous biochemical example appears to be related to the bacterial transcription factor CREB, which has been shown to require Mg^{2+} to properly bind its 10-bp target.[31]

4.3.4 Divalent cation-induced discrimination is ionic strength-dependent and increases with the size of the TALE central repeat domain

We further sought to understand the effects of divalent cations on TALE binding to target and random DNA sequences. Here, we studied the effect of increasing the overall ionic strength (via increasing divalent cation concentration) on TALE binding. Our results show that TALE binding affinity generally decreases upon increasing the concentration of divalent cations (**Figure 4.11**), which is similar to TALE binding behavior in the presence of monovalent salt. However, we found that the decrease in DNA binding affinity to off-target sites was significantly larger when the increase in total ionic strength resulted from a higher proportion of divalent cations (as opposed to

monovalent cations). Based on these results, we further studied the effect of increasing the proportion of divalent cations in solution while maintaining the total ionic strength at a constant value. In these experiments, we prepared a series of solutions at constant (total) ionic strength (90 or 120 mM), while increasing the proportion of divalent cations Mg^{2+} in solution (**Figures 4.12-4.14**).

Our results show that TALE binding specificity increases for target DNA (corresponding to an increase in the difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$) with increasing proportion of Mg^{2+} for all full length TALEs studied in this work (TALEs with 11.5, 15.5, and 21.5 repeats). For shorter TALEs (11.5 and 15.5 repeats), differences in binding affinity between random and target sites were minimal until > 40% of the total ionic strength was contributed by divalent Mg^{2+} cations (**Figures 4.13** and **4.14**). In contrast, TALEs with 21.5 repeats showed higher specificity for target sites (where the difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$ is more than 500-fold) when the proportion of Mg^{2+} reached similar levels (**Figure 4.12**), and intermediate specificity was observed for TALEs with 21.5 repeats (where the difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$ is between 10 and 50 fold) for lower proportions of Mg^{2+}. At higher total ionic strengths with moderate proportions of Mg^{2+} (similar to physiological conditions), TALEs bound target DNA with extremely high specificity (**Figure 4.15**), consistent with prior studies that utilize TALEs for specific genomic edits. Interestingly, in buffered solutions containing only Mg^{2+} as the primary cation (40 mM $MgSO_4$), TALEs with 21.5 repeats bind to target DNA with extremely high specificity (the difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$ is approaches infinity, effectively as $k_{d,\ non\text{-}specific}$ is too large to measure) (**Figure 4.16**). Overall, these results suggest that TALE binding specificity increases with

higher concentrations of divalent cations, in contrast to proportional increases in ionic strength provided by monovalent cations only.

4.3.5 TALE NTR binding to DNA is impaired by divalent cations

In order to elucidate the effect of divalent cations on target site discrimination for TALEs, we further probed the role of divalent cations on TALE NTR binding (**Figures 4.17** and **4.18**). In prior work, it has been clearly demonstrated that the NTR is critical for nucleation of TALE binding events.[13,24,29] Furthermore, we have previously shown that a truncation mutant containing only the TALE NTR is capable of binding and diffusing one-dimensionally along DNA (Chapters 2 and 3) in reduced ionic strength conditions in the presence of monovalent salt[29]. In this work, we observed that Mg^{2+} entirely suppresses the ability of the NTR to bind to DNA, even in low concentrations of the divalent cation (**Figure 4.17b**). In particular, the binding affinity of the TALE NTR to DNA was significantly attenuated even at very low total ionic strength in the presence of 2.5 mM $MgSO_4$ (with 10 mM total ionic strength) when compared to pure monovalent salt conditions (**Figure 4.17a**). As expected, a TALE construct that lacked the NTR (CRD+CTR only) showed no binding affinity for target or random DNA (**Figure 4.19**). Taken together with prior results, these data help to elucidate the mechanism of TALE binding to target DNA[13,29]. The TALE NTR is essential for initial binding to DNA, and this initial non-specific binding is driven primarily by electrostatic interactions. Our results suggest that addition of monovalent salt and/or an increase in overall ionic strength can weaken the initial binding interactions, however, divalent cations greatly suppress NTR binding to DNA. From this view, the NTR is responsible for nucleating binding to DNA, followed by interactions of the TALE CRD with DNA via primarily by non-electrostatic (hydrogen-bonding) interactions. TALEs with shorter CRDs have a

greater proportion of their overall binding energy conferred via the NTR, in comparison to TALEs with longer CRDs. Thus, when NTR binding is significantly attenuated by the presence of divalent cations, the overall binding affinity of shorter TALEs will be more strongly impacted than the binding affinity of longer TALEs, which is consistent with the results we present here.

4.3.6 Competitor DNA attenuates TALE binding in monovalent but not divalent salt conditions

We next utilized a DNA competitor assay for fluorescence anisotropy in which a large molar excess (100-fold) of unlabelled random DNA was added to 1 nM of fluorescently labelled probe DNA (**Figure 4.20**). We hypothesized that in the presence of a large excess of competitor DNA, TALEs that exhibit a relatively weak non-specific binding interaction will be sequestered away from target (probe) DNA, as they are unable to discriminate between target and random DNA. In pure monovalent salt conditions, the 21.5 repeat TALE binding to both target and random DNA was substantially attenuated by the presence of unlabelled random competitor DNA (**Figure 4.21a**). In the presence of monovalent salt, therefore, there appears to be minimal differences in the behavior of TALE binding to target versus random DNA. Interestingly, the results for TALE binding in monovalent salt are sharply contrasted by results TALE binding to target DNA in the presence of 10 mM $MgCl_2$ (with equivalent total ionic strength as monovalent salt binding experiments). Notably, TALE binding to target DNA is only minimally impacted by the large excess of random competitor DNA in the presence of Mg^{2+} divalent cations (**Figure 4.21b**).

4.3.7 TALE binding to target DNA originates at the 5' region

With newly established conditions critical for sufficient TALE sequence specificity, we sought to probe the sensitivity of TALEs for engineered target site mutations. We designed three new substrates with this goal in mind. The first two substrates, termed hybrid random 1 and hybrid random 2 (**Figure 4.1**), contained ~50% sequence match at either the NTR-proximal or NTR-distal regions of the binding sequence, and complete sequence mismatch at the other half of the substrate. The third substrate, termed SIFTED random, was designed using the readout from a recently developed platform for predicting TALE specificity, Specificity Interface for TAL-Effector Design (SIFTED)[32]. Inputting the TALE target sequence of our 21.5 repeat TALE construct into SIFTED provides a position weight matrix (PWM) detailing the effects of specific base pair mismatches on the binding affinity of the TALE relative to its affinity for its full target match. We took this PWM and determined what combination of base pairs would yield the lowest possible affinity score. In this way the SIFTED substrate we designed was engineered such that the 21.5 repeat TALE construct had the lowest possible affinity for it. We measured 21.5 repeat TALE binding to these three new substrates under conditions that permitted measurable off-target binding but still induced discrimination between target and non-target binding (100 mM KCl, 5 mM $MgCl_2$) (**Figure 4.22**). Our results showed that there was only a discernable difference in binding affinity between hybrid random 2, the substrate with NTR-distal mismatches, and the other random substrates (hybrid random 1, SIFTED random, and the original random substrates). In this way we observe that TALEs are appear to be more tolerant of NTR-distal mismatches, but demonstrate little difference in binding once mismatches are introduced at NTR-proximal positions. This result is in good agreement with several

other studies suggesting that TALEs display a polarity in binding, and best tolerate NTR-distal mismatches[22,33].

4.3.8 Single molecule visualization of non-specific TALE search with Mg^{2+}

Motivated by the intriguing effects of divalent cations on TALE specificity observed in bulk-level fluorescence anisotropy experiments, we next sought to directly visualize TALE binding to target DNA sequences. Here, our experiments focus on understanding TALE binding dynamics in the presence of magnesium, and we compare these results to TALE dynamics in the presence of purely monovalent salt conditions.

For these experiments, we utilize our TIRF-based SMFM assay to observe single TALE interactions with single-tethered DNA templates. We utilized single-tethered DNA extended under constant flow as to match results described in 4.3.8. We began by studying the dynamics of the 21.5 repeat TALE construct specifically labeled with Cy3 along 48.5 kbp lambda-phage DNA containing zero target binding sites. We observed a marked decrease in TALE binding on nonspecific regions of DNA when 5 mM of $MgCl_2$ was included in a standard imaging buffer (130 mM KCl, 50 mM MOPS pH = 8.1) (**Figure 4.23**), which is consistent with our results fluorescence anisotropy data showing decreased nonspecific binding affinity of TALEs in the presence of divalent cations. After introducing TALE proteins (250 pM protein concentration into the sample chamber for two minutes, we imaged the dynamics of TALEs bound to DNA templates. We generally observed one or two TALEs (at most) bound to a single DNA template at any time for the standard imaging buffer containing 130 mM KCl, 5 mM $MgCl_2$ (150 mM total added ionic strength) (**Figure 4.23**). Interestingly, we observed high levels of TALE binding (5 or more TALEs bound to the same DNA template) when the imaging buffer is

comprised of 150 mM KCl only (**Figure 4.24**). Moreover, in the presence of purely monovalent salt, TALEs remain bound and searching for longer periods of time.

As in Chapter 2, we determine the non-specific DNA binding lifetimes of the 21.5 repeat TALE (**Figure 4.25**). Interestingly, distributions of binding lifetimes for the 21.5 repeat TALE in 90 mM KCl and 5 mM $MgCl_2$ are fit well to a single exponential decay with a characteristic binding time t_1 = 0.31± 0.02 seconds. These data are in contrast to binding lifetime distributions for the same TALE construct in 110 mM KCl (equivalent in total ionic strength to 90 mM KCl and 5 mM $MgCl_2$), which are fit best to a double exponential decay with characteristic binding lifetimes t_1 = 0.87 and t_2 = 7.9 seconds, respectively. This difference in lifetime distributions is due in part to the significant decrease in long binding events that persist for several seconds (or longer) when magnesium is included in the imaging buffer. Given that magnesium ions associate more tightly with the DNA phosphate backbone than potassium ions[7], we conjecture that the decrease in binding lifetime and reduction in longer non-specific binding events is largely due to the higher energetic cost incurred when magnesium ions must be displaced from the DNA by TALEs. Taken together with our understanding of the wrapped helical conformation of TALEs engaged in long-range non-specific search (Chapter 3), we posit that the additional energetic cost of displacing divalent ions reduces the probability that TALEs are able to fully encircle DNA and thus assume a conformation that supports long-range searching events.

4.3.9 Single molecule visualization of TALE specific binding to target sites

In Chapters 2 and 3 and Section 4.3.6, we described our single molecule fluorescence imaging approach to studying TALE non-specific search. In these studies,

we utilize DNA substrates that lack a target binding site for our TALE constructs. In order to study TALE binding to specific target sites using SMFM, graduate student colleagues Zhanar Abil and Xiong Xiong, working in Professor Huimin Zhao's lab, generated DNA templates containing an array of 8 binding sites (the 22 bp target of the 21.5 repeat TALE studied in this thesis: TAGCAACCTCAAACAGACACCAT (**Figure 4.26**) arranged in tandem, providing a concentrated 'hot spot' for localizing specifically bound TALEs. In order to directly visualize TALE target binding, we introduced Cy3-labeled 21.5 repeat TALEs into sample chambers containing several singly tethered 8x binding site DNA substrates and observed their search and binding dynamics under continuous laser illumination. We were unable to observe TALE-DNA interactions in the absence of flow due to poor ratios of double tethered:single tethered DNA substrates for this 8x binding site substrate. Achieving these results with double tethered 8x substrates is an area of ongoing work within our lab. After observing several minutes of protein binding and search dynamics within one field of view, we introduced Sytox Orange intercalating dye at low concentrations (~1-5 nM) in order to visualize the locations of the DNA molecules within the field of view. This enables us to unambiguously correlate protein behavior with approximate DNA location.

Motivated by our results from fluorescence anisotropy experiments described above, we observed TALE target localization and binding in both the presence and absence of magnesium. In both conditions, we observe that TALEs readily locate their target site following one-dimensional diffusion along DNA substrates (**Figures 4.27 and 4.28**). Upon successful target localization, TALEs remain bound for upwards of several minutes in some cases. Under the conditions of constant illumination by CW laser, we

observed photobleaching of Cy3 dyes, which may occur on a shorter timescale than TALE unbinding under these solution conditions. Additionally, with eight binding sites arranged in tandem, it is possible that dye-dye self-quenching between two or more bound TALE-Cy3 species may occur. In the case of multiple TALEs bound to the 8x array, individual TALE proteins would be positioned within ~2-5 nm of one another and thus possibly in range for intermolecular energy transfer between Cy3 dyes, which could reduce the duration of fluorescence localization at the target site.

The vast majority of TALEs that engage in 1-D search eventually locate the binding sites and remain stably bound immediately upon reaching the array. However, quite interestingly, we also observe occasional bypass of the binding sites under both purely monovalent and mixed monovalent/divalent salt conditions (**Figures 4.29 and 4.30**). Given that eight binding sites are arranged in tandem on this DNA template, we were surprised to observe occasional bypass of the binding array. We can rationalize this behavior by considering our proposed two-state search mechanism for TALE search, in particular by noting that bypass behavior can be reconciled by considering the stochastic nature of the 'check' mode. If TALE checking events occur with a characteristic frequency lower than the inverse time scale required for 1-D diffusion past the 8x binding array, it is possible for a TALE to 'miss' the binding array. For this particular case, the 8x array of binding sites encompasses ~220 bp, giving a timescale for bypass equal to approximately 100 milliseconds, assuming a 1-D diffusion coefficient of 5.9 Mbp^2/sec under these solution conditions. Given that most TALEs bind one of the array's target sites upon first encounter, the distribution of TALE 'checking' frequencies is likely

centered around values greater than 10 s^{-1}, but with some finite population at or below 10 s^{-1}.

In addition to observing TALE target localization, we also observed occasional unbinding and release from binding sites. In these events, TALEs, bound to their target site under both sets of ionic conditions, release under flow and continue on a 1-D search along the remainder of the DNA templates (**Figures 4.31** and **4.32**). As we were unable to observe target binding in the absence of flow, it is not clear how much the applied flow contributed to target release, and how much is a result of the inherently finite binding lifetime of the specifically-bound TALE complex.

Specifically bound TALEs represent potential roadblocks for other TALEs engaging in non-specific 1-D search. Our findings from Chapter 3 strongly suggest that TALEs adopt a wrapped helical conformation during their search process and thus would be incapable of TALE-TALE bypass once wrapped around DNA. Indeed, we observe several instances of TALE diffusion to the end of single-tethered DNA templates followed by apparent TALE protein accumulation at the untethered end of the DNA template. This accumulation is likely due to the extended single stranded DNA (ssDNA) overhangs at the DNA ends, which may represent high barriers to TALE diffusion in contrast to double stranded DNA (dsDNA). These events are clear from the bright ends near kymographs. Furthermore, we also observe TALE-TALE collisions occurring under both sets of ionic conditions (**Figures 4.33** and **4.34**). In these occurrences, specifically bound TALEs represent apparently immovable barriers, as the colliding TALE that is engaged in target search is deflected in the opposite direction. Based on these

preliminary data, we conjecture that the force required to eject a specifically bound TALE is greater than that imparted by a colliding TALE undergoing target search.

4.4 Concluding Remarks

This chapter presents work on the characterization of a divalent cation effect on TALE specificity, in addition to direct visualization of TALE target localization and binding. Prior to our work, the effects of divalent cations on TALE specificity had not been reported. We utilized a fluorescence anisotropy assay to test a variety of DNA substrates in many different solution conditions, identifying the presence of low concentrations of certain divalent cations as a key factor in the ability of TALEs to discriminate between target and non-target sequences. This effect is not evident in a large number of other sequence specific DNA binding proteins. When considering site-specific DBPs, their binding energy can be broken down into two main contributors: electrostatic and non-electrostatic binding energy. Given that DNA is characterized by its negatively charged phosphate backbone, and that this negative charge density is evenly distributed regardless of local sequence, it holds that sequence specificity is not driven by electrostatic interactions. Instead, it is the non-electrostatic binding interactions that primarily allow for a DBP to differentiate between specific and non-specific sequences. These include hydrogen bonding to the nitrogenous bases of DNA, as well as van der Waals interactions. It would hold, therefore, that disruption of electrostatic binding interactions would induce a greater difference between binding affinity of DBPs to specific and non-specific sequences.

Indeed, there are several reported cases of DBP sequence specificity being masked at low ionic strength, only to see significant sequence discrimination as the electrostatic

charge screening is increased. In these cases, however, the sufficient electrostatic screening is satisfied by higher concentrations of monovalent salts. One such example is tumor suppressor p53. A fluorescence anisotropy study of p53 binding to half and full target sites, as well as non-specific substrates, showed that not until 175 mM total ionic strength was apparent target discrimination achieved[28]. The authors included only low concentrations of imidazole buffer and primarily NaCl as the relevant ions, and thus no divalent cations.

A smaller number of studies have highlighted roles for divalent cations in the specificity of sequence-specific DBPs. The most notable examples are the class of basic leucine zipper proteins (B-ZIPs), including CREB and Fos-Jun[31], as well as the prokaryotic transcription factor NikR[34]. In the case of B-ZIP proteins, no discernable sequence specificity is observed at 150 mM KCl like TALEs, but introduction of 10 mM $MgCl_2$ nearly abolishes non-specific binding while only slightly impacting target binding[31]. While a specific justification for divalent cation-induced specificity was not given for CREB, crystal structures of this B-ZIP protein show that there is a hexahydrated magnesium ion bound between the two helices of the zipper domain, and thus this magnesium may stabilize the protein structure[35]. This stabilization may then provide a structure that favors specific binding over non-specific binding. Several restriction enzymes, including BamHI[3,36], EcoRV[37], and EcoRI[4], have also been shown to require Mg^{2+} or Ca^{2+} for sequence specific binding, though such divalent cations are also required for their enzymatic activity.

Similarities between other DBPs showing a divalent cation effect on specificity and TALEs persist only to the extent that B-ZIP proteins have helix-loop-helix domains[35]

which are relatively similar to TALE repeats' helix-loop-helix structure. TALEs' closest structural relative, MTERF1/2, has no reported dependence on divalent cations for binding specificity[38]. Of the three main studies providing TALE crystal structures, only one study included magnesium (or any other divalent cation) in the crystallization process, and the authors reported no sites for magnesium binding. A survey of studies that include *in vitro* binding assays of TALEs with both target and non-target substrates reveals a general trend: those that include magnesium in binding solutions tend to show greater target binding specificity. The vast majority of TALE studies, however, include only an *in vivo* assay and by virtue of the cellular interior, also include millimolar concentrations of both magnesium and calcium[7].

TALE-DNA binding has been largely overlooked from an in-depth, *in vitro* perspective. The vast majority of studies on TALEs either ignore it, focusing only on the results of *in vivo* nuclease activity to comment on factors affecting TALE-DNA binding, or give it minimal thought with basic gel-shift (EMSA) assays designed only to confirm protein-DNA binding. In this work we applied a systematic approach to the study of TALE-DNA interactions, focusing on the effects of different ions at a variety of concentrations. The most striking result of our study was the strong dependence on the presence of certain divalent cations to induce TALE target sequence discrimination. From our previous work and the work of others, we know that the TALE NTR is critical to nucleation of DNA binding events. In this work, we show that binding of the NTR to DNA is significantly attenuated by divalent cations. It is possible that these certain divalent cations interact specifically with the NTR structure, reducing the access of its many positively charged residues to the negatively charged DNA backbone. It is also

possible that divalent cation binding to the DNA occurs such that it impairs TALE/NTR contact in ways that monovalent cations cannot.

We observe that the effects of divalent cations are greatest for the longest TALE (a 21.5 repeat construct in our work), and decrease as the TALE CRD is shortened. From this observation and the effects on NTR binding observed, we propose a general model for the divalent cation effect on TALE specificity. Divalent cations strongly affect NTR binding, and the NTR contributes a more significant proportion of the overall binding affinity for shorter TALEs than longer TALEs. Longer TALEs are able to overcome the reduction in NTR and non-specific binding affinity by their larger CRD regions. Thus, it holds that the effects of divalent cations can be thought of as a balancing act; they increase specificity by decreasing affinity for nonspecific DNA.

In order to better understand the effects of divalent cations on TALE-DNA interaction, we utilized our previously developed single molecule imaging assay to directly observe TALE-DNA dynamics. As expected, we directly visualized a marked decrease in TALE occupation on extended DNA substrates when only 5 mM $MgCl_2$ was included in a 130 mM KCl imaging buffer. In addition to reducing the number of TALE-DNA interactions, we quantified the binding lifetimes of TALEs on non-specific DNA substrates and found that the presence of 5 mM $MgCl_2$ significantly reduced the duration of TALE non-specific search events. In particular, the number of long-lived (>2-3 seconds) events was greatly attenuated. As these events are in all likelihood mediated by the ability of TALEs to fully encircle DNA substrates, the presence of magnesium may serve to induce a significant energetic barrier to TALEs' ability to adopt such a conformation.

We generated long DNA substrates that were engineered to contain an array of eight TALE binding sites (specific for the 21.5 repeat TALE construct) at a precise location. We observed TALEs readily diffuse, carrying out their facilitated search, and subsequently localize at this binding array. This behavior was observed in both pure monovalent and mixed monovalent/divalent cation imaging buffers, suggesting that the enhanced specificity resulting from divalent cations is primarily driven by a decrease in non-specific binding affinity. We also observed target bypass under both conditions, suggesting that the TALE 'check' frequency may be occurring on timescales of ~ 100 milliseconds. Finally, as specifically-bound TALEs represent stationary obstacles on DNA substrates, we observed TALE-TALE collisions that suggest the collisional energy is not typically sufficient to dislodge target-bound TALEs. Taken together, our single molecule imaging results suggest that TALE target binding is partially reaction-driven, and thus not entirely diffusion-driven. These observations further support the two-state search mechanism and loosely wrapped conformation of searching TALEs described in Chapters 2 and 3, given that searching TALEs occasionally miss their target and can be deflected by other bound TALEs.

4.5 Figures and Tables

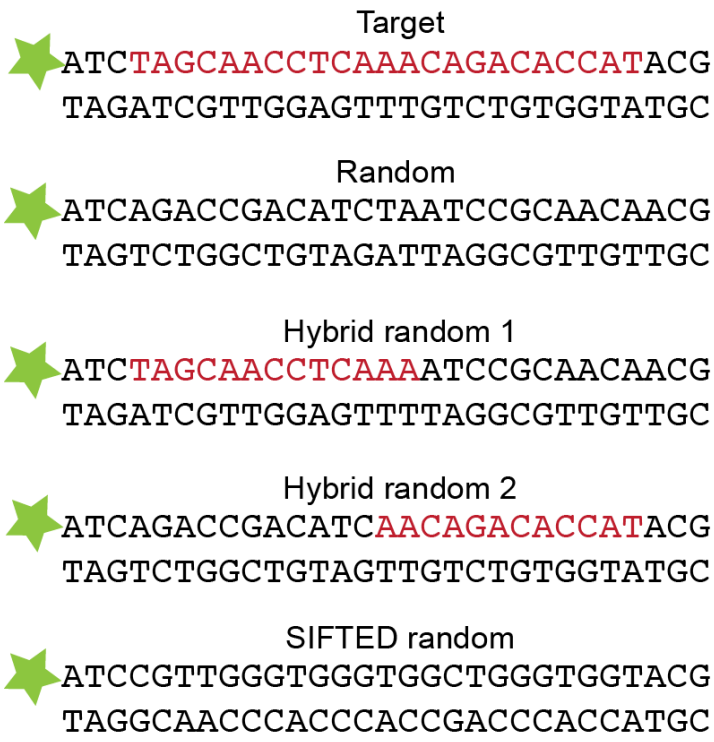

Figure 4.1 DNA substrates used in fluorescence anisotropy measurements. The binding site of the 21.5 repeat TALE construct is shown in red text, and non-specific elements are shown in black. The green star represents the fluorescein molecule conjugated to the DNA that is used for anisotropy measurement. The SIFTED random substrate was designed using the Specificity Interface for TAL-Effector Design (SIFTED) tool, developed by the Joung and Bulyk labs[32]. Briefly, the binding site of the 21.5 TALE construct used in our work was inputted into the SIFTED tool, and a position weight matrix (PWM) generated, which described the predicted relative effects of sequence mismatches on binding affinity. We used this PWM to determine the lowest possible score and thus 'most' random substrate for our 21.5 repeat construct.

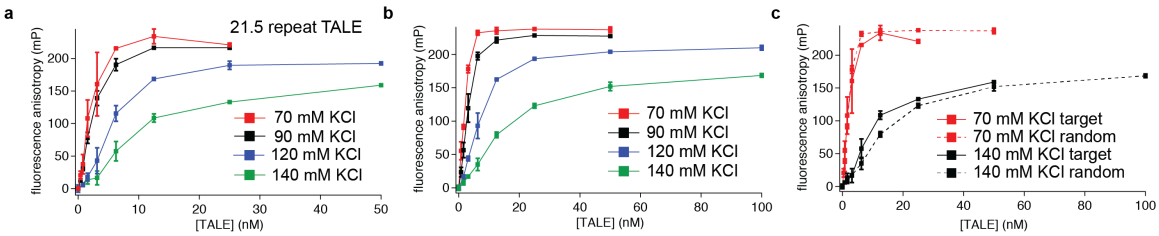

Figure 4.2 Effects of increasing concentrations of monovalent ions on 21.5 repeat TALE binding. The binding of the 21.5 repeat TALE construct to (a) target and (b) random DNA sequences was measured via fluorescence anisotropy. Binding was measured in 20 mM Tris buffer with different amounts of KCl added. DNA substrate was held constant at 1 nM and concentrations of the 21.5 repeat TALE were modulated. Comparing the binding of both constructs directly at high and low ionic strength (c) reveals that even at high ionic strength (140 mM KCl) there is little difference between target and random binding.

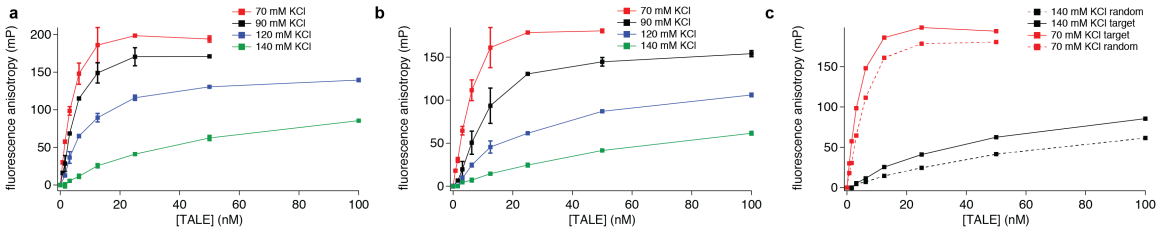

Figure 4.3 Effects of increasing concentrations of monovalent ions on 15.5 repeat TALE binding. The binding of the 15.5 repeat TALE construct to (a) target and (b) random DNA sequences was measured via fluorescence anisotropy. Binding was measured in 20 mM Tris buffer with different amounts of KCl added. DNA substrate was held constant at 1 nM and concentrations of the 15.5 repeat TALE were modulated. Comparing the binding of both constructs directly at high and low ionic strength (c) reveals that even at high ionic strength (140 mM KCl) there is little difference between target and random binding.

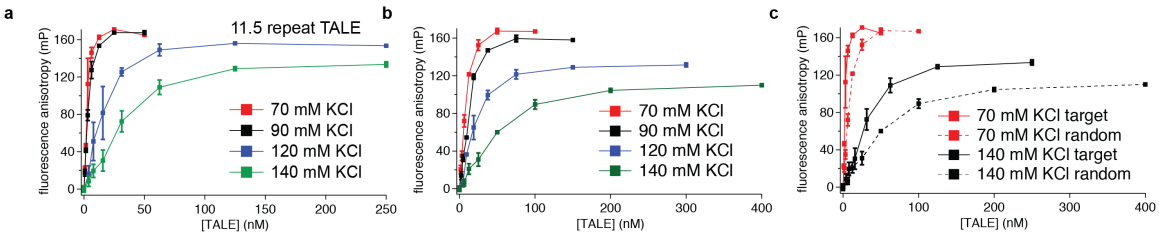

Figure 4.4 Effects of increasing concentrations of monovalent ions on 11.5 repeat TALE binding. The binding of the 11.5 repeat TALE construct to (a) target and (b) random DNA sequences was measured via fluorescence anisotropy. Binding was measured in 20 mM Tris buffer with different amounts of KCl added. DNA substrate was held constant at 1 nM and concentrations of the 11.5 repeat TALE were modulated. Comparing the binding of both constructs directly at high and low ionic strength (c) reveals that at high ionic strength (140 mM KCl) there is a slight difference between target and random binding.

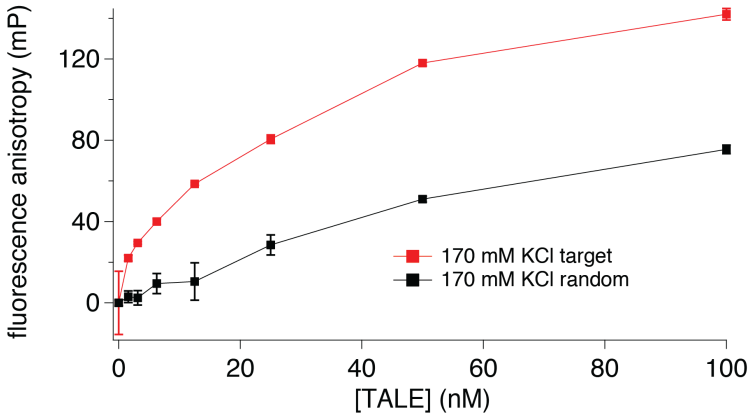

Figure 4.5 Binding of the 21.5 repeat TALE construct to target and random DNA substrates at a supraphysiological concentration of KCl. The binding of the 21.5 repeat TALE construct to target (red) and random (black) DNA sequences was measured via fluorescence anisotropy. Binding was measured in 20 mM Tris buffer with 150 mM KCl added. DNA substrate was held constant at 1 nM and concentrations of the 21.5 repeat TALE were modulated.

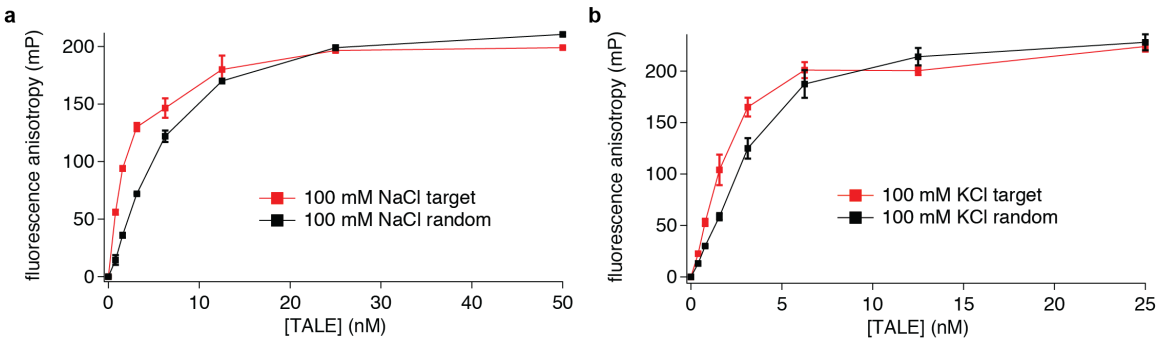

Figure 4.6 Effects of changing the monovalent cation on 21.5 repeat TALE binding. The binding of the 21.5 repeat TALE construct to target (red) and random (black) DNA was measured via fluorescence anisotropy. Binding was measured in 20 mM Tris buffer with 100 mM (a) NaCl or (b) KCl. No appreciable difference was observed in the binding behavior of the TALE to target or random DNA when the monovalent cation was changed from K^+ to Na^+.

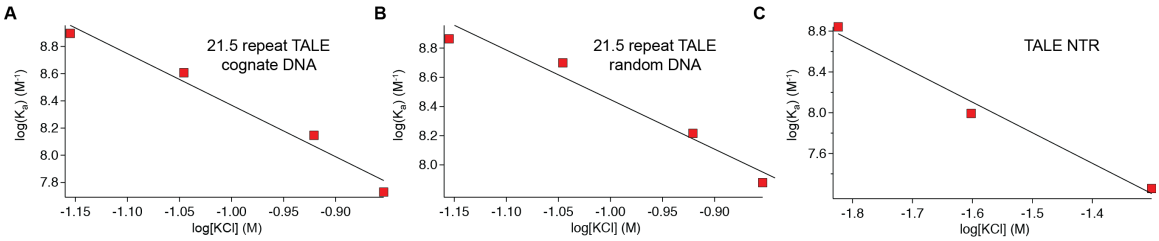

Figure 4.7 Binding affinity of TALE proteins as a function of monovalent salt concentration. Log-log plots of binding affinity versus KCl concentration for the 21.5 repeat TALE binding to: (A) random DNA and (B) target DNA. Counter-ion condensation theory provides a framework for interpreting the monovalent salt dependence of TALE binding, demonstrating a higher proportion of non-electrostatic binding energy for target DNA. (C) Log-log plot of binding affinity versus KCl concentration for the TALE NTR, which suggests that electrostatics dominate the binding of the NTR to DNA.

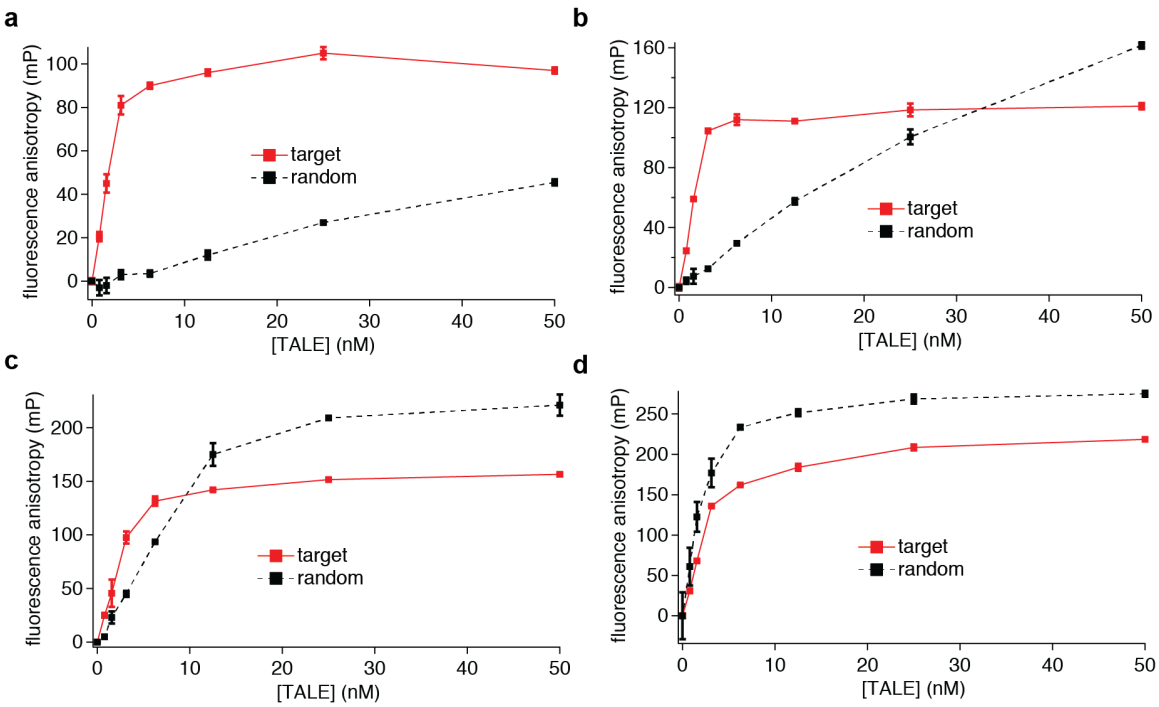

Figure 4.8 Effects of different divalent cations on 21.5 repeat TALE binding. The binding of the 21.5 repeat TALE construct to target (red) and random (black) DNA was measured via fluorescence anisotropy. Binding was measured in 20 mM Tris buffer with 100 KCl and 10 mM divalent cation. DNA concentration was held constant at 1 nM. Divalent cations tested included (a) magnesium, (b) calcium, (c) strontium, and (d) zinc. The chloride salt of each cation was used. These results demonstrate that both magnesium and calcium, the two divalent cations in highest concentrations within cells, induce the greatest difference in binding specificity for the 21.5 repeat TALE construct.

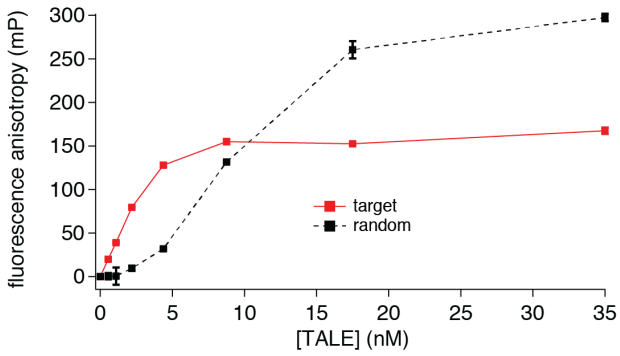

Figure 4.9 Effect of manganese on 21.5 repeat TALE binding. The binding of the 21.5 repeat TALE construct to target (red) and random (black) DNA was measured via fluorescence anisotropy. Binding was measured in 20 mM Tris buffer with 100 KCl and 10 mM manganese cation. DNA concentration was held constant at 1 nM. The chloride salt of each manganese was used. These results demonstrate that manganese has only a moderate effect on the binding specificity for the 21.5 repeat TALE construct.

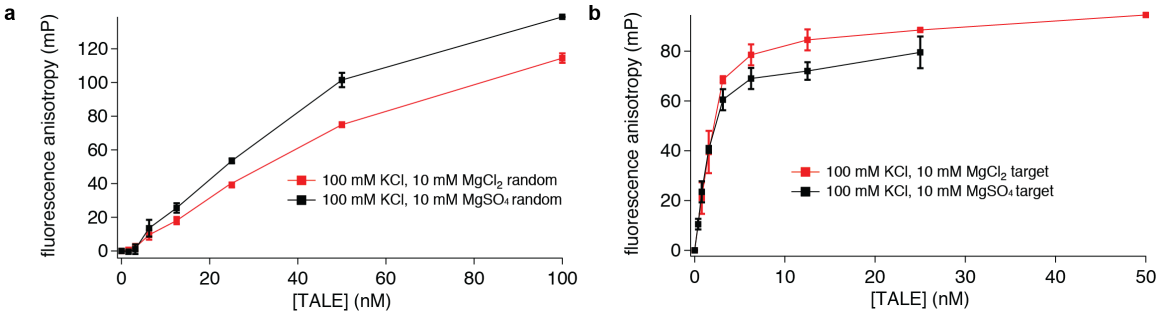

Figure 4.10 Effect of divalent versus monovalent anions on 21.5 repeat TALE binding. The binding of the 21.5 repeat TALE construct to (a) random and (b) target DNA was measured via fluorescence anisotropy. Binding was measured in 20 mM Tris buffer with 100 KCl and 10 mM magnesium. DNA concentration was held constant at 1 nM. The chloride salt (red in plots) and sulfate salt (black in plots) were compared. These results demonstrate that divalent anions, in contrast to divalent cations, have little effect on TALE binding when compared to monovalent anions.

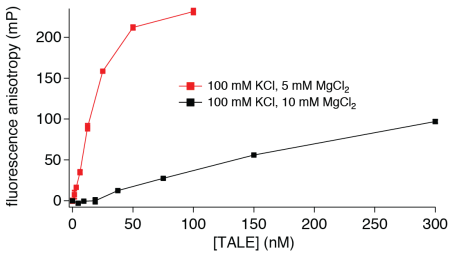

Figure 4.11 Increasing the overall ionic strength of solution via addition of $MgCl_2$ significantly reduces nonspecific binding affinity of the 21.5 repeat TALE. Binding of the 21.5 repeat TALE to the target DNA substrate was measured using fluorescence anisotropy, in 20 mM Tris buffer and 100 mM KCl with the DNA concentration held constant at 1 nM. $MgCl_2$ was added to 5 mM (red line) and 10 mM (black line) final concentrations. While binding to the random substrate is decreased by any increase in ionic strength, the increase here is much greater than for monovalent salt.

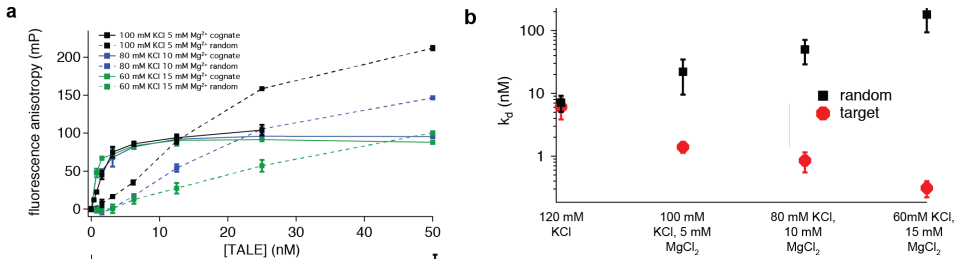

Figure 4.12 Increasing the proportion of $MgCl_2$ relative to KCl while holding total ionic strength constant increases the difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$ for the 21.5 repeat TALE construct. (a) The binding of the 21.5 repeat TALE to target (solid lines) and random (dashed lines) DNA was measured in 20 mM Tris buffer, with the total ionic strength of added salts (not including the buffer) held constant at 120 mM. DNA was held constant at 1 nM while the TALE concentration was varied. This 120 mM of added salt was a varied proportion of KCl:$MgCl_2$. (b) As the proportion of ionic strength contributed by $MgCl_2$ was increased, the difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$ increased significantly as $k_{d,\ specific}$ decreased slightly while $k_{d,\ non\text{-}specific}$ increased rapidly.

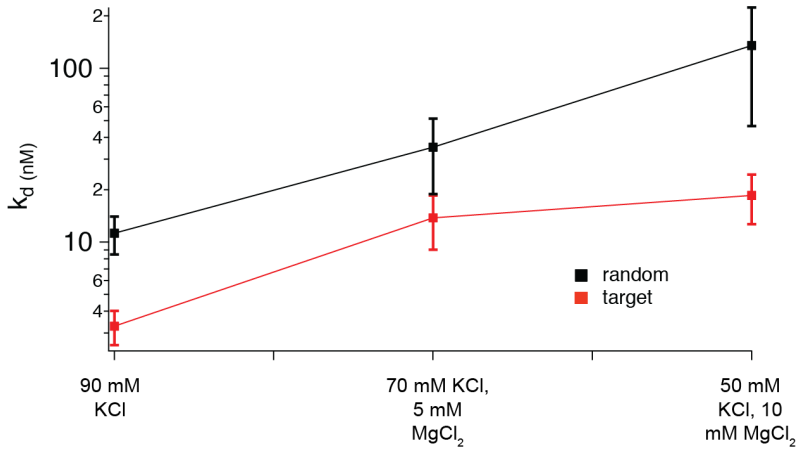

Figure 4.13 Increasing the proportion of $MgCl_2$ relative to KCl while holding total ionic strength constant increases the difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$ for the 15.5 repeat TALE construct. The binding of the 15.5 repeat TALE to target and random DNA was measured in 20 mM Tris buffer, with the total ionic strength of added salts (not including the buffer) held constant at 90 mM. DNA was held constant at 1 nM while the TALE concentration was varied. This 90 mM of added salt was a varied proportion of KCl:$MgCl_2$. As the proportion of ionic strength contributed by $MgCl_2$ was increased, the difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$ increased as $k_{d,\ non\text{-}specific}$ increased more quickly than $k_{d,\ specific}$.

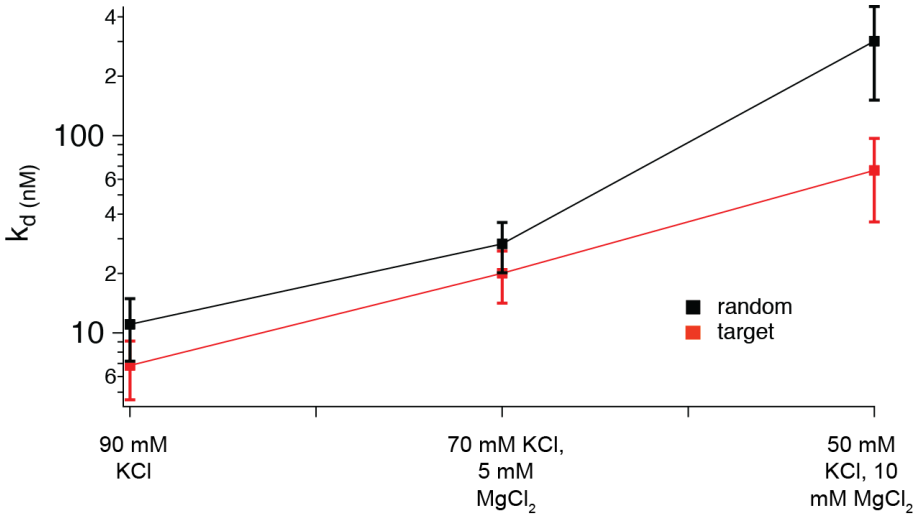

Figure 4.14 Increasing the proportion of $MgCl_2$ relative to KCl while holding total ionic strength constant increases the difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$ for the 11.5 repeat TALE construct. The binding of the 11.5 repeat TALE to target and random DNA was measured in 20 mM Tris buffer, with the total ionic strength of added salts (not including the buffer) held constant at 90 mM. DNA was held constant at 1 nM while the TALE concentration was varied. This 90 mM of added salt was a varied proportion of KCl:$MgCl_2$. As the proportion of ionic strength contributed by $MgCl_2$ was increased, the difference between $k_{d,\ specific}$ and $k_{d,\ non\text{-}specific}$ increased as $k_{d,\ non\text{-}specific}$ increased more quickly than $k_{d,\ specific}$.

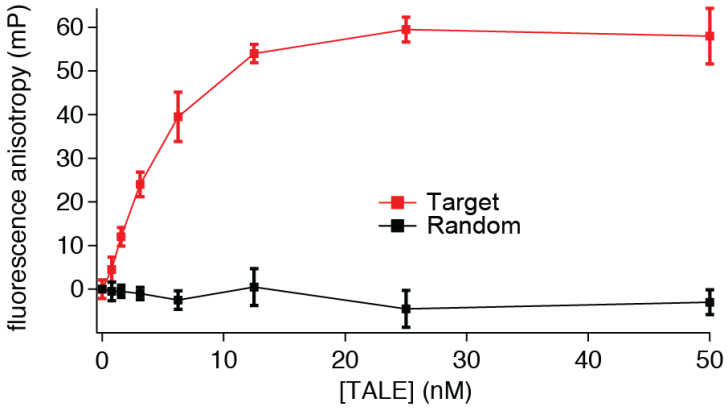

Figure 4.15 21.5 repeat TALE binding at physiological conditions shows high specificity. Binding of the 21.5 repeat TALE to the target DNA substrate was measured using fluorescence anisotropy, in 20 mM Tris buffer, 150 mM KCl, 10 mM $MgCl_2$. The DNA concentration was held constant at 1 nM. While binding to the random substrate (black) is destroyed in this condition, the binding to the target substrate (red) remains strong.

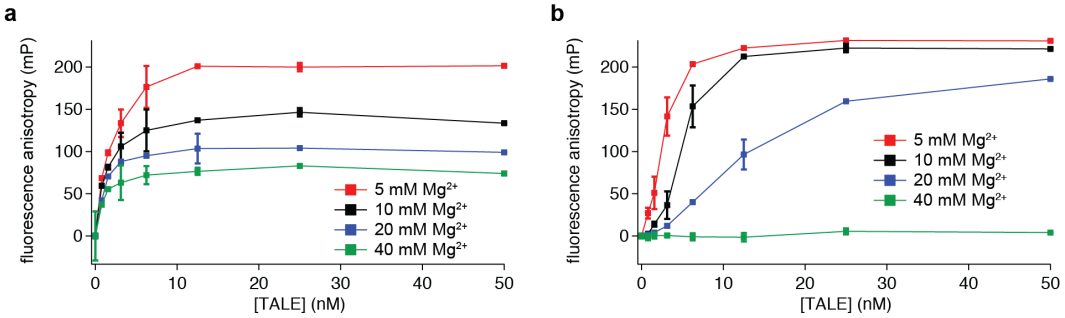

Figure 4.16 Binding of the 21.5 repeat TALE in the presence of $MgCl_2$ but not monovalent salt. Binding of the 21.5 repeat TALE to the (a) target and (b) random DNA substrate was measured via fluorescence anisotropy in 20 mM Tris buffer and increasing concentrations of $MgCl_2$. No KCl or other monovalent salt was added. DNA concentration was held constant at 1 nM. While binding to target DNA showed a decrease in binding constant, binding to random DNA was rapidly abolished beyond 20 mM $MgCl_2$.

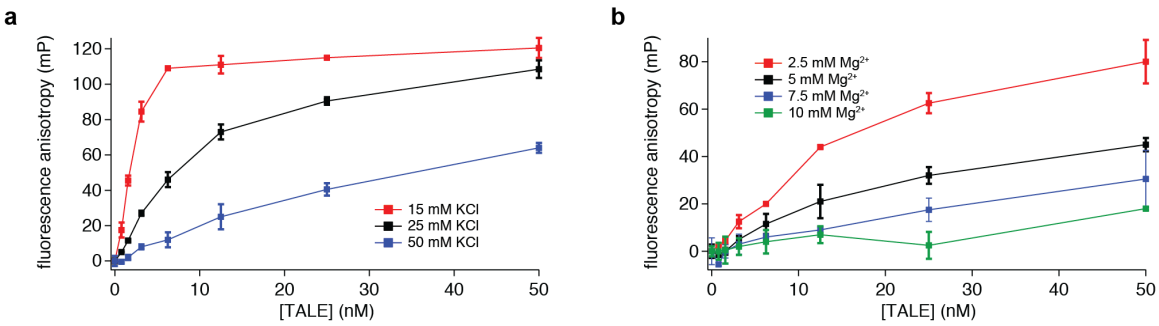

Figure 4.17 Binding of the TALE NTR to random DNA in (a) KCl and (b) $MgCl_2$. Binding of the TALE NTR only construct (without a CRD or CTR) to the random DNA substrate was measured via fluorescence anisotropy in 20 mM Tris buffer and increasing concentrations of either (a) KCl or (b) $MgCl_2$. DNA concentration was held constant at 1 nM. While binding to showed a steady decline at moderate KCl concentrations, binding was more rapidly diminished in $MgCl_2$, even taking into account the total ionic strength.

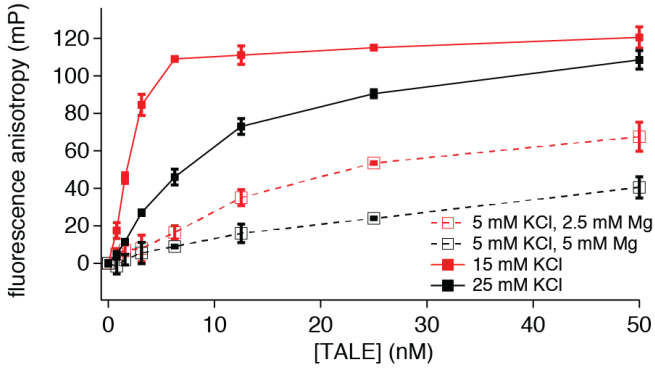

Figure 4.18 Binding of the TALE NTR to random DNA in KCl and mixed KCl/$MgCl_2$ solutions. Binding of the TALE NTR only construct (without a CRD or CTR) to the random DNA substrate was measured via fluorescence anisotropy in 20 mM Tris buffer with either KCl only added (solid lines) or KCl and $MgCl_2$ added (dashed lines). Two ionic strengths were tested, 15 mM and 25 mM. Even a very small addition of $MgCl_2$ (2.5 mM) greatly diminishes NTR binding.

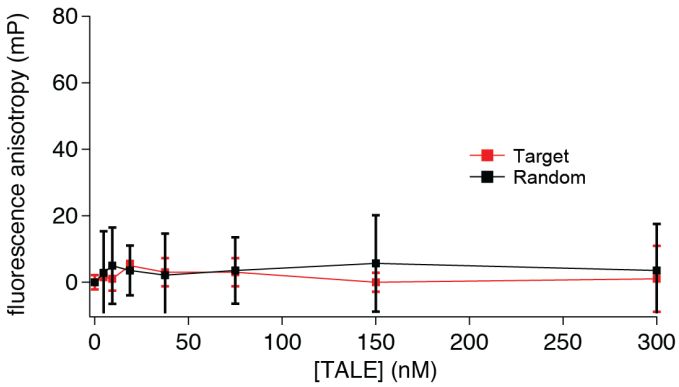

Figure 4.19 The TALE CRD+CTR-only construct shows no binding activity. Binding of the 21.5 repeat TALE CRD and CTR (the 21.5 repeat construct without the NTR) to the target DNA substrate (red line) and random DNA substrate (black line) was measured using fluorescence anisotropy, in 20 mM Tris buffer, 100 mM KCl, 5 mM $MgCl_2$. The DNA concentration was held constant at 1 nM. No binding was observed to either substrate, with similar results in the absence of $MgCl_2$. This result highlights the criticality of the NTR in nucleating initial binding and/or maintaining the structural integrity of the full-length TALE.

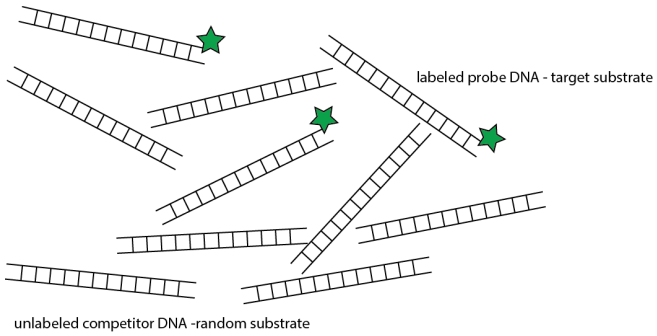

Figure 4.20 Schematic illustration of competitor binding assay. A large excess (100-fold, or 100 nM total) of the random DNA substrate is added to the standard 1 nM target DNA substrate. The random DNA has no fluorescent dye conjugated, and thus only serves to sequester TALEs away from the reporter substrate, here the labeled target substrate.

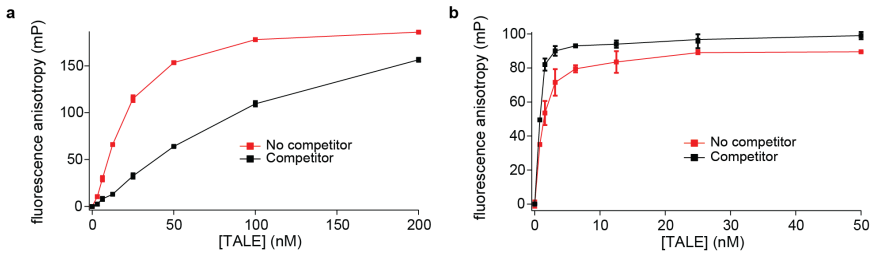

Figure 4.21 Binding of the 21.5 repeat TALE construct to the target DNA substrate in the presence of excess unlabeled random DNA substrate. The binding of the 21.5 repeat TALE to the target substrate was measured in the presence (black lines) and absence (red lines) of 100 nM random DNA that had no fluorescent dye attached. The labeled target substrate was held constant at 1 nM concentration. Binding was measured via fluorescence anisotropy in 20 mM Tris buffer with either (a) 100 mM KCl or (b) 100 mM KCl and 10 mM $MgCl_2$. While the nonspecific random substrate is able to sequester away TALE with only KCl in solution, when $MgCl_2$ was added, there was no significant increase in the measured dissociation constant for target binding.

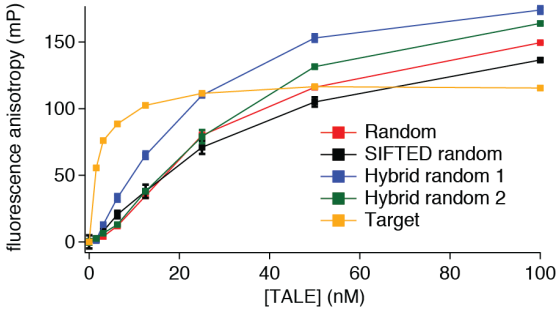

Figure 4.22 Binding of the 21.5 repeat TALE construct to a variety of DNA substrates. Binding of the 21.5 repeat TALE to the target DNA substrate (red line) and random DNA substrate (gold line), the random substrate (red), a random substrate generated via the SIFTED platform (black), and two half-target match sequences (blue and green) was measured using fluorescence anisotropy, in 20 mM Tris buffer, 100 mM KCl, 5 mM $MgCl_2$. The DNA concentration was held constant at 1 nM.

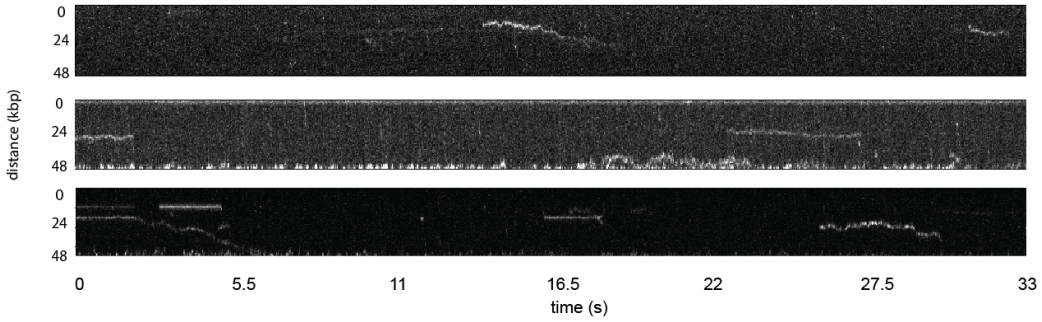

Figure 4.23 Representative kymographs of the 21.5 repeat TALE construct diffusing on non-specific DNA templates (lambda-phage DNA, 48.5 kbp) at 250 pM protein concentration in mixed monovalent/divalent cation buffer. The Cy3-conjugated 21.5 repeat TALE at 250 pM concentration was introduced into flow cells containing single tethered non-specific DNA (0 binding sites) in the presence of fluid flow for 2 minutes. The imaging buffer (50 mM MOPS, pH 8.1) was comprised of 130 mM KCl and 5 mM $MgCl_2$, providing total added ionic strength of 150 mM.

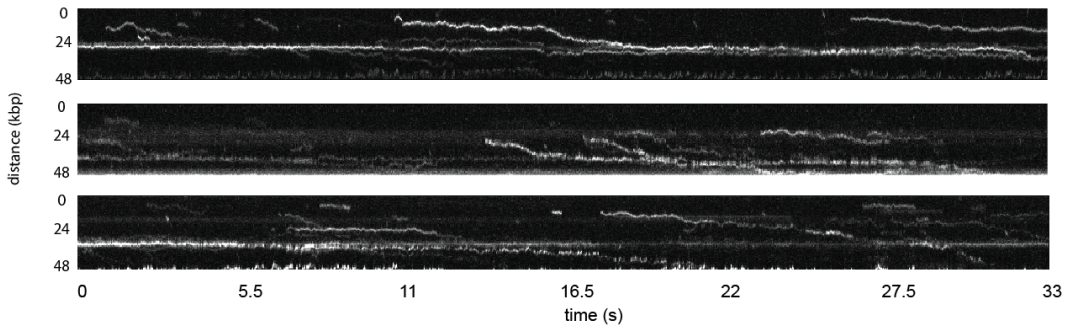

Figure 4.24 Representative kymographs of the 21.5 repeat TALE construct diffusing on non-specific DNA templates (lambda-phage DNA, 48.5 kbp) at 250 pM protein concentration in monovalent cation buffer. The Cy3-conjugated 21.5 repeat TALE at 250 pM concentration was introduced into flow cells containing single tethered non-specific DNA (0 binding sites) in the presence of fluid flow for 2 minutes. The imaging buffer (50 mM MOPS, pH 8.1) was comprised of 150 mM KCl, providing total added ionic strength of 150 mM.

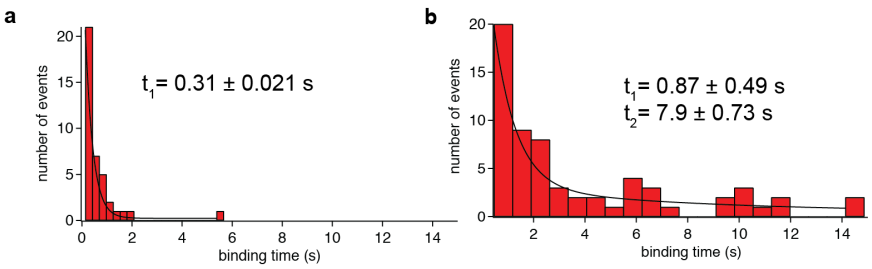

Figure 4.25 Binding lifetime distributions for the 21.5 repeat TALE construct on non-specific DNA at 110 mM total added ionic strength. (a) Distribution of binding lifetimes of the 21.5 repeat TALE in 90 mM KCl and 5 mM $MgCl_2$, which is best fit to a single exponential decay function. (b) Distribution of binding lifetimes of the 21.5 repeat TALE in 110 mM KCl, which is best fit to a double exponential decay function.

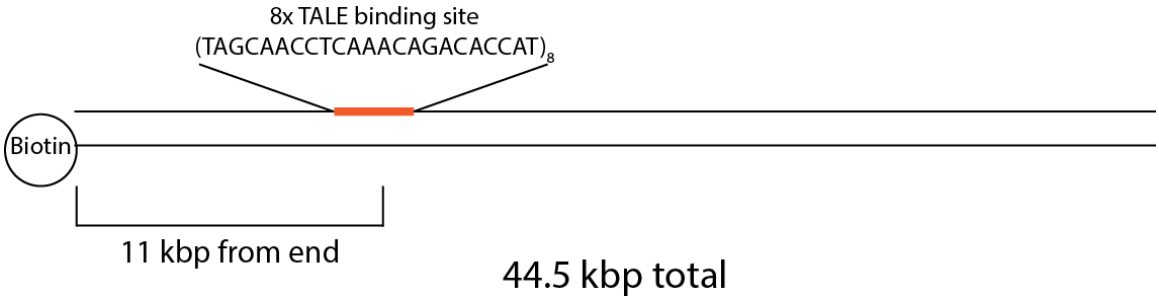

Figure 4.26 Schematic overview of specific DNA template used in single molecule studies. This DNA template has an 8x array of binding sites for the 21.5 repeat TALE construct arranged in tandem. The binding array is located ~11 kbp from the tether point, approximately one quarter of the distance to the end of the 44.5 kbp substrate.

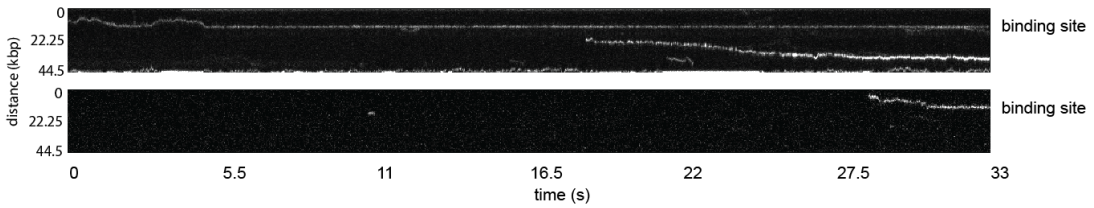

Figure 4.27 Kymographs depicting TALE target search and localization in the presence of 110 mM KCl only (no divalent salt). The 21.5 repeat TALE (labeled with Cy3) was added to the microfluidic sample cell. The 8x binding array DNA templates were extended under flow and the motion of TALEs on DNA was recorded. In these examples, the proteins bind DNA at a non-target site and then carry out 1-D search until arriving at a target site. In the top example, we observe a long-lived non-specific search event, beginning at ~17 seconds near 25 kbp from the tether, and this event does not yield successful target site localization. In both example traces shown here, we observe that TALEs appear to recognize their binding site upon first encounter.

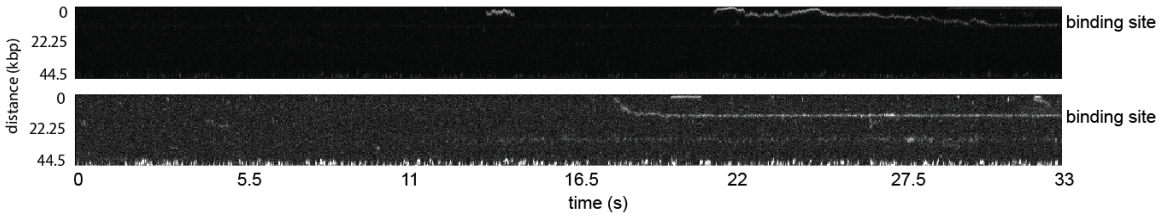

Figure 4.28 Kymographs depicting TALE target search and localization in the presence of 90 mM KCl and 5 mM $MgCl_2$. The 21.5 repeat TALE (labeled with Cy3) was added to the microfluidic sample cell. The 8x binding array DNA templates were extended under flow and the motion of TALEs on DNA was recorded. In both examples shown here, the proteins bind DNA at a non-target site and then carry out 1-D search until arriving at their target site.

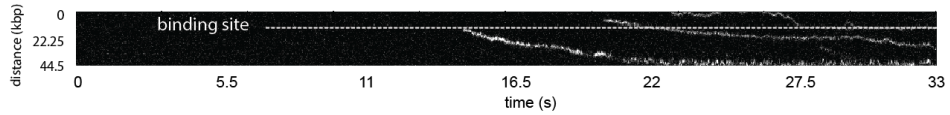

Figure 4.29 Kymograph depicting TALE target search and unsuccessful target localization in the presence of 110 mM KCl. The 21.5 repeat TALE (labeled with Cy3) was added to the microfluidic sample cell. The 8x binding array DNA templates were extended under flow and the motion of TALEs on DNA was recorded. In this example, proteins bind DNA at a non-target site around t = 20 seconds, followed by 1-D search without binding to the target site. Around t = 14.5 seconds, we observe sudden release of a TALE from the target site, which could be due to Cy3 dye quenching due to dye-dye interactions with nearby Cy3-TALEs. Additionally, this event could be binding of the TALE at an off-target site such that it misses the 8x binding array and simply diffuses away.

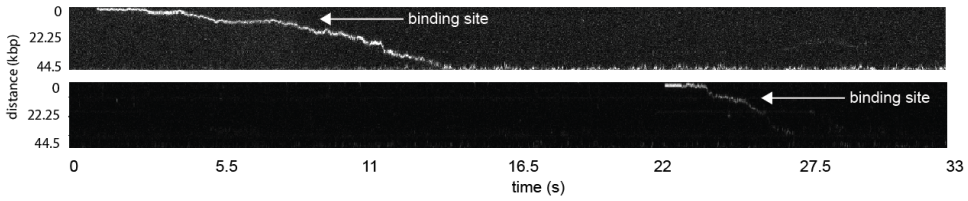

Figure 4.30 Kymographs depicting TALE target search and unsuccessful or temporary target localization in the presence of 90 mM KCl and 5 mM $MgCl_2$. The 21.5 repeat TALE (labeled with Cy3) was added to the microfluidic sample cell. The 8x binding array DNA templates were extended under flow and the motion of TALEs on DNA was recorded. In this example, TALE proteins initially bind DNA at non-target sites, and then carry out a 1-D search. In both cases shown here, however, the TALEs either transiently bind a target site for a few seconds or less or completely fail to stably localize at the target.

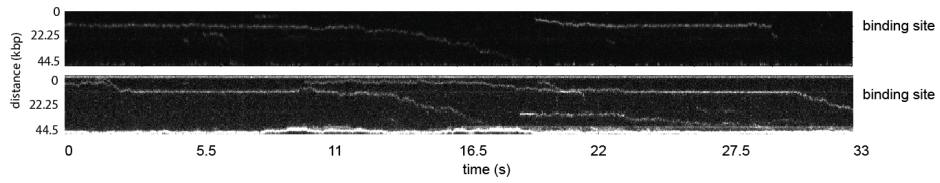

Figure 4.31 Kymographs depicting TALE target search, target binding, and target release in the presence of 110 mM KCl. The 21.5 repeat TALE (labeled with Cy3) was added to the microfluidic sample cell. The 8x binding array DNA templates were extended under flow and the motion of TALEs on DNA was recorded. In this example, the proteins are bound at target sites, but then release and continue to carry out non-specific search. In the top example, the second target release event occurring at ~30 seconds is followed by a brief non-specific search event.

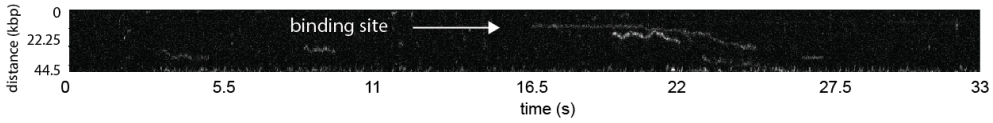

Figure 4.32 Kymograph depicting TALE target release in the presence of 90 mM KCl and 5 mM $MgCl_2$. The 21.5 repeat TALE (labeled with Cy3) was added to the microfluidic sample cell. The 8x binding array DNA templates were extended under flow and the motion of TALEs on DNA was recorded. In this example, there are brief non-specific binding events in the initial 15 seconds of the experiment, followed by the rapid appearance of a specifically bound TALE at the target site array. This target-bound TALE is then released from the target site around t = 20 seconds, followed by 1-D diffusion for several seconds before either releasing from the DNA template or undergoing a photobleaching event. Around t = 21.5 seconds, there appears to be a collision between two non-specifically bound TALEs, the TALE released and another TALE.

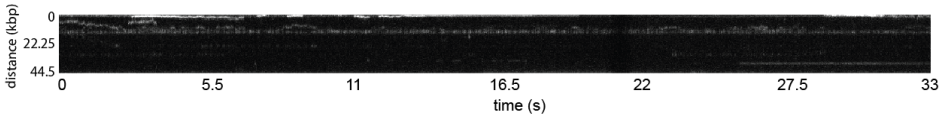

Figure 4.33 Kymograph depicting TALE-TALE collisions in the presence of 110 mM KCl. The 21.5 repeat TALE (labeled with Cy3) was added to the microfluidic sample cell. The 8x binding array DNA templates were extended under flow and the motion of TALEs on DNA was recorded. In this example, there is at least one TALE bound at the target site continuously throughout the entire duration of the experiment (33 seconds). Numerous other TALEs non-specifically bind DNA at different time points and engage in a 1-D search process. Several of these proteins appear to collide with the target-bound TALE and often reverse direction, possible 'bouncing off' the target-bound TALE.

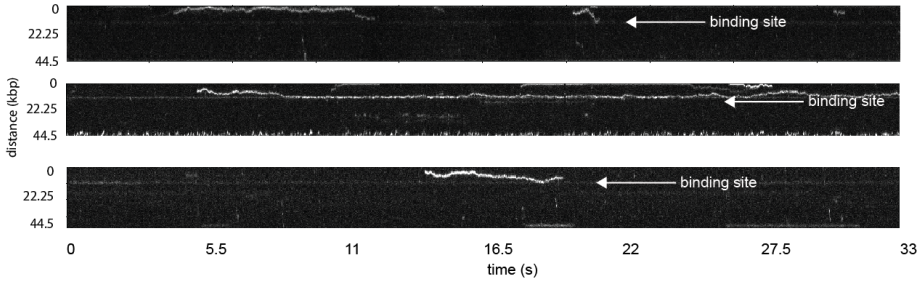

Figure 4.34 Kymographs depicting putative TALE-TALE collisions in the presence of 90 mM KCl and 5 mM $MgCl_2$. The 21.5 repeat TALE (labeled with Cy3) was added to the microfluidic sample cell. The 8x binding array DNA templates were extended under flow and the motion of TALEs on DNA was recorded. In these examples, there is at least one TALE bound at the target site continuously throughout the duration of the imaging experiment (33 seconds). Numerous other TALEs non-specifically bind DNA at different time points and engage in a 1-D search process. Several of these proteins appear to collide with the target-bound TALE and often reverse direction (i.e. 'bouncing off' the target-bound TALE)

Table 4.1 Binding constants for the 21.5 repeat TALE and target DNA substrate.

[KCl] (mM)	[$MgCl_2$] (mM)	K_d (nM)	K_d error (± nM)
70	0	1.37	0.45
90	0	2	0.5
120	0	6.08	2.24
140	0	13.23	4.03
170	0	30.06	6.94
100	5	1.4	0.27
100	10	0.98	0.16
80	10	0.85	0.3
60	15	0.31	0.09
0	5	1.2	0.2
0	10	0.72	0.15
0	20	0.45	0.07
0	40	0.4	0.11

Table 4.2 Binding constants for the 21.5 repeat TALE and random DNA substrates.

[KCl] (mM)	[$MgCl_2$] (mM)	DNA	K_d (nM)	K_d error (± nM)
70	0	random	1.27	0.42
90	0	random	2.47	0.82
120	0	random	7.13	2.03
140	0	random	18.67	3.82
170	0	random	124.31	37.98
100	5	random	29.11	12.56
100	10	random	802.59	327.37
80	10	random	49.94	20.98
60	15	random	179.92	86.15
0	5	random	2.19	0.81
0	10	random	5.43	2.95
0	20	random	23.51	8.35
0	40	random	--	--
100	5	SIFTED	45.36	4.54
100	5	1AB	26.12	4.35
100	5	2AB	54.9	13.6

Table 4.3 Binding constants for the 15.5 repeat TALE and target DNA substrate.

[KCl] (mM)	[$MgCl_2$] (mM)	K_d (nM)	K_d error (± nM)
70	0	2.11	0.4
90	0	3.28	0.73
120	0	7.97	0.97
140	0	24.54	0.58
70	5	13.76	4.73
50	10	18.53	5.88

Table 4.4 Binding constants for the 15.5 repeat TALE and random DNA substrate.

[KCl] (mM)	[$MgCl_2$] (mM)	K_d (nM)	K_d error (± nM)
70	0	4.88	1.16
90	0	11.24	2.77
120	0	50.4	7.92
140	0	181.4	20.91
70	5	35.08	16.21
50	10	134.96	88.52

Table 4.5 Binding constants for the 11.5 repeat TALE and target DNA substrate.

[KCl] (mM)	[$MgCl_2$] (mM)	K_d (nM)	K_d error (± nM)
70	0	1.88	0.64
90	0	6.83	2.25
120	0	14.66	3.66
140	0	38	10.38
70	5	20.05	5.9
50	10	66.65	30.15

Table 4.6 Binding constants for the 11.5 repeat TALE and random DNA substrate.

[KCl] (mM)	[$MgCl_2$] (mM)	K_d (nM)	K_d error (± nM)
70	0	7.89	1.84
90	0	11.05	3.86
120	0	20.71	5.44
140	0	55.84	11.4
70	5	28.27	8.07
50	10	301.24	289.1

Table 4.7 Results derived from counter-ion condensation theory and linear regression fits to plots of log(K_a) versus log[KCl]. These results summarize the number of ions displaced (*N*) and number of phosphate groups contacted by TALEs (*Z*) when binding the specified DNA substrates, as well as the energetic contributions from non-electrostatic interactions.

Parameter	21.5 repeat TALE, target DNA	21.5 repeat TALE, random DNA	TALE NTR
N	3.3 ± 0.5	3.8 ± 0.4	3.0 ± 0.4
Z	5.2	6	4.7
log(K^a_{nel})	5.1 ± 0.5	4.6 ± 0.5	3.3 ± 0.5
ΔG_{nel} (kcal/mol)	-7	-6.2	-4.5

Table 4.8 Binding constants for the 21.5 repeat TALE construct and target/random DNA substrate with various divalent cations.

[KCl] mM	**[divalent cation] (mM)**	**divalent cation**	**DNA**	**K_d (nM)**	**K_d error (± nM)**
100	5	Ca2+	Target	0.89	0.34
100	5	Ca2+	Random	76.62	7.07
100	5	Zn2+	Target	2.17	0.45
100	5	Zn2+	Random	1.48	0.16
100	5	Sr2+	Target	2	0.5
100	5	Sr2+	Random	8.87	2.68

Table 4.9 Binding constants for the TALE NTR and target DNA substrate.

[KCl] (mM)	**[$MgCl_2$] (mM)**	**K_d (nM)**	**K_d error (± nM)**
50	0	55.37	9.77
25	0	10.18	1.06
15	0	1.44	0.44
5	5	55.9	19.23
5	2.5	26.06	6.32
0	2.5	21.72	4.02
0	5	27.59	5.23
0	7.5	72.34	54.93
0	10	--	--

4.6 References

1. Loayza, D., Parsons, H., Donigian, J., Hoke, K. & de Lange, T. DNA Binding Features of Human POT1: a nonamer 5'-TAGGGTTAG-3' minimal binding site, sequence specificity, and internal binding to multimeric sites. *J. Biol. Chem.* **279,** 13241–13248 (2004).

2. Yang, Y., Sass, L. E., Du, C., Hsieh, P. & Erie, D. A. Determination of protein-DNA binding constants and specificities from statistical analyses of single molecules: MutS-DNA interactions. *Nucleic Acids Res.* **33,** 4322–4334 (2005).

3. Jen-jacobson, L., Engler, L. E., Ames, J. T., Kurpiewski, M. R. & Grigorescu, A. Thermodynamic Parameters of Specific and Nonspecific Protein-DNA Binding. *Supramol. Chem.* **12,** 143–160 (2000).

4. Terry, B. J., Jack, W. E., Rubin, R. A. & Modrich, P. Thermodynamic parameters governing interaction of EcoRI endonuclease with specific and nonspecific DNA sequences. *J. Biol. Chem.* **258,** 9820–9825 (1983).

5. Takeda, Y., Ross, P. D. & Mudd, C. P. Thermodynamics of Cro protein-DNA interactions. *Proc. Natl. Acad. Sci.* **89,** 8180–8184 (1992).

6. Privalov, P. L., Dragan, A. I. & Crane-Robinson, C. Interpreting protein/DNA interactions: Distinguishing specific from non-specific and electrostatic from non-electrostatic components. *Nucleic Acids Res.* **39,** 2483–2491 (2011).

7. Chiu, T. K. & Dickerson, R. E. 1 A crystal structures of B-DNA reveal sequence-specific binding and groove-specific bending of DNA by magnesium and calcium. *J. Mol. Biol.* **301,** 915–945 (2000).

8. Xiong, K. & Blainey, P. C. Molecular sled sequences are common in mammalian proteins. *Nucleic Acids Res.* gkw035 (2016). doi:10.1093/nar/gkw035

9. Cravens, S. L., Hobson, M. & Stivers, J. T. Electrostatic properties of complexes along a DNA glycosylase damage search pathway. *Biochemistry* **53,** 7680–7692 (2014).

10. Yin, P. *et al.* Specific DNA-RNA Hybrid Recognition by TAL Effectors. *Cell Rep.* **2,** 707–713 (2012).

11. Doyle, E. L. *et al.* TAL effector specificity for base 0 of the DNA target is altered in a complex, effector- and assay-dependent manner by substitutions for the tryptophan in cryptic repeat -1. *PLoS One* **8,** 1–17 (2013).

12. Römer, P., Recht, S. & Lahaye, T. A single plant resistance gene promoter engineered to recognize multiple TAL effectors from disparate pathogens. *Proc.*

Natl. Acad. Sci. U. S. A. **106,** 20526–20531 (2009).

13. Gao, H., Wu, X., Chai, J. & Han, Z. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.* **22,** 1716–20 (2012).

14. Juillerat, A. *et al.* Comprehensive analysis of the specificity of transcription activator-like effector nucleases. *Nucleic Acids Res.* **42,** 5390–5402 (2014).

15. Wicky, B. I. M., Stenta, M. & Dal Peraro, M. TAL Effectors Specificity Stems from Negative Discrimination. *PLoS One* **8,** e80261 (2013).

16. Bedell, V. M. *et al.* In vivo genome editing using a high-efficiency TALEN system. *Nature* **491,** 114–8 (2012).

17. Ding, Q. *et al.* A TALEN Genome-Editing System for Generating Human Stem Cell-Based Disease Models. *Cell Stem Cell* **12,** 238–251 (2013).

18. Sung, Y. H. *et al.* Knockout mice created by TALEN-mediated gene targeting. *Nat. Biotechnol.* **31,** 23–24 (2013).

19. Zu, Y. *et al.* TALEN-mediated precise genome modification by homologous recombination in zebrafish. *Nat Methods* **10,** 329–331 (2013).

20. Christian, M. L. *et al.* Targeting G with TAL Effectors: A Comparison of Activities of TALENs Constructed with NN and NK Repeat Variable Di-Residues. *PLoS One* **7,** (2012).

21. Mahfouz, M. M. *et al.* De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 2623–8 (2011).

22. Meckler, J. F. *et al.* Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic Acids Res.* **41,** 4118–4128 (2013).

23. Valton, J. *et al.* Overcoming transcription activator-like effector (TALE) DNA binding domain sensitivity to cytosine methylation. *J. Biol. Chem.* **287,** 38427–38432 (2012).

24. Schreiber, T. *et al.* Refined Requirements for Protein Regions Important for Activity of the TALE AvrBs3. *PLoS One* **10,** e0120214 (2015).

25. Richter, A., Streubel, J. & Boch, J. *TALENs*. *TALENs, Methods in Molecular Biology* **1338,** (Springer New York, 2016).

26. O'Gorman, R. B., Dunaway, M. & Matthews, K. S. DNA binding characteristics of lactose repressor and the trypsin-resistant core repressor. *J. Biol. Chem.* **255,**

10100–10106 (1980).

27. Record, M. T., deHaseth, P. L. & Lohman, T. M. Interpretation of monovalent and divalent cation effects on the lac repressor-operator interaction. *Biochemistry* **16,** 4791–4796 (1977).

28. Weinberg, R. L., Veprintsev, D. B. & Fersht, A. R. Cooperative binding of tetrameric p53 to DNA. *J. Mol. Biol.* **341,** 1145–1159 (2004).

29. Cuculis, L., Abil, Z., Zhao, H. & Schroeder, C. M. Direct observation of TALE protein dynamics reveals a two-state search mechanism. *Nat. Commun.* **6,** 7277 (2015).

30. Privalov, P. L., Dragan, A. I. & Crane-Robinson, C. Interpreting protein/DNA interactions: distinguishing specific from non-specific and electrostatic from non-electrostatic components. *Nucleic Acids Res.* **39,** 2483–2491 (2011).

31. Moll, J. R., Acharya, A., Gal, J., Mir, A. a & Vinson, C. Magnesium is required for specific DNA binding of the CREB B-ZIP domain. *Nucleic Acids Res.* **30,** 1240–1246 (2002).

32. Rogers, J. M. *et al.* Context influences on TALE–DNA binding revealed by quantitative profiling. *Nat. Commun.* **6,** 7440 (2015).

33. Guilinger, J. P. *et al.* Broad specificity profiling of TALENs results in engineered nucleases with improved DNA-cleavage specificity. *Nat. Methods* **11,** 429–435 (2014).

34. Dosanjh, N. S., Hammerbacher, N. a. & Michel, S. L. J. Characterization of the Helicobacter pylori NikR-PureA DNA Interaction: Metal ion requirements and sequence specificity. *Biochemistry* **46,** 2520–2529 (2007).

35. Schumacher, M. A., Goodman, R. H. & Brennan, R. G. The structure of a CREB bZIP somatostatin CRE complex reveals the basis for selective dimerization and divalent cation-enhanced DNA binding. *J. Biol. Chem.* **275,** 35242–35247 (2000).

36. Viadiu, H. & Aggarwal, A. K. Structure of BamHI Bound to Nonspecific DNA. *Mol. Cell* **5,** 889–895 (2000).

37. Thielking, V. *et al.* Magnesium(2+) confers DNA binding specificity to the EcoRV restriction endonuclease. *Biochemistry* **31,** 3727–3732 (1992).

38. Jiménez-Menéndez, N. *et al.* Human mitochondrial mTERF wraps around DNA through a left-handed superhelical tandem repeat. *Nat. Struct. Mol. Biol.* **17,** 891–893 (2010).

Chapter 5: Conclusions

5.1 Summary of Main Findings

The overarching goal of the work presented in this thesis is to develop a detailed understanding of the search mechanism and target specificity of transcription activator-like effector (TALE) proteins. We accomplished this goal by utilizing single molecule fluorescence microscopy (SMFM) in order to extract molecular-level details of TALE-DNA interactions in real time. We also combined these results with bulk biochemical techniques that allowed for assaying a large number of conditions that impact TALE binding.

In Chapter 2, we describe the ability of TALEs to diffuse one dimensionally along long DNA templates. These results present the first direct observation of facilitated search by TALE proteins along DNA. We characterize both the diffusive speed of TALEs during non-specific search and their binding lifetimes, and based on these results, we proposed a two-state search mechanism for TALE search. Individual TALE diffusion trajectories show an unexpected level of heterogeneity, highlighting possible 'search' and 'check' activities that may help TALEs to overcome the search speed/stability paradox. Distributions of TALE binding lifetimes are best described by multi-exponential decays, indicating that there are likely multiple binding modes for TALEs. We prepared TALE truncation mutants containing the TALE N-terminal region (NTR), and we observe that this construct exhibits only short, rapid 1-D diffusion in contrast to the heterogeneous diffusive behavior of full-length TALEs. Combined with our observation that an NTR-deficient TALE mutant is incapable of biding DNA, these findings highlight the key role of the NTR in nucleating initial TALE binding, as well as its role in mediating non-

specific searching events. We extended our results to include TALEs with different numbers of repeats (overall, we studies 11.5, 15.5, and 21.5 repeat TALEs) and we find that diffusion coefficients do not scale according to hydrodynamic drag models when the number of repeats was increased. We hypothesize that longer repeat domains in the CRD influence the frequency and/or duration of local sequence checking events in the context of the two-state model for TALE search.

In Chapter 3, we uncover new details regarding the conformation of TALEs during their search process and the microscopic trajectory they follow during 1-D non-specific search of DNA. We observed the ability of TALEs to diffuse under supraphysiological ionic strengths, wherein we observe 1-D diffusion coefficients larger than any previously reported for DNA-binding proteins, even those that are substantially smaller in size than TALEs[1]. We next utilized a hydrodynamic flow assay to characterize the relative association strength of non-specifically bound TALEs to DNA. Under supraphysiological ionic strengths, we were able to push and trap diffusing TALEs against the chemical linkages of our dual-tethered DNA substrates, confirming that these TALEs fully encircle DNA substrates and providing evidence that TALEs are with some frequency effectively dissociated from DNA during search despite their superhelical conformation. Under physiological ionic strength, we observe that this bias persists, which provides further evidence that TALEs remain only partially associated (via transiently forming electrostatic interactions) with DNA during their search, despite their wrapped, superhelical conformation. By modulating the size of the fluorescent probes conjugated to TALEs, we determined a probe size scaling relationship for 1-D diffusion coefficients, and we find that TALEs follow a primarily rotation-free search trajectory, in

contrast to a large number of other DNA binding proteins undergoing non-specific search[2].

In Chapter 4, we describe the effects of divalent cations on TALE binding. We demonstrate that certain divalent cations significantly enhance TALE binding specificity by simultaneously enhancing target binding while significantly attenuating non-specific binding. We studied several different TALE constructs (11.5, 15.5, and 21.5 repeat length CRDs, as well as the NTR-only variant), and our results show that longer TALEs are more susceptible to the divalent cation-mediated specificity. Divalent cations significantly attenuate NTR binding, and given the NTR's demonstrated role in nucleating initial binding events[3,4], we hypothesize that disruption of NTR binding is a contributing factor to divalent cation mediated specificity. Our results suggest that specificity is balanced by decreasing the NTR-mediated nonspecific binding affinity of TALEs while maintaining the CRD-mediated specific binding, which increases as CRD length is similarly increased. Although divalent cations are well characterized as important cofactors in many biological processes, there have been only limited reports of their contributions to increasing specificity in DNA-binding proteins outside of nucleases and restriction enzymes.

We further characterized the effects of divalent cations on TALE-DNA interactions using our SMFM assay. We directly observe a significantly decreased TALE occupancy on DNA templates in the presence of low concentrations of divalent cations. Furthermore, the divalent cations reduce the characteristic binding lifetimes of TALEs on non-specific DNA. We further studied the specific binding behavior of TALEs along DNA templates containing the target binding sites. Here, we observed TALE target

search and localization at target sites under both monovalent and mixed monovalent/divalent cation conditions. We also observed preliminary evidence of TALE target bypass, target release, and TALE-TALE collisions, and we present these findings in the context of a rotationally de-coupled search and two-state search model developed in Chapters 2 and 3. Taken together, the results of Chapter 4 provide insight as to how TALE specificity can vary depending on the design of the TALE (CRD length) and the conditions present in the intracellular environment.

5.2 Future Outlook

Future work in understanding TALE search dynamics at the single molecule level can be categorized into three broad categories: investigation of intramolecular conformational dynamics during the search process, single molecule kinetic studies on TALE target and near-target binding, and *in vivo* characterization of TALE intranuclear search behavior.

Our results to date have led us to propose a two-state search model wherein TALEs bind DNA nonspecifically, followed by 1-D diffusion in a rotationally decoupled search process. Our results support at model in which TALEs remain primarily out of phase with the DNA major or minor groove, likely adopting an extended helical conformation during the non-specific search process. In order for TALEs to locate and stably bind their target sites, we reason that TALE proteins are able to compress along their helical axis, thereby engaging in a checking mode. This conformational change is supported by recent molecular dynamics studies[5], and helps to reconcile the search speed/stability paradox for TALEs. Careful incorporation of intramolecular fluorescent donor and acceptor molecules within the TALE CRD could enable single molecule FRET

readout of this conformational change. In this way, the conformation of TALEs (either extended or compressed superhelix) could be monitored while TALEs are simultaneously tracked during their search process. The results presented in this thesis support a model in which heterogeneity in TALE search dynamics arises due to conformational changes within the TALE structure. However, in the absence of an assay with higher spatial resolution such as smFRET, we are unable to provide direct evidence of this behavior and are thus unable to characterize the frequency and duration of any 'checking' events. Furthermore, the ability to directly read out TALE conformation during search would allow for correlation of checking and searching behaviors with solution conditions (such as ionic strength or presence of divalent cations) and local DNA sequence. Key questions include: do TALEs enter the 'check' mode more often in G-rich sequences, or near partial cognate sites?

In Chapter 4, we describe the direct observation of TALE target binding under a variety of conditions. We show that TALEs are capable of locating their target sites via a 1-D facilitated search and perhaps more rarely by apparent 3-D collisions with little, if any facilitated search. The spatial resolution of our technique is limited to ~65 nm, which is well below the diffraction limit of light and permits us to observe long-range TALE search on extended DNA substrates. Nevertheless, this spatial resolution spans >200 base pairs. Thus, it is not possible to use the current assay to unambiguously determine whether TALEs remain stationary at their target sites, or whether they fluctuate between nearby target sites in the 8x binding array, perhaps sampling both target and near-target locations. To address these issues, single molecule FRET would provide the requisite sub-nanometer resolution of TALE dynamics. Instead of

incorporating the FRET donor pair entirely within the TALE repeats, placing donor and acceptor on the TALE and DNA molecules, respectively, would permit direct read out of short range TALE binding behavior. These smFRET measurements would allow for the interaction of TALEs with a large variety of DNA substrates to be evaluated. Similar to using bulk anisotropy measurements to study the effects of NTR-distal versus NTR-proximal sequence mismatches, smFRET studies would allow for systematic incorporation of target mismatches and evaluation of their effects on the kinetics of TALE binding. Our bulk fluorescence anisotropy measurements have permitted the observation of a previously unreported divalent cation effect and have enabled clear discrimination between target and random substrate binding. However, anisotropy measurements only provide readout of ensemble-averaged measurements at equilibrium. Single molecule FRET would enable separation of the factors contributing to apparent binding constants, revealing both the frequency and duration of TALE-DNA interactions, as well as the stability of TALE target binding.

This thesis has focused on a bottom-up, arguably reductionist view of TALE-DNA dynamics. Using both bulk and single molecule *in vitro* studies, we have revealed many details of the TALE search process. Such a reductionist view is necessary to fully understand how individual factors affect TALE behavior, in particular allowing for varying a single variable at a time and assessing at the minimal system for TALE search: DNA and protein. With a detailed picture of TALE search and target binding established using this approach, a logical next step is to add layers of complexity by investigating *in vivo* TALE search dynamics and target binding. This approach would allow for a top down view of TALE behavior. TALEs have been utilized in *in vivo* imaging studies,

albeit only as markers of specific chromosomal locations[6–9]. To date, there have been no reported investigations of the dynamics of individual TALEs within living cells. Recent investigation of the dynamics of individual CRISPR/Cas9 protein-RNA complexes[10], along with studies of mammalian transcription factors *in vivo*[11–15] provide a picture of the insights that can be gained from studying TALE binding *in vivo* at the single molecule level. Key areas that *in vivo* TALE imaging would address include residence times, search speed, and target binding stability within different chromatin structures.

5.3 References

1. Xiong, K. & Blainey, P. C. Molecular sled sequences are common in mammalian proteins. *Nucleic Acids Res.* gkw035 (2016). doi:10.1093/nar/gkw035

2. Blainey, P. C. *et al.* Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.* **16,** 1224–1229 (2009).

3. Schreiber, T. *et al.* Refined Requirements for Protein Regions Important for Activity of the TALE AvrBs3. *PLoS One* **10,** e0120214 (2015).

4. Gao, H., Wu, X., Chai, J. & Han, Z. Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell Res.* **22,** 1716–20 (2012).

5. Wan, H., Hu, J. ping, Li, K. shun, Tian, X. hong & Chang, S. Molecular Dynamics Simulations of DNA-Free and DNA-Bound TAL Effectors. *PLoS One* **8,** (2013).
6. Pederson, T. Repeated TALEs. *Nucleus* **5,** 28–31 (2014).

7. Yuan, K., Shermoen, A. W. & O'Farrell, P. H. Illuminating DNA replication during Drosophila development using TALE-lights. *Curr. Biol.* **24,** R144–R145 (2014).

8. Thanisch, K. *et al.* Targeting and tracing of specific DNA sequences with dTALEs in living cells. *Nucleic Acids Res.* **42,** e38–e38 (2014).

9. Miyanari, Y., Ziegler-Birling, C. & Torres-Padilla, M.-E. Live visualization of chromatin dynamics with fluorescent TALEs. *Nat. Struct. Mol. Biol.* **20,** 1321–1324 (2013).

10. Knight, S. C. *et al.* Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350,** 823–6 (2015).

11. Gebhardt, J. C. M. *et al.* Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. *Nat. Methods* **10,** 421–6 (2013).

12. Normanno, D. *et al.* Probing the target search of DNA-binding proteins in mammalian cells using TetR as model searcher. *Nat. Commun.* **6,** 7357 (2015).

13. Caccianini, L., Normanno, D., Izeddin, I. & Dahan, M. Single molecule study of non-specific binding kinetics of LacI in mammalian cells. *Faraday Discuss.* **00,** 1–8 (2015).

14. Morisaki, T., Müller, W. G., Golob, N., Mazza, D. & McNally, J. G. Single-molecule analysis of transcription factor binding at transcription sites in live cells. *Nat. Commun.* **5,** 4456 (2014).

15. Mazza, D., Abernathy, A., Golob, N., Morisaki, T. & McNally, J. G. A benchmark for chromatin binding measurements in live cells. *Nucleic Acids Res.* **40,** 1–13 (2012).