STUDY AND ENGINEERING OF NUCLEIC ACID-BINDING REPEAT PROTEINS

BY

ZHANAR ABIL

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biochemistry
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Huimin Zhao, Chair
Associate Professor Charles M. Schroeder
Associate Professor Raven H. Huang
Associate Professor Lin-Feng Chen

Abstract

From DNA replication and repair to transcription and translation, all aspects of molecular biology are governed by interactions between proteins and nucleic acids. Some of the nucleic acid-binding proteins have evolved repetitive structures and a modular sequence recognition mode. In particular, DNA-binding transcription activator-like effectors (TALE) and RNA-binding Pumilio/fem-3 binding factor homology (PUF) proteins were shown to bind their target sequences in a single repeat/base fashion. This modular recognition nature allowed researchers to rationally design these proteins for novel specificities and utilize them in diverse biotechnological applications such as genome engineering, gene expression regulation, and imaging of specific nucleic acids.

The first part of my dissertation was dedicated to the study and engineering of TALE. Using the published specificity code, a pair of TALE nucleases was engineered to site-specifically cut the Cystic Fibrosis Transmembrane Regulator gene (*CFTR)* carrying the Δ508F mutation responsible for the most cases of cystic fibrosis. Coexpression of these proteins in human cell lines allowed homologous recombination-mediated repair of the episomal and chromosomal reporter gene interrupted with the sequence fragment from *CFTR*. In the next project, the binding dynamics of TALE on DNA substrates was investigated using single molecule fluorescence microscopy. For the first time, the binding dynamics and 1-D diffusion of TALE proteins along DNA were directly visualized. The data strongly suggest that TALE searches its target using the facilitated diffusion mechanism, with a combination of 1-D and 3-D diffusion before it reaches its target sequence. This study can contribute to improved rational design of these proteins for biomedical applications.

The aim of the second part of my dissertation was to engineer new PUF-based biosynthetic tools. First, the Golden Gate cloning method was implemented for an efficient one-step assembly of designer PUF proteins. To this end, a repeat module library that is potentially capable of generating a PUF domain with any desired specificity was created. The assembled novel PUF variants exhibited high *in vitro* binding efficiencies to cognate RNA sequences, corroborating the applicability of the modular assembly approach for PUF engineering. Next, the PUF domain was fused to a post-transcriptional regulator, which allowed for a sequence-specific reporter and endogenous gene repression in human cell lines. This work was a demonstration of the efficacy of a synthetic PUF-based gene expression regulator.

In the last project, a PUF-based system for intracellular directional transport of mRNA was developed in mammalian cells. The biosynthetic device consists of a "motor", which provides a directional movement, and the PUF protein that dimerizes with the "motor" in a ligand-dependent manner. The system allowed RNA-sequence and motor-specific transport and colocalization of PUF and its RNA cargo. Currently, this PUF-motor system is being implemented for the transport as well as axonal and dendritic local protein translation of reporter and endogenous genes in rat neurons. This prototypical synthetic device should allow easy and controlled intracellular mRNA transport regulation in eukaryotes for applications in basic science and therapeutics.

Acknowledgements

I would like to thank my dear Alma Mater for honoring me with this life-changing opportunity and letting me meet and learn from some of the greatest minds of today. I will be always grateful to her for giving me this precious chance to further pursue my dreams.

I am very grateful to all the past and current members of my PhD committee Dr. Maria Spies, Dr. Daniel W. Pack, Dr. Raven H. Huang, Dr. Lin-Feng Chen, Dr. Charles M. Schroeder, and Dr. Huimin Zhao for their kind support, helpful advice, and challenging inquiry.

I would like to give special thanks to my collaborators Dr. Charles M. Schroeder, Luke Cuculis, Dr. Laura F. Gumy and Dr. Casper Hoogenraad, who were very generous in sharing their valuable expertise, brought new perspective, and made the most challenging projects possible.

My sincere acknowledgements also go to Dr. Barbara Pilas, Ben Montez, Dr. Mayandi Sivaguru, Dr. Xudong Guan, Dr. Chen Zhang, Dr. Peter Yau, Dr. Brian S. Imai, and Dr. Sandra McMasters for training, advising, and helping me with so many different technologies required for the completion of this work, and to Shawna M. Smith and Jeff Goldberg for their cordial and skillful coordination.

I am greatly indebted to Dr. Michael J. McLachlan and Dr. Zengyi Shao for their patient mentorship in the beginning of my PhD work. Equally, my gratitude goes to my labmates Dr. Jing Liang and Dr. Han Xiao for helping me learn even further, Ryan E. Cobb for his help with writing, and Stella Xinzi Wu for her help with several of my projects.

I am grateful to my friends Dr. Dawn T. Eriksen, Rachel J. Waldemer, Kori L. Dunn, Hui-Chia Yu, Koh Eun Narm, and Dr. Guiomar S. Carrón who were always there to help and encourage.

I would like to give special thanks to my fiancé Dr. Carl A. Denard for his enormous help, advice, and inspiration, for sharing the happiest moments along this path, for giving me courage in those of the worst, and for letting me grow as a person. This work wouldn't be possible without his continuous support.

Finally, I would like to thank my family – my brother and sister-in-law for their continuous inspiration and support. My special gratitude is to my father, whose efforts made it possible to continue my education, as well as my mother, who, at my youngest age, ignited my passion to the world of scientific inquiry and inspired me to be where I am today.

Table of Contents

Chapter 1. Introduction[1]

1.1 Synthetic Biology and Repeat Proteins

The field of synthetic biology aims to contribute to a variety of health and environmental challenges including the production of biofuels and biomaterials, drug discovery, drug production, development of new vaccines, as well as combatting infectious diseases and cancer. The main principle of synthetic biology is to use well-characterized functional modules to build biological systems with novel functions. The functional modules that it develops and utilizes are usually protein-coding or regulatory DNA parts, which are cloned from various organisms or synthesized, and assembled together into genetic circuits.

Engineering devices for synthetic biology commonly requires the use of modules for regulation of gene expression on DNA or RNA levels. Emulating nature, these modules are often developed from nucleic acid-binding proteins linked to various functional domains following the modular design principle, where each domain in a polypeptide chain functions independently from the rest of the protein. Highly sophisticated synthetic biology endeavors would require equally complex gene circuits, which in turn would require multiple orthogonal gene expression regulators. There are a limited number of orthogonal nucleic acid-binding proteins in nature, however, and thus scaffold DNA- and RNA-binding proteins (DBPs and

[1] Portions of this chapter were reprinted (adapted) with permission from (Abil *et al. Mol Pharm* 2015, 12(2), pp322-331) Copyright (2015) American Chemical Society and (Sun *et al. Biotechnol J* 2012 7(9), pp1074-1087) Copyright (2012) John Wiley and Sons.

RBPs, respectfully) with easily reprogrammable specificity are required. Using the same modular design fashion, precision-targeted DNA-binding endonucleases can also be constructed for genome modification purposes, as will be discussed later. A single precise genome modification in an organism would also require a highly specific DNA-binding domain (DBD), easily reprogrammable for a sequence of interest for each new genetic locus. Engineering artificial proteins with desired nucleic acid sequence specificity is therefore one of the important challenges of synthetic biology.

It was shown that transcription factors with new DNA-binding specificities can be selected[1] from a library of zinc fingers[2]. Subsequently, numerous new transcription factors and endonucleases were successfully created using zinc fingers (reviewed in 3 and 4). However, the DNA-binding specificity in zinc finger proteins is not fully reprogrammable and certain selection strategies have to be implemented for isolation of a zinc finger domain with desired specificity[5-6]. Thus, engineering of proteins for nucleic acid specificity has not been straightforward until recently, with the implementation of proteins with repetitive structure.

Proteins with repetitive structure, or repeat proteins, are widespread in nature, and include ankyrin, armadillo, WD, HEAT, leucine-rich repeat proteins, tetratricopeptide repeat TPR proteins, pentatricopeptide repeat (PPR) proteins, Pumilio and fem3 mRNA-binding factor (PUF) homology proteins, and transcription activator-like effector (TALE) proteins[7-13]. It is generally believed that repeat proteins evolve through coding sequence duplication within genes, and consecutive divergence of each repeat sequence from the neighboring repeats. The structure of

such proteins is dominated by short-range interactions that usually stack the tandem arrays of repeats in an elongated non-globular structure with a joined hydrophobic core[8, 14]. The repeat domain is usually flanked and stabilized by special repeats, or 'pseudorepeats', that shield the hydrophobic core from the solvent[8, 14]. This architecture, in principle, allows for addition, exchange, and removal of repeats in an out of the repeat domain without compromising the overall structure, but varying the surface for ligand interaction. The extended architecture of repeat proteins can thus allow for modular engineering of these proteins for specific binding to other macromolecules, such as proteins, as was shown with ankyrin repeats[15-16] and TPR proteins[17].

A striking revelation was that some of the repeat proteins, such as PUF and TALE proteins, bind nucleic acids in a modular fashion, where one repeat interacts with one base, with certain amino acid positions in each repeat dictating the specificity[18-21]. This modular mode of binding to DNA prompted to develop TALE proteins as a valuable tool for engineering DNA-binding proteins with custom specificity[22-23]. Similarly, PUF domains have been successfully reprogrammed for altered RNA specificity[18-19], an accomplishment that has not been previously achieved with any other RNA-binding protein. Although other repetitive nucleic acid-binding proteins, such as PPR proteins also hold promise for engineering of custom RNA-binding proteins[24], in this dissertation, I focused on the PUF protein because PUF proteins have been studied more extensively. Cas9 proteins that site-specifically recognize DNA via RNA-hybridization also hold promise in engineering custom DNA-binding proteins[25]; however, it was beyond the scope of this work.

1.2 TALEs as Designer DNA-Binding Proteins

TALE proteins were first discovered in a plant bacterial pathogen of the genus *Xanthomonas*, which uses them as transcription activators of host plant genes[26]. The pathogen injects the proteins into the plant cells via a type III secretion system. These proteins are translocated into the plant cell nucleus and used to recognize target promoters and activate the expression of genes that are important for pathogen virulence.

1.2.1 TALE DNA-Recognition Code, Structure, and Preparation

The number of repeats in natural TALEs varies between 1.5 and 33.5, with 15.5–19.5 being the most common[26-27]. However, it was shown that at least 6.5 repeats are required for gene induction in promoter activation studies[21]. While the usual repeat length is 34 amino acids, the last repeat is typically shorter, with only 20 amino acids, and is therefore referred to as a 'half-repeat'[26]. The DNA recognition code of TALE was deciphered in two independent studies[20-21]. While most of the amino acids are highly conserved across repeats, two hypervariable amino acids at positions 12 and 13, often referred to as the repeat-variable diresidue (RVD), confer the specificity of each repeat to the corresponding nucleotide in the target DNA sequence (Figure 1.1). Thus, RVDs NI and HD show strong preference for adenine and cytosine, respectively, while both NG and HG favor recognition of thymine. Some other RVDs are less specific: NS has the least definitive nucleotide preference, while NN recognizes both guanine and adenine. Thus, there is no RVD with specific recognition of guanine, which must be taken into account when designing artificial TALEs. The N-terminal repeat of TALEs binds to the 5′ end of the target DNA

sequence, while each consecutive repeat binds a single base downstream of the same DNA strand.

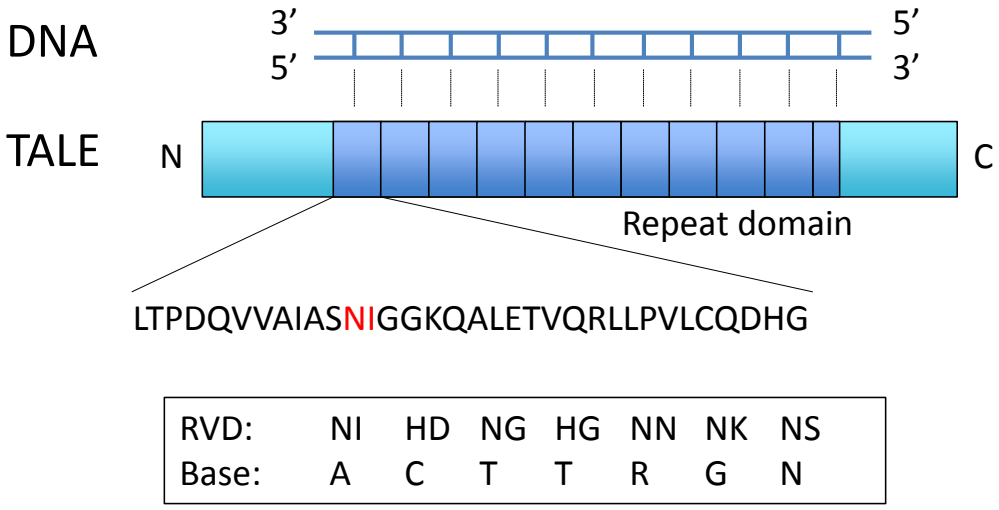


Figure 1.1 Schematic of TALE binding to dsDNA. Recognition code is shown in a box. R represents A or G. N represents any of the four nucleobases. Amino acid residues in red represent RVD at positions 12 and 13.

The crystal structures of an 11.5-repeat artificially engineered TALE dHax3 in both DNA-free and DNA-bound conformations[28] and a 23.5-repeat natural TALE PthXo1 from the rice pathogen *Xanthomonas oryzae* bound to its DNA target[29] improved our understanding of TALE architecture and its DNA recognition. These structures revealed two α-helices that span residues 3–11 and 14(15)–33 in each repeat and a flexible loop containing the RVD between the two helices. The RVD is positioned in such a way that it interacts with the major groove of target DNA. The resulting right-handed superhelical structure of a TALE tracks along the major groove of the B-form DNA duplex. The two reports agree that residue at position 12 of each repeat does not directly contact the DNA, but indirectly contributes to the stability of the binding conformation and allows residue 13 to specifically interact

with a corresponding nucleotide. Interestingly, comparison of TALE structures in DNA-bound and DNA-free forms revealed a significant conformational change along the axis of the TALE superhelix, which might be important for the function of TALE[28].

The modular recognition mode of TALE allows for easy design of custom DBD. However, the highly repetitive nature of the modules makes it difficult to mutagenize the protein using the traditional methods of recombinant DNA technology. The development of the Golden Gate cloning method[30] abolished the need for costly artificial DNA synthesis and allowed for an easy and efficient in-house assembly of TALEs[31-39]. Using this method, up to ten repeats can be efficiently assembled in a single digestion-ligation reaction. The technique is based on the property of type IIS restriction endonucleases to cleave outside their recognition sequences, which results in unique non-palindromic 4-bp overhangs. These overhangs can be re-ligated in an ordered fashion, with a simultaneous removal of the recognition sites.

In another approach, Reyon and coworkers[40] developed the fast ligation-based automatable solid-phase high-throughput (FLASH) system to synthesize TALEs in large scale, which significantly decreases the TALE synthesis cost (less than $200 per pair). An alternative high-throughput TALE synthesis approach was proposed in a method called fairyTALE by Liang *et al*., wherein a liquid-phase Golden Gate-based approach allowed efficient synthesis of TALEs at material cost of $10 per TALE pair[41]. With the combined effort from several laboratories, TALE assembly toolkits are now available through Addgene and have been made accessible to the public.

Moreover, commercial services for custom TALE assembly and activity validation are also conveniently available from Cellectis (Romainville, France), PNA Bio (Thousand Oaks, CA, USA) and Life Technologies (Grand Island, NY, USA).

1.2.2 TALE Nuclease (TALEN) Applications

Using the modular design principle of DNA-binding proteins, it was shown that the activation domain of TALE proteins could be replaced with a non-specific nuclease domain FokI. Such a chimeric protein can function as a highly specific endonuclease, which can be utilized for genome editing purposes in basic and applied science[31, 34, 42-46]. The principle of using endonucleases for gene editing lies in the ability of most organisms to activate native DNA repair machineries to repair the site-specifically introduced double strand break (DSB). Major repair pathways are non-homologous end joining (NHEJ)[47] and homologous recombination (HR)[48]. NHEJ is an error-prone process, which often leads to micro-deletions or micro-insertions that often result in frame shift mutagenesis and a consequent gene knockout. HR is less error-prone, and in native conditions, is achieved using a sister chromatid as a template for the repair of the damaged chromosome.

In traditional gene targeting, DSB-independent HR was utilized to mutagenize a gene of interest by providing an exogenous homology donor. The frequency of HR in such conditions was very low (10^{-4} – 10^{-7}). However, it was later discovered that the introduction of DSBs via site-specific endonucleases increases the efficiency of HR by several orders of magnitude[49-50]. This site-specific endonuclease-mediated gene targeting approach is referred to as gene editing.

Since FokI functions as a dimer[51], TALENs are designed as heterodimers,

where two TALE-FokI fusions are designed to bind opposite strands of target DNA, in a head-to-head orientation. We and other groups performed optimization of the TALEN architecture to achieve the best efficiency in gene editing. It was shown that TALE truncations that retain 20–60 amino acid residues of the original C-terminal residues and ~150 residues at the N terminal region had a higher TALEN activity compared to the full-length TALE sequences[42, 46, 52]. These reports also agree that longer C-terminal configurations (~60 compared to ~20 residues) have a higher preference for longer spacer lengths between the two protein binding sites (12–30 bp compared to 10–16 bp)[46, 52].

From the natural TALE-target promoter pairs listed by Moscou and Bogdanove[20], Cermak and co-workers[31] developed a program that calculated frequencies of natural TALE binding site preferences. Based on these calculations, they listed several guidelines for choosing an optimal synthetic TALEN target site. Thus, besides the favorable thymine at position –1[20-21], they suggest that thymine at position 1 (the first position at the 5′ end of the target sequence) is disfavored, adenine at position 2 is disfavored, and guanine at the last and penultimate positions are disfavored, and a thymine at the last position (when the last half repeat contains the RVD NG) is highly favored[31].

The first proof-of-concept for using custom-designed TALENs for genome modification in a trans-kingdom host was shown using a LacZ reporter in *Saccharomyces cerevisiae*[53-54]. Soon after, in cultured mammalian cells, successful targeted gene disruption (up to 21% efficiency) as well as editing (up to 16% efficiency) of endogenous human neurotrophin 3 (*NTF3*) and chemokine (C-C motif)

receptor 5 (*CCR5)* genes were reported[46]. In another study, the hypoxanthine phosphoribosyltransferase 1 gene (*HPRT1)* was knocked out in a mammalian cell line[31].

Compared to immortalized tumor cell lines, human pluripotent stem cells are relatively resistant to gene targeting. This limitation can be addressed by using the TALEN technology, as was shown with TALENs engineered for recognition of *PPP1R12C*, *OCT4* and *PITX3* genes in human pluripotent cells[43]. With the help of a gene trap or an autonomous selection, they reported targeting in >50% of the selected colonies for *OCT4*, ~50% for *PPP1R12C*, and 1–10% for *PITX3*.

The first genome editing in an animal was shown by Tesson and coworkers, who knocked out the immunoglobulin M (*IgM)* gene by injecting TALEN-coding nucleic acids into one-cell rat embryos[55]. When the nucleases were introduced by transfection with coding DNA, the frequency of mutated animals was 9%, whereas when nucleases were introduced to cells via transfection of coding mRNA, the frequency of modified animals was increased to 59%.

The wide range of variations in the modification efficiencies reported to date might be due to the varying affinities of different TALEs to target DNA, differences in amenability of the cells and tissues for the uptake of foreign genetic material, and the cell cycle status of the cells. Additionally, chromosomal context and epigenetic modifications determine the chromatin accessibility of tailored DNA endonucleases, which may play a major role in the genomic modification efficiency. It is therefore important to investigate the influence of each of these factors on genome modification rates and develop strategies to efficiently manipulate them. Despite big

variations, the efficiency of targeted genome engineering in most studies is low (<1%); therefore, a powerful and seamless selection strategy would be valuable to enrich the targeted events.

1.3 Designer RNA-Binding Proteins

PUF family proteins are named after the founding members Pumilio from *Drosophila melanogaster* [56] and fem3-binding factor (FBF) from *Caenorhabditis elegans*[57]. Since then, PUF homology proteins have been found in all eukaryotic organisms[58-59]. PUF domains usually bind to the 3′ UTR of target genes, and regulate mRNA stability, translation, and localization (reviewed in references [59-60]), pre-rRNA processing[61], and even recognition of viral infections[62-63].

1.3.1 PUF Domain: Reprogrammable RNA-Binding Protein

PUF family proteins are identified by the presence of 8 imperfectly repeated 36 amino acid motifs (PUF repeats), which, together with flanking conserved sequences, form a sequence-specific RNA-binding domain (RBD), the PUF domain[64-65]. The first crystal structures of PUF domains revealed that the eight PUF repeats are very similar structurally, and each of them comprises a triangle-shaped structural unit that contains three helical segments[66-67]. These repeated units pack together to form a crescent-shaped right-handed superhelix with a continuous hydrophobic core[66-67].

The crystal structure of a PUF domain from human Pumilio 1 (HsPUM1) protein in complex with RNA revealed that RNA is bound to the concave surface of the protein; moreover, each repeat interacts with a single RNA base, suggesting that RNA recognition of the PUF domain is highly modular[18]. The N-terminal repeat of

the PUF domain interacts with the 3′ end of the 8-base target mRNA, with each subsequent repeat interacting with the adjacent RNA base immediately upstream; thus the PUF domain binds RNA in what can be described as 'antiparallel' configuration[18]. The structure revealed that each repeat establishes a sequence-specific interaction with a Watson-Crick edge of the RNA base via amino acid side chains at positions 12 and 16 in each repeat (Figure 1.2). The amino acid side chains at position 13 in each repeat form stacking interactions with adjacent RNA bases[18]. RNA-recognition code was proposed based on this co-crystal structure, where NYxxQ (residue positions 12-16, x-any residue) recognizes uracil, CRxxQ recognizes adenine, and SNxxE recognizes guanine. No code for cytosine was apparent from this structure, although repeat 4 could accept any base.

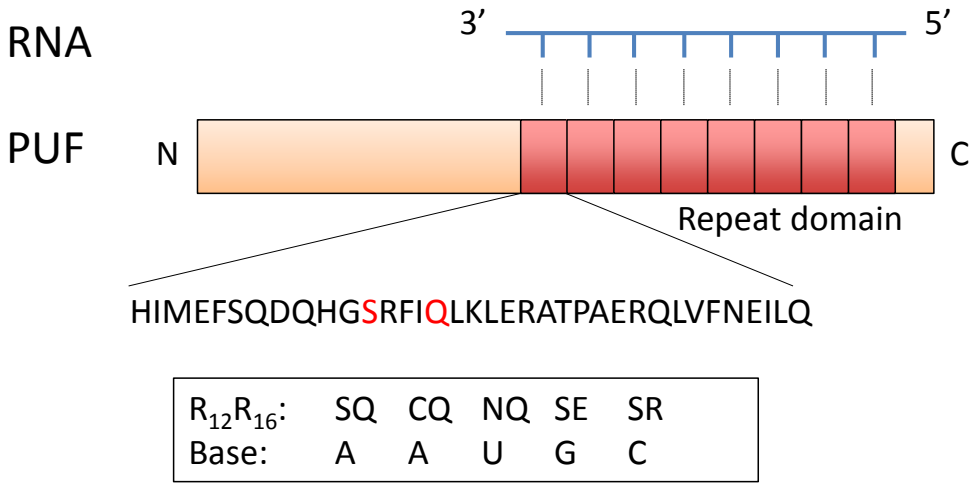


Figure 1.2 Schematic of PUF binding to ssRNA. Recognition code is shown in a box. Amino acid residues in red, as well as $R_{12}R_{16}$ represent key amino acids at positions 12 and 16.

Such elegant RNA recognition pattern suggested that PUF domains with designed specificities could be created via site-directed mutagenesis. Indeed, mutagenesis of only 2-3 residues in a repeat was necessary to predictably alter PUF

domain specificity in HsPUM1, as confirmed by *in vitro* binding assays[18-19, 68]. Recently, this PUF specificity code was expanded using data gathered from a high throughput sequencing approach. These data allowed determination of the specificity conferred by various combinations of the three amino acid residues introduced into the seventh repeat of FBF-2[69]. This elegant approach, termed SEQRS, combined *in vitro* selection, high-throughput sequencing of RNA and sequence specificity landscapes. Multiple highly specific combinations were discovered for the recognition of uracil, adenine, and guanine, some of which were *de novo* designed amino acid combinations. Thus, the *de novo* designed CFxxQ and CYxxE amino acid combinations were more specific for adenine than any other natural combination[69].

Identification of cytosine-recognition code of PUF repeats expanded the RNA-binding specificity of PUF domains to recognize any RNA target sequence[70-71]. In order to select for a PUF repeat with specificity for cytosine, yeast three-hybrid system was used, where the interaction between the PUF domain and target RNA was linked to a life-death selection in *S. cerevisiae*. From a pool of PUF domain variants with randomized amino acid positions 12 and 16 in repeat 6, PUF repeats with the SYxxR amino acid combination were selected[70]. In a similar study, arginine at position 16 was likewise selected; while in the position 12, other small or nucleophilic side chains were also selected besides serine[71]. This code was functional in other PUF repeats[70-71], and PUF domains with multiple cytosine-recognizing repeats were shown to be active[70].

1.3.2 PUF Domain in Applications

Most RBPs have modular configuration: RBDs are physically separate and function independently from effector domains[72]. Emulating nature, artificial RBPs can be created in a similar fashion, by combining effector domains of choice to RBD's with desired specificity[73-77]. These designer RBPs can be implemented in the regulation of a broad range of biological processes that involve RNA metabolism.

The first application of a PUF domain-based engineered RBP was live-cell tracking of mRNA. Visualization of RNA in live cells have previously been possible by tagging the RNA of interest with multiple copies of recognition elements and the use of RNA-binding proteins fused to fluorescent proteins like GFP[78]. Although successful in detecting and tracking RNA in live ells, this system requires prior tagging of RNA of interest, which may alter RNA abundance and localization. Application of reprogrammable RNA-binding proteins such as PUF domains for this purpose abolishes the need of tagging the RNA of interest. To this end, split-protein systems were implemented, where two PUF domains were used to reconstitute enhanced green fluorescent protein (EGFP) or Venus from two nonfunctional parts that were brought to proximity on target RNA[79]. Such split protein system reduces background fluorescence and improves the signal-to-noise ratio compared with full-length GFP fusions, where it is difficult to distinguish RNA-bound protein GFP from RNA-free protein. This approach allowed to track mitochondrial RNA[79] or single molecules of β-actin mRNA[80] in cultured mammalian cells. In a similar strategy, split mCitrine-PUF system was used to track tobacco mosaic virus RNA in plant cells[81].

The ability to control gene expression is necessary for both fundamental research as well as in applications such as cell fate manipulation and metabolic engineering. Gene expression is regulated at multiple RNA processing steps, including capping, polyadenylation, splicing, degradation, editing, and translocation. By fusing corresponding functional domains to RBDs such as PUF, one can envision development of tools capable of sequence-specific manipulation of RNA almost at any processing step. One such tool has been developed to inhibit translation via regulation of the poly(A) tail of mRNA. To inhibit translation of microinjected reporter mRNA in *Xenopus* oocytes, FBF-2, a PUF domain found in *C. elegans,* was fused to *Xenopus* CAF1 protein, which has deadenylation activity that shortens poly(A) tails as well as an intrinsic translational repression activity[82]. The system induced deadenylation and decreased translation of reporter luciferase mRNA. In another study, a PUF domain alone, targeted to the 5′ untranslated region (UTR) of an open reading frame, was used as a steric block in the translation initiation pathway to inhibit the translation of the downstream reporter luciferase gene[83].

The use of a PUF scaffold for sequence-specific enhancement of gene expression in living systems was first demonstrated in *Xenopus* oocytes. FBF-2 was fused with GLD2, a cytoplasmic poly(A) polymerase, and the chimeric protein was shown to direct polyadenylation of reporter RNA, enhance translation of microinjected luciferase mRNA, and direct polyadenylation of endogenous ribosomal protein L1 mRNA[82]. In another elegant study, the PUF domain, which was targeted to the 5′ UTR of an open reading frame, was fused to the eukaryotic translation initiation factor 4E (eIF4E) to enhance translation of reporter luciferase

mRNA[83]. Furthermore, light-inducible heterodimerization of PUF and eIF4E through light-sensitive protein partners was demonstrated, thus allowing light-inducible translation activation[83]. Alternatively, the PUF domain of FBF-2 protein that was mutagenized for specific binding to a new target sequence was fused with a segment of *S. cerevisiae* poly(A)-binding protein, which is known to stimulate translation. This chimera protein, referred to as 'neo-activator', increased endogenous cyclin B1 abundance in a cancer cell line, and rendered the cell line hypersensitive to chemotherapeutic drugs[69].

Another strategy by which gene expression is regulated in nature is regulation of mRNA abundance. In many organisms, this strategy is achieved through the RNA interference (RNAi) pathway, where a multi-component complex cuts and degrades target RNA[84]. An alternative synthetic system for RNA cleavage termed artificial site-specific RNA endonuclease (ASRE) was designed using a PUF domain as an RNA-recognition scaffold fused to a non-specific endonuclease domain PilT N-terminus[85]. ASRE targeting the endogenous *LacZ* transcript in *Escherichia coli* and a mitochondrial gene in human cultured cells were designed and used to sequence-specifically degrade target mRNA and reduce corresponding protein levels.

Alternative splicing is an important step in eukaryotic mRNA processing that allows for expression of multiple isoforms of proteins with distinct functions. The first attempt at engineering splicing factors with designed sequence specificities and activities was reported using a PUF domain as an RNA-binding scaffold[86]. These engineered splicing factors (ESFs) were constructed from a wild type or modified PUF domain of HsPUM1 fused to a glycine-rich domain of hnRNP A1 (Gly-PUF) or

the arginine-serine-rich domain of ASF/SF2 (RS-PUF). Consistent with the reported activities of the tethered domains, Gly-PUF type ESF repressed inclusion of the cassette exon containing the cognate target sequence of PUF, while RS-PUF type ESF promoted inclusion of the same cassette exon in reporter constructs[70, 86]. Additionally, these ESFs could also regulate alternative use of splice sites. For example, RS-PUF increased the use of an upstream 3′ splice site, whereas Gly-PUF promoted the use of a downstream 5′ splice site. The designer ESF was shown to be able to shift splicing of endogenous pre-mRNA of anti-cancer target Bcl-x towards the short Bcl-xS isoform, thereby promoting apoptosis[86]; as well as promote splicing towards the anti-angiogenic isoform b of endogenous VEGFA gene[70] in cultured cancer cells.

1.3.3 Potential use of PPR Proteins in RBP Engineering

PPR are a huge class of proteins found primarily in mitochondria and chloroplasts of land plants, where they participate in RNA cleavage, splicing, degradation, editing and translation[87-88]. The RNA-binding motif of PPR consists of 2-30 degenerate repeats that are approximately 35-amino acid-long[89] and are organized in a helix-turn-helix structure[90-91]. Like PUF proteins, PPR proteins bind RNA in a 1:1 correspondence between repeats and bases, and recognize single stranded RNA in a modular fashion[92-95].

Computational, biochemical, and structural analyses agree that residue 5 (residue delimitation according to reference [92]) in each repeat contributes in distinguishing between purines and pyrimidines[93-95]. Residue 35, which is located close to the residue 5 in 3-dimensional space, allows the discrimination of adenine

from guanine, and cytosine from uracil. *In vitro* binding assays confirmed that altering these two amino acids could change the RNA sequence specificity according to the code[93]. Another amino acid, which is located at position 2, is sandwiched between two adjacent bases in a stacking interaction[92, 96], and thus is analogous to the residue at position 13 in PUF proteins. Unlike the PUF domain, the PPR protein binds RNA in a parallel configuration, i.e. the N-terminal repeat of PPR binds to the 5′-end of the target RNA sequence, and the following repeats bind a stretch of consecutive bases downstream.

Due to only a recent confirmation of the RNA-recognition code of PPR via a co-crystal structure[92], PPR proteins have not yet been used as widely as PUF proteins in reprogramming specificity to RNA or engineering of designer RBP. However, unlike PUF proteins, PPR proteins differ widely in their recognized RNA sequences, which may be an indication of a greater plasticity and amenability in engineering specificity.

1.4 Scope and Goals of This Dissertation

The first part of this dissertation is dedicated to the study and engineering of TALE proteins. In Chapter 2, the engineering of TALEN for targeting Cystic Fibrosis Transmembrane Regulator gene (*CFTR)* is discussed. The Δ508F mutation in exon 10 of *CFTR* is responsible for the most cases of cystic fibrosis[97], and was chosen as the target site for the engineered TALEN. Expression of the engineered TALEN in human cell lines allowed HR-mediated repair of the episomally and chromosomally located reporter green fluorescent protein gene interrupted with the sequence fragment from *CFTR*.

In Chapter 3, we described the binding dynamics of TALE on DNA substrates measured by single molecule fluorescence microscopy. For the first time, the binding dynamics and one-dimensional diffusion of TALE proteins along DNA were directly visualized. The data strongly suggest that the TALE protein searches its target using the facilitated diffusion mechanism, as was reported for many other transcription factors. This study can contribute to improved rational design of TALEs for biomedical applications.

The aim of the second part of my dissertation was to engineer new PUF-based biosynthetic tools. Chapter 4 discusses implementation of the Golden Gate cloning method for an efficient one-step assembly of designer PUF proteins. To this end, a repeat module library that is potentially capable of generating a PUF domain with any desired specificity was created. The assembled novel PUF variants exhibited high *in vitro* binding efficiencies to cognate RNA sequences, corroborating the applicability of the modular assembly approach for PUF engineering. In Chapter 5, we described the development of a designer translation inhibitor. The PUF domain was fused to human TTP, which allowed for a sequence-specific reporter and endogenous gene repression in human cell lines. This work was a demonstration of the efficacy of a synthetic PUF-based gene expression regulator.

In Chapter 6, a PUF-based system for intracellular directional transport of mRNA was described. The synthetic biology device consists of a “motor”, which provides a directional movement, and the PUF protein that dimerizes with the “motor” in a ligand-dependent manner. The system allowed RNA-sequence and motor-specific transport and colocalization of PUF and its RNA cargo. The PUF-

motor system was implemented for the reporter and endogenous mRNA in transport in rat hippocampal neurons. This prototypical synthetic device should allow easy and controlled intracellular mRNA transport regulation in eukaryotes for applications in basic science and therapeutics.

1.5 References

(1) Pomerantz, J. L.; Sharp, P. A.; Pabo, C. O., Structure-based design of transcription factors. *Science* **1995,** *267*, 93-6.
(2) Rebar, E. J.; Pabo, C. O., Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science* **1994,** *263*, 671-3.
(3) Klug, A., The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annual review of biochemistry* **2010,** *79*, 213-31.
(4) Palpant, N. J.; Dudzinski, D., Zinc finger nucleases: looking toward translation. *Gene therapy* **2013,** *20*, 121-7.
(5) Greisman, H. A.; Pabo, C. O., A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* **1997,** *275*, 657-61.
(6) Sepp, A.; Choo, Y., Cell-free selection of zinc finger DNA-binding proteins using in vitro compartmentalization. *Journal of molecular biology* **2005,** *354*, 212-9.
(7) Main, E. R.; Jackson, S. E.; Regan, L., The folding and design of repeat proteins: reaching a consensus. *Current opinion in structural biology* **2003,** *13*, 482-9.
(8) Main, E. R.; Lowe, A. R.; Mochrie, S. G.; Jackson, S. E.; Regan, L., A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Current opinion in structural biology* **2005,** *15*, 464-71.
(9) Grove, T. Z.; Cortajarena, A. L.; Regan, L., Ligand binding by repeat proteins: natural and designed. *Current opinion in structural biology* **2008,** *18*, 507-15.
(10) Tewari, R.; Bailes, E.; Bunting, K. A.; Coates, J. C., Armadillo-repeat protein functions: questions for little creatures. *Trends in cell biology* **2010,** *20*, 470-81.
(11) Boersma, Y. L.; Pluckthun, A., DARPins and other repeat protein scaffolds: advances in engineering and applications. *Current opinion in biotechnology* **2011,** *22*, 849-57.
(12) Villamil, M. A.; Liang, Q.; Zhuang, Z., The WD40-repeat protein-containing deubiquitinase complex: catalysis, regulation, and potential for therapeutic intervention. *Cell biochemistry and biophysics* **2013,** *67*, 111-26.
(13) Filipovska, A.; Rackham, O., Modular recognition of nucleic acids by PUF, TALE and PPR proteins. *Molecular bioSystems* **2012,** *8*, 699-708.
(14) Forrer, P.; Binz, H. K.; Stumpp, M. T.; Pluckthun, A., Consensus design of repeat proteins. *Chembiochem : a European journal of chemical biology* **2004,** *5*, 183-9.
(15) Binz, H. K.; Amstutz, P.; Kohl, A.; Stumpp, M. T.; Briand, C.; Forrer, P.; Grutter, M. G.; Pluckthun, A., High-affinity binders selected from designed ankyrin repeat protein libraries. *Nature biotechnology* **2004,** *22*, 575-82.

(16) Stefan, N.; Martin-Killias, P.; Wyss-Stoeckle, S.; Honegger, A.; Zangemeister-Wittke, U.; Pluckthun, A., DARPins recognizing the tumor-associated antigen EpCAM selected by phage and ribosome display and engineered for multivalency. *Journal of molecular biology* **2011,** *413*, 826-43.
(17) Cortajarena, A. L.; Kajander, T.; Pan, W.; Cocco, M. J.; Regan, L., Protein design to understand peptide ligand recognition by tetratricopeptide repeat proteins. *Protein engineering, design & selection : Protein engineering, design & selection* **2004,** *17*, 399-409.
(18) Wang, X.; McLachlan, J.; Zamore, P. D.; Hall, T. M., Modular recognition of RNA by a human pumilio-homology domain. *Cell* **2002,** *110*, 501-12.
(19) Cheong, C. G.; Hall, T. M., Engineering RNA sequence specificity of Pumilio repeats. *Proceedings of the national academy of sciences of the United States of America* **2006,** *103*, 13635-9.
(20) Moscou, M. J.; Bogdanove, A. J., A simple cipher governs DNA recognition by TAL effectors. *Science* **2009,** *326*, 1501.
(21) Boch, J.; Scholze, H.; Schornack, S.; Landgraf, A.; Hahn, S.; Kay, S.; Lahaye, T.; Nickstadt, A.; Bonas, U., Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **2009,** *326*, 1509-12.
(22) Bogdanove, A. J.; Voytas, D. F., TAL effectors: customizable proteins for DNA targeting. *Science* **2011,** *333*, 1843-6.
(23) Mussolino, C.; Cathomen, T., TALE nucleases: tailored genome engineering made easy. *Current opinion in biotechnology* **2012**.
(24) Shen, C.; Wang, X.; Liu, Y.; Li, Q.; Yang, Z.; Yan, N.; Zou, T.; Yin, P., Specific RNA recognition by designer pentatricopeptide repeat protein. *Molecular plant* **2015**.
(25) Doudna, J. A.; Charpentier, E., Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **2014,** *346*, DOI: 10.1126/science.1258096.
(26) Boch, J.; Bonas, U., Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annual review of phytopathology* **2010,** *48*, 419-36.
(27) Scholze, H.; Boch, J., TAL effectors are remote controls for gene activation. *Current opinion in microbiology* **2011,** *14*, 47-53.
(28) Deng, D.; Yan, C.; Pan, X.; Mahfouz, M.; Wang, J.; Zhu, J. K.; Shi, Y.; Yan, N., Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **2012,** *335*, 720-3.
(29) Mak, A. N.; Bradley, P.; Cernadas, R. A.; Bogdanove, A. J.; Stoddard, B. L., The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **2012,** *335*, 716-9.
(30) Engler, C.; Kandzia, R.; Marillonnet, S., A one pot, one step, precision cloning method with high throughput capability. *PloS one* **2008,** *3*, e3647.
(31) Cermak, T.; Doyle, E. L.; Christian, M.; Wang, L.; Zhang, Y.; Schmidt, C.; Baller, J. A.; Somia, N. V.; Bogdanove, A. J.; Voytas, D. F., Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res* **2011,** *39*, e82.
(32) Huang, P.; Xiao, A.; Zhou, M.; Zhu, Z.; Lin, S.; Zhang, B., Heritable gene targeting in zebrafish using customized TALENs. *Nature biotechnology* **2011,** *29*, 699-700.

(33) Geissler, R.; Scholze, H.; Hahn, S.; Streubel, J.; Bonas, U.; Behrens, S. E.; Boch, J., Transcriptional activators of human genes with programmable DNA-specificity. *PloS one* **2011,** *6*, e19509.
(34) Li, T.; Huang, S.; Zhao, X.; Wright, D. A.; Carpenter, S.; Spalding, M. H.; Weeks, D. P.; Yang, B., Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic Acids Res* **2011,** *39*, 6315-25.
(35) Morbitzer, R.; Elsaesser, J.; Hausner, J.; Lahaye, T., Assembly of custom TALE-type DNA binding domains by modular cloning. *Nucleic Acids Res* **2011,** *39*, 5790-9.
(36) Sander, J. D.; Cade, L.; Khayter, C.; Reyon, D.; Peterson, R. T.; Joung, J. K.; Yeh, J. R., Targeted gene disruption in somatic zebrafish cells using engineered TALENs. *Nature biotechnology* **2011,** *29*, 697-8.
(37) Weber, E.; Gruetzner, R.; Werner, S.; Engler, C.; Marillonnet, S., Assembly of designer TAL effectors by Golden Gate cloning. *PloS one* **2011,** *6*, e19722.
(38) Zhang, F.; Cong, L.; Lodato, S.; Kosuri, S.; Church, G. M.; Arlotta, P., Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **2011,** *29*, 149-53.
(39) Sanjana, N. E.; Cong, L.; Zhou, Y.; Cunniff, M. M.; Feng, G.; Zhang, F., A transcription activator-like effector toolbox for genome engineering. *Nature protocols* **2012,** *7*, 171-92.
(40) Reyon, D.; Tsai, S. Q.; Khayter, C.; Foden, J. A.; Sander, J. D.; Joung, J. K., FLASH assembly of TALENs for high-throughput genome editing. *Nature biotechnology* **2012,** *30*, 460-5.
(41) Liang, J.; Chao, R.; Abil, Z.; Bao, Z.; Zhao, H., FairyTALE: a high-throughput TAL effector synthesis platform. *ACS synthetic biology* **2014,** *3*, 67-73.
(42) Mussolino, C.; Morbitzer, R.; Lutge, F.; Dannemann, N.; Lahaye, T.; Cathomen, T., A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res* **2011,** *39*, 9283-93.
(43) Hockemeyer, D.; Wang, H.; Kiani, S.; Lai, C. S.; Gao, Q.; Cassady, J. P.; Cost, G. J.; Zhang, L.; Santiago, Y.; Miller, J. C.; Zeitler, B.; Cherone, J. M.; Meng, X.; Hinkley, S. J.; Rebar, E. J.; Gregory, P. D.; Urnov, F. D.; Jaenisch, R., Genetic engineering of human pluripotent cells using TALE nucleases. *Nature Biotechnology* **2011,** *29*, 731-734.
(44) Wood, A. J.; Lo, T. W.; Zeitler, B.; Pickle, C. S.; Ralston, E. J.; Lee, A. H.; Amora, R.; Miller, J. C.; Leung, E.; Meng, X.; Zhang, L.; Rebar, E. J.; Gregory, P. D.; Urnov, F. D.; Meyer, B. J., Targeted genome editing across species using ZFNs and TALENs. *Science* **2011,** *333*, 307.
(45) Mahfouz, M. M.; Li, L.; Shamimuzzaman, M.; Wibowo, A.; Fang, X.; Zhu, J. K., De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks. *Proceedings of the national academy of sciences of the United States of America* **2011,** *108*, 2623-8.
(46) Miller, J. C.; Tan, S.; Qiao, G.; Barlow, K. A.; Wang, J.; Xia, D. F.; Meng, X.; Paschon, D. E.; Leung, E.; Hinkley, S. J.; Dulay, G. P.; Hua, K. L.; Ankoudinova, I.; Cost, G. J.; Urnov, F. D.; Zhang, H. S.; Holmes, M. C.; Zhang, L.; Gregory, P. D.; Rebar, E. J., A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **2011,** *29*, 143-8.

(47) Waters, C. A.; Strande, N. T.; Wyatt, D. W.; Pryor, J. M.; Ramsden, D. A., Nonhomologous end joining: a good solution for bad ends. *DNA repair* **2014,** *17*, 39-51.
(48) Spies, M.; Fishel, R., Mismatch Repair during Homologous and Homeologous Recombination. *Cold Spring Harbor perspectives in biology* **2015,** *7*.
(49) Rouet, P.; Smih, F.; Jasin, M., Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proceedings of the national academy of sciences of the United States of America* **1994,** *91*, 6064-8.
(50) Smih, F.; Rouet, P.; Romanienko, P. J.; Jasin, M., Double-strand breaks at the target locus stimulate gene targeting in embryonic stem cells. *Nucleic acids research* **1995,** *23*, 5012-9.
(51) Bitinaite, J.; Wah, D. A.; Aggarwal, A. K.; Schildkraut, I., FokI dimerization is required for DNA cleavage. *Proceedings of the national academy of sciences of the United States of America* **1998,** *95*, 10570-5.
(52) Sun, N.; Liang, J.; Abil, Z.; Zhao, H., Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Molecular bioSystems* **2012,** *8,* 1255-63.
(53) Christian, M.; Cermak, T.; Doyle, E. L.; Schmidt, C.; Zhang, F.; Hummel, A.; Bogdanove, A. J.; Voytas, D. F., Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **2010,** *186*, 757-61.
(54) Li, T.; Huang, S.; Jiang, W. Z.; Wright, D.; Spalding, M. H.; Weeks, D. P.; Yang, B., TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic acids research* **2011,** *39*, 359-72.
(55) Tesson, L.; Usal, C.; Menoret, S.; Leung, E.; Niles, B. J.; Remy, S.; Santiago, Y.; Vincent, A. I.; Meng, X.; Zhang, L.; Gregory, P. D.; Anegon, I.; Cost, G. J., Knockout rats generated by embryo microinjection of TALENs. *Nature biotechnology* **2011,** *29*, 695-6.
(56) Lehmann, R.; Nusslein-Volhard, C., The maternal gene nanos has a central role in posterior pattern formation of the Drosophila embryo. *Development* **1991,** *112*, 679-91.
(57) Ahringer, J.; Kimble, J., Control of the sperm-oocyte switch in *Caenorhabditis elegans* hermaphrodites by the fem-3 3' untranslated region. *Nature* **1991,** *349*, 346-8.
(58) Wickens, M.; Bernstein, D. S.; Kimble, J.; Parker, R., A PUF family portrait: 3'UTR regulation as a way of life. *Trends in genetics : Trends in genetics* **2002,** *18*, 150-7.
(59) Quenault, T.; Lithgow, T.; Traven, A., PUF proteins: repression, activation and mRNA localization. *Trends in cell biology* **2011,** *21*, 104-12.
(60) Miller, M. A.; Olivas, W. M., Roles of Puf proteins in mRNA degradation and translation. *Wiley interdisciplinary reviews. RNA* **2011,** *2*, 471-92.
(61) Qiu, C.; McCann, K. L.; Wine, R. N.; Baserga, S. J.; Hall, T. M., A divergent Pumilio repeat protein family for pre-rRNA processing and mRNA localization. *Proceedings of the national academy of sciences of the United States of America* **2014,** *111*, 18554-9.
(62) Un Huh, S.; Paek, K. H., Role of Arabidopsis Pumilio RNA binding protein 5 in virus infection. *Plant signaling & behavior* **2013,** *8*, e23975.

(63) Narita, R.; Takahasi, K.; Murakami, E.; Hirano, E.; Yamamoto, S. P.; Yoneyama, M.; Kato, H.; Fujita, T., A novel function of human Pumilio proteins in cytoplasmic sensing of viral infection. *PLoS pathogens* **2014,** *10*, e1004417.
(64) Zamore, P. D.; Williamson, J. R.; Lehmann, R., The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins. *RNA* **1997,** *3*, 1421-33.
(65) Zhang, B.; Gallegos, M.; Puoti, A.; Durkin, E.; Fields, S.; Kimble, J.; Wickens, M. P., A conserved RNA-binding protein that regulates sexual fates in the C. elegans hermaphrodite germ line. *Nature* **1997,** *390*, 477-84.
(66) Edwards, T. A.; Pyle, S. E.; Wharton, R. P.; Aggarwal, A. K., Structure of Pumilio reveals similarity between RNA and peptide binding motifs. *Cell* **2001,** *105*, 281-9.
(67) Wang, X.; Zamore, P. D.; Hall, T. M., Crystal structure of a Pumilio homology domain. *Molecular cell* **2001,** *7*, 855-65.
(68) Abil, Z.; Denard, C. A.; Zhao, H., Modular assembly of designer PUF proteins for specific post-transcriptional regulation of endogenous RNA. *Journal of biological engineering* **2014,** 8, DOI: 10.1186/1754-1611-8-7.
(69) Campbell, Z. T.; Valley, C. T.; Wickens, M., A protein-RNA specificity code enables targeted activation of an endogenous human transcript. *Nature structural & molecular biology* **2014,** *21*, 732-8.
(70) Dong, S.; Wang, Y.; Cassidy-Amstutz, C.; Lu, G.; Bigler, R.; Jezyk, M. R.; Li, C.; Hall, T. M.; Wang, Z., Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. *The Journal of biological chemistry* **2011,** *286*, 26732-42.
(71) Filipovska, A.; Razif, M. F.; Nygard, K. K.; Rackham, O., A universal code for RNA recognition by PUF proteins. *Nature chemical biology* **2011,** *7*, 425-7.
(72) Lunde, B. M.; Moore, C.; Varani, G., RNA-binding proteins: modular design for efficient function. *Nature reviews. Molecular cell biology* **2007,** *8*, 479-90.
(73) Mackay, J. P.; Font, J.; Segal, D. J., The prospects for designer single-stranded RNA-binding proteins. *Nature structural & molecular biology* **2011,** *18*, 256-61.
(74) Chen, Y.; Varani, G., Engineering RNA-binding proteins for biology. *The FEBS journal* **2013,** *280*, 3734-54.
(75) Wang, Y.; Wang, Z.; Tanaka Hall, T. M., Engineered proteins with Pumilio/fem-3 mRNA binding factor scaffold to manipulate RNA metabolism. *The FEBS journal* **2013,** *280*, 3755-67.
(76) Choudhury, R.; Wang, Z., Manipulation of RNA using engineered proteins with customized specificity. *Advances in experimental medicine and biology* **2014,** *825*, 199-225.
(77) Filipovska, A.; Rackham, O., Designer RNA-binding proteins: New tools for manipulating the transcriptome. *RNA biology* **2011,** *8*, 978-83.
(78) Querido, E.; Chartrand, P., Using fluorescent proteins to study mRNA trafficking in living cells. *Methods in cell biology* **2008,** *85*, 273-92.
(79) Ozawa, T.; Natori, Y.; Sato, M.; Umezawa, Y., Imaging dynamics of endogenous mitochondrial RNA in single living cells. *Nature methods* **2007,** *4*, 413-9.
(80) Yamada, T.; Yoshimura, H.; Inaguma, A.; Ozawa, T., Visualization of nonengineered single mRNAs in living cells using genetically encoded fluorescent probes. *Analytical chemistry* **2011,** *83*, 5708-14.

(81) Tilsner, J.; Linnik, O.; Christensen, N. M.; Bell, K.; Roberts, I. M.; Lacomme, C.; Oparka, K. J., Live-cell imaging of viral RNA genomes using a Pumilio-based reporter. *The Plant journal : for cell and molecular biology* **2009,** *57*, 758-70.
(82) Cooke, A.; Prigge, A.; Opperman, L.; Wickens, M., Targeted translational regulation using the PUF protein family scaffold. *Proceedings of the national academy of sciences of the United States of America* **2011,** *108*, 15870-5.
(83) Cao, J.; Arha, M.; Sudrik, C.; Schaffer, D. V.; Kane, R. S., Bidirectional regulation of mRNA translation in mammalian cells by using PUF domains. *Angewandte Chemie* **2014,** *53*, 4900-4.
(84) Ghildiyal, M.; Zamore, P. D., Small silencing RNAs: an expanding universe. *Nature reviews. Genetics* **2009,** *10*, 94-108.
(85) Choudhury, R.; Tsai, Y. S.; Dominguez, D.; Wang, Y.; Wang, Z., Engineering RNA endonucleases with customized sequence specificities. *Nature communications* **2012,** *3*, 1147.
(86) Wang, Y.; Cheong, C. G.; Hall, T. M.; Wang, Z., Engineering splicing factors with designed specificities. *Nature methods* **2009,** *6*, 825-30.
(87) Nakamura, T.; Yagi, Y.; Kobayashi, K., Mechanistic insight into pentatricopeptide repeat proteins as sequence-specific RNA-binding proteins for organellar RNAs in plants. *Plant & cell physiology* **2012,** *53*, 1171-9.
(88) Barkan, A.; Small, I., Pentatricopeptide repeat proteins in plants. *Annual review of plant biology* **2014,** *65*, 415-42.
(89) Small, I. D.; Peeters, N., The PPR motif - a TPR-related motif prevalent in plant organellar proteins. *Trends in biochemical sciences* **2000,** *25*, 46-7.
(90) Howard, M. J.; Lim, W. H.; Fierke, C. A.; Koutmos, M., Mitochondrial ribonuclease P structure provides insight into the evolution of catalytic strategies for precursor-tRNA 5' processing. *Proceedings of the national academy of sciences of the United States of America* **2012,** *109*, 16149-54.
(91) Ringel, R.; Sologub, M.; Morozov, Y. I.; Litonin, D.; Cramer, P.; Temiakov, D., Structure of human mitochondrial RNA polymerase. *Nature* **2011,** *478*, 269-73.
(92) Yin, P.; Li, Q.; Yan, C.; Liu, Y.; Liu, J.; Yu, F.; Wang, Z.; Long, J.; He, J.; Wang, H. W.; Wang, J.; Zhu, J. K.; Shi, Y.; Yan, N., Structural basis for the modular recognition of single-stranded RNA by PPR proteins. *Nature* **2013,** *504*, 168-71.
(93) Barkan, A.; Rojas, M.; Fujii, S.; Yap, A.; Chong, Y. S.; Bond, C. S.; Small, I., A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS genetics* **2012,** *8*, e1002910.
(94) Takenaka, M.; Zehrmann, A.; Brennicke, A.; Graichen, K., Improved computational target site prediction for pentatricopeptide repeat RNA editing factors. *PloS one* **2013,** *8*, e65343.
(95) Yagi, Y.; Hayashi, S.; Kobayashi, K.; Hirayama, T.; Nakamura, T., Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. *PloS one* **2013,** *8*, e57286.
(96) Fujii, S.; Bond, C. S.; Small, I. D., Selection patterns on restorer-like genes reveal a conflict between nuclear and mitochondrial genomes throughout angiosperm evolution. *Proceedings of the national academy of sciences of the United States of America* **2011,** *108*, 1723-8.

(97) Kerem, B.; Rommens, J. M.; Buchanan, J. A.; Markiewicz, D.; Cox, T. K.; Chakravarti, A.; Buchwald, M.; Tsui, L. C., Identification of the cystic fibrosis gene: genetic analysis. *Science* **1989,** *245*, 1073-80.

Chapter 2. Engineering of TALEN for Editing of the Cystic Fibrosis-Associated Mutation

2.1 Introduction

Cystic fibrosis (CF) is a chronic autosomal recessive disorder with an incidence of about one in 2,500 and symptoms including elevated sweat chloride concentrations, exocrine pancreatic insufficiency, progressive obstructive lung disease, and male infertility[1-2]. The disease is caused by mutations in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene[3-4]. Thousands of mutations in the *CFTR* gene have been reported, and the most common one is a 3 bp deletion in exon 10 found in almost 70% of patients[5]. This genotype causes the deletion of phenylalanine at the 508 position (F508del), resulting in misfolding and biosynthetic arrest of CFTR[6-7]. Due to its monogenic nature, CF is a potentially ideal target for genome-based therapies.

While the initial demonstration that transduction of *CFTR* cDNA could correct the CF phenotype in cultured cells was promicing[8], in patients with CF, adenovirus-mediated gene transfer of *CFTR* cDNA demonstrated only transient correction of the chloride transport defect[9]. One reason of a transient effect of the aforesaid approach could be the absence of the endogenous gene expression regulation of heterogenous *CFTR*.

Due to its ability to permanently correct a genetic deficiency, targeted genetic manipulation using custom-designed nucleases has become a powerful approach

recently[10-13]. This approach is based on the observation that induction of a double-strand break (DSB) in the target DNA sequence increases the homologous recombination efficacy by several orders of magnitude[14-15]. Thus, it was demonstrated that co-delivery of a site-specific nuclease with a donor plasmid bearing locus-specific homology arms can efficiently correct a genetic abnormality[16].

Transcription activator-like effector (TALE) protein fusions with nonspecific endonuclease *Fok*I (TALEN) are one of the recently developed potent tools used for targeted genome modification. TALE proteins are transcription activators that were found to be secreted into host plant cells by plant-pathogenic *Xanthomonas* bacteria[17-19]. The central part of the TALE gene is a highly repetitive DNA binding motif, consisting of nearly identical repeats of 33-35 amino acids[20-21], while the last repeat is only a half repeat consisting of 19-20 amino acids[20]. The target sequence recognition by TALE proteins is governed by two hypervariable amino acids at positions 12 and 13 referred to as "repeat variable di-residue" (RVD)[22-23]. The preference of certain nucleobases in the target sequence by a succession of RVD follows a simple code and a modular 1 repeat : 1 base recognition pattern. Thus, the RVD NI prefers adenine, NG thymine, HD a cytosine, while NN recognizes either adenine or a guanine[22-23]. A TALE with a desired specificity can be designed by assembling the individual repeats by following this code[22]. The aim of this project was to design a TALEN specific for the *CFTR* gene to permit a highly specific and efficient homologous recombination event and correction of the *CFTR* gene in airway epithelial cells with the F508del genotype of *CFTR*.

2.2 Results

2.2.1 Design and Optimization of TALENs Using the Episomal GFP Reporter

Because the FokI endonuclease functions as a dimer[24], a TALEN heterodimer (TALEN-CFTR-L and TALEN-CFTR-R) was designed, where each TALE was constructed to target a 20 bp sequence separated by a 16 bp spacer. Each RVD was chosen in accordance with the TALE DNA recognition code[22-23] (Figure 2.1 a, b and c). The trinucleotide deletion responsible for the F508del was excluded from the TALEN-CFTR-L target site, so that it binds to the mutant F508del locus of *CFTR*, and not the wild type *CFTR* locus. Each TALE was fused with a nonspecific DNA cleavage domain of FokI endonuclease via a flexible G_4S linker. The TALE were constructed using the AvrXa10 TALE [25] as a template. The N-terminal region (NTR) of the TALEs were designed to exclude the first 152 amino acids in the native TALE, as it was shown that this region specifies the transport of the protein into plant cells and is not essential for other functions[26]. To study the effect of the C-terminal length of TALE on the efficacy of TALEN heterodimerization and function, the C-terminal region (CTR) was truncated to various lengths. Our TALEN constructs contained the first 20, 31, 62 amino acids of the AvrXa10 TALE, or the first 31 amino acids followed by a long (G_4S)x6 linker (CTRs C20, C31, C62, and C31-GS, respectively). The TALE-FokI fusions L and R were cloned separately under the cytomegalovirus (CMV) promoter.

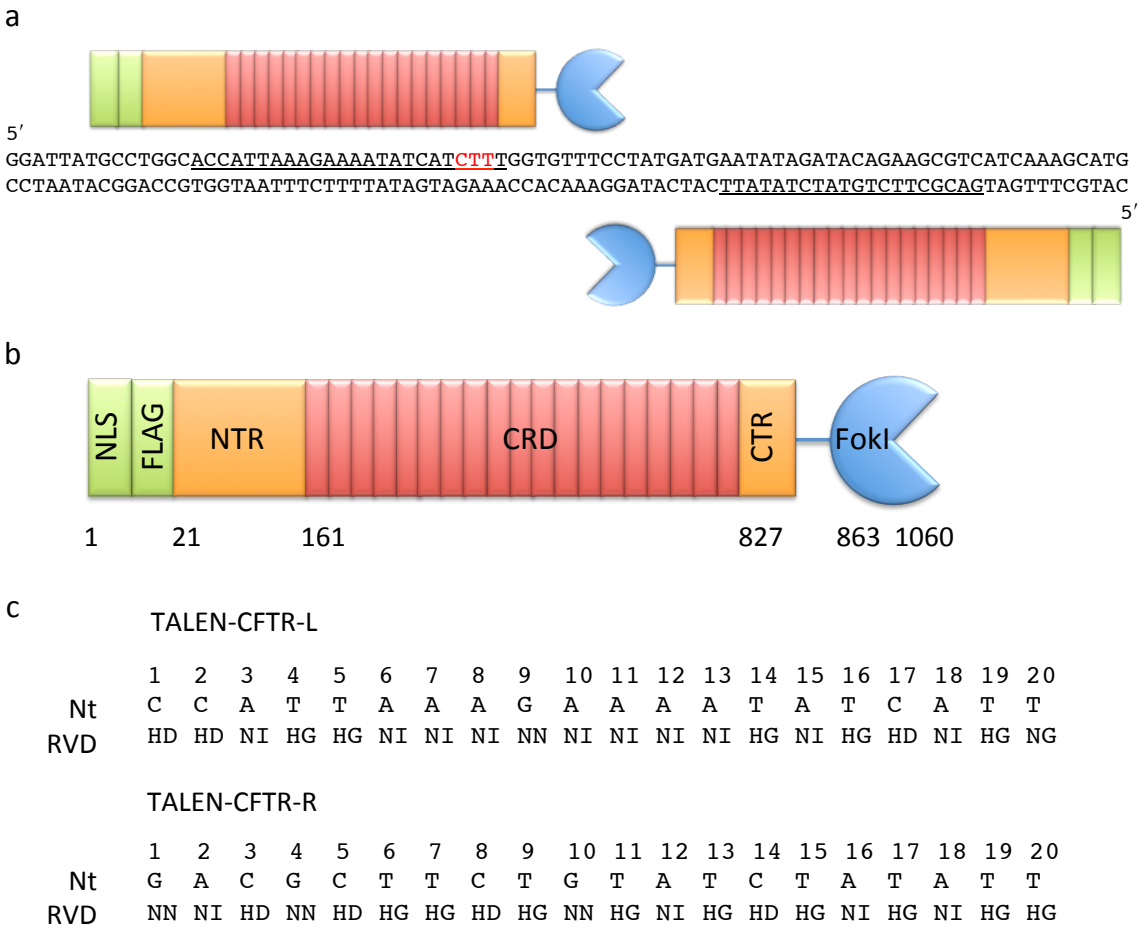


Figure 2.1 Design of TALEN-CFTR. **(a)** Schematic of *CFTR* exon 10 sequence fragment and TALEN-CFTR. Red sequence, F508del mutation. Underlined sequence, TALEN-binding sites. **(b)** Schematic of TALEN. NLS, SV40 nuclear localization signal. FLAG, FLAG protein tag. NTR, N-terminal region. CRD, Central repeat domain. CTR, C-terminal region. **(c)** The 5'-3' sequences of TALEN-CFTR binding sites as well as RVD array of the TALEN, listed from the N-terminus to the C-terminus of the proteins.

To test the gene targeting efficacy of the designer TALEN, *EGFP*-gene targeting reporter was constructed, where the *EGFP* gene under the CMV promoter was interrupted by a stop codon followed by a 56 bp sequence from the *CFTR* locus containing the F508del (Figure 2.2 a). The EGFP reporter was inserted into a plasmid containing a donor EGFP sequence with no promoter and lacking the first

37 bp at the 5′ of its sequence. Upon introduction of a double stranded break adjacent to the F508del, the cell's homologous recombination machinery should facilitate the repair of the reporter EGFP using the donor EGFP sequence (Figure 2.2 b). The donor-reporter EGFP plasmid, as well as the TALEN-CFTR-L and TALEN-CFTR-R plasmids were used to co-transfect the HeLa cell line, and 72 hours later, the percentage of EGFP-positive cells was measured using flow cytometry. The *EGFP* correction efficacy was estimated as the percentage of EGFP-positive cells.

a

b

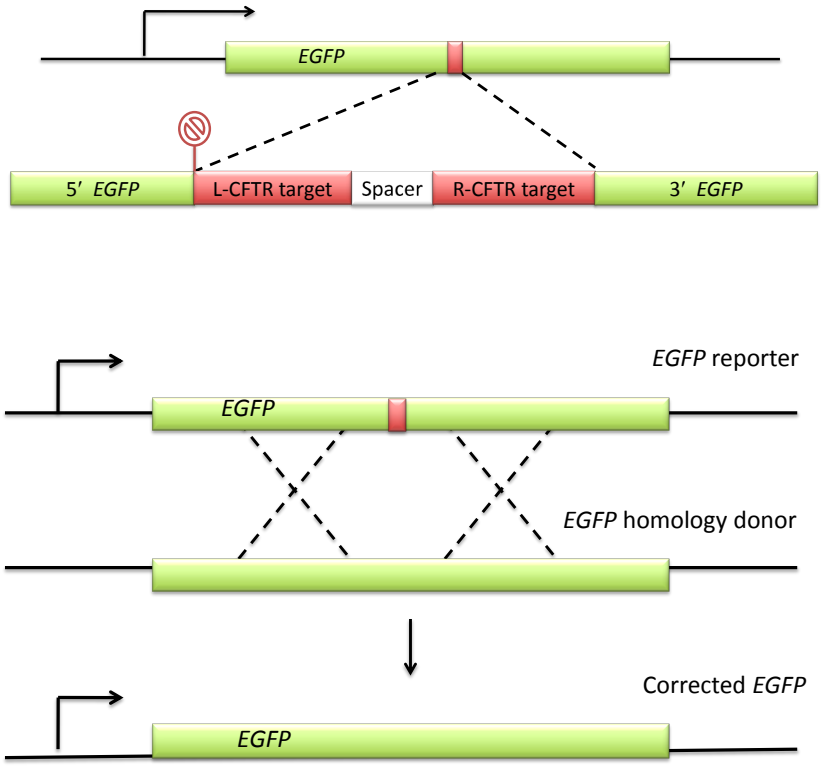


Figure 2.2 Schematic of *EGFP* gene targeting reporter system. **(a)** The structure of the *EGFP*-interrupting insert, with TALEN-binding sites (red), and a 16-nt spacer (white). Stop codon position is indicated with a "Stop" sign. **(b)** Schematic of *EGFP* correction through a homologous recombination event.

In the presence of TALENs, 4.2% to 8.5 % EGFP-positive cells were detected (Figure 2.3), while in the absence of TALENs, the percentage of eGFP-positive cells was undetectably small (data not shown), indicating that successful gene editing events took place due to the endonuclease activity of designer TALENs. We also observed that TALENs with longer CTRs generally exhibited slightly increased correction efficacies compared to shorter CTRs (Figure 2.3), whereas TALENs with the C31-GS truncations demonstrated correction efficacies considerably lower than C62 truncations, despite the same total number of amino acids after the CRD. The worse performance of C31-GS constructs could be due to the shorter effective length of C31-GS, due to a more compact structure of the random coil the GS linker assumes. It is possible that the C62 tail forms a more rigid structure, allowing enough length for the two FokI domains to reach each other, bridging the 16 bp-spacer between the two TALE recognition sites.

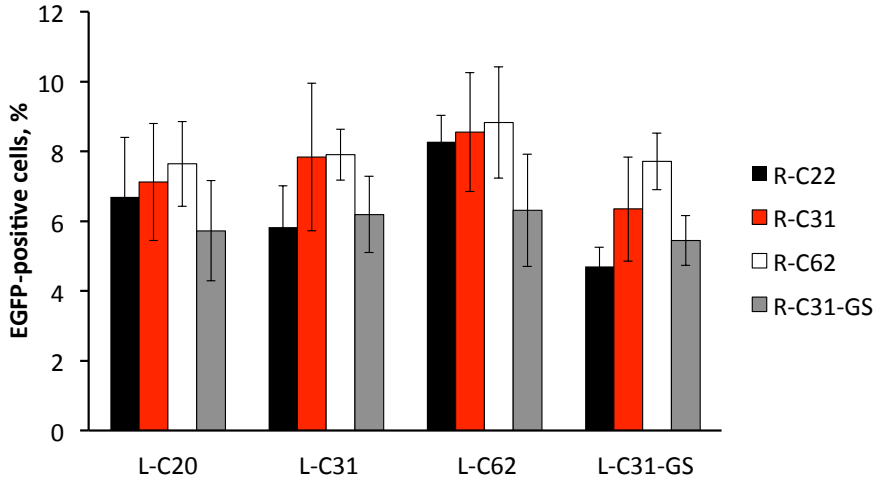


Figure 2.3 Analysis of TALEN performance with various CTR lengths, measured by episomal reporter correction. Reporter correction in HeLa cells was estimated by quantifying EGFP-positive cells via flow cytometry.

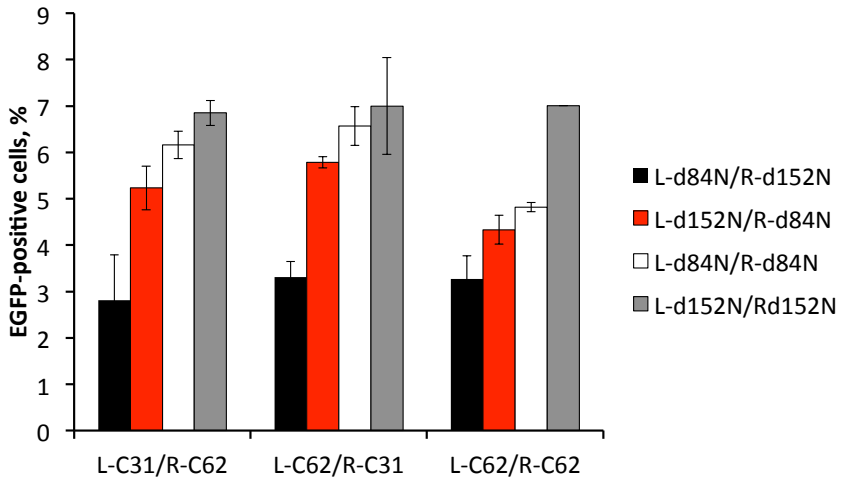


Figure 2.4 Analysis of TALEN performance with various NTR and CTR lengths, measured by episomal reporter correction. Reporter correction in HeLa cells was estimated by quantifying EGFP-positive cells via flow cytometry.

In order to confirm that deletion of a portion of NTR does not compromise the function of the designer TALEN, we constructed TALENs with longer NTR, truncated to only the first 84 amino acids (d84N). Only TALEN configurations that previously showed highest EGFP correction efficacies (L-C31/RC-62, L-C62/R-C31, and L-C62/R-C62) were analyzed with additional NTR lengths. TALENs with longer NTR did not demonstrate an increase in activity compared to TALENs with the d152N truncation of NTR (Figure 2.4), thus confirming that NTR between the amino acids 84 and 152 of the original TALE was not essential for the DNA-binding activity of the TALE. The reason behind the reduced activity of TALENs with longer NTR, however, remains unknown.

2.2.2 Optimization of TALENs Using the Chromosomal GFP Reporter

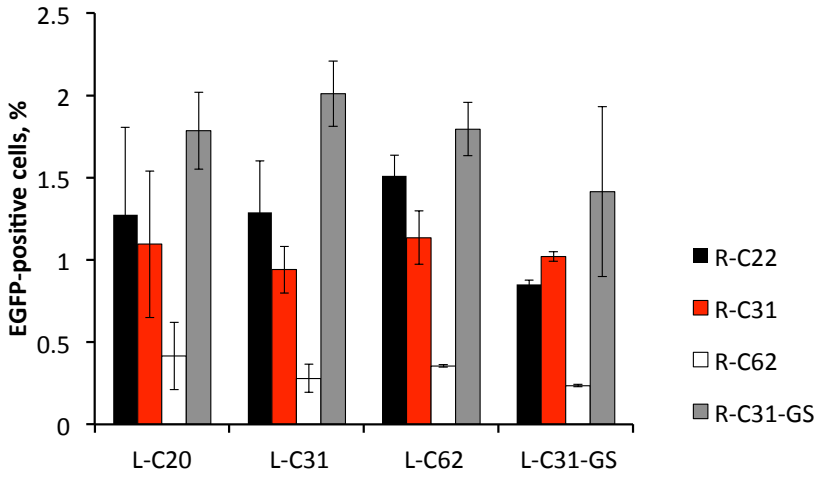


Figure 2.5 Analysis of TALEN performance with various CTR lengths, measured by chromosomal reporter correction. Reporter correction in HeLa cells was estimated by quantifying EGFP-positive cells via flow cytometry.

Although high efficacy episomal correction of reporter *EGFP* was demonstrated above, the correction of a chromosomal gene could be less efficient due to fewer reporter sequences present per cell, as well as lower accessibility of the reporter gene to cleavage due to higher order chromosome structure. To test the chromosomal gene targeting efficacy of the designer TALEN, the *EGFP* reporter disrupted by a 56 bp sequence as above was retrovirally integrated into a HeLa cell line genome. The HeLa cell line with the integrated sequences was co-transfected with pCMV5-TEN-CFTR-L, pCMV5-TEN-CFTR-R, and the homology donor plasmid, and the percentage of EGFP-positive cells was measured by flow cytometry. The highest percentage of EGFP-positive cells in these conditions was ~2 %, a significant drop from that observed for episomal correction. In addition, contrary to our observations for episomal correction, chromosomal correction efficacy of *EGFP* was

lowest when TALENs with longest CTR were used, and highest when TALENs with the C31-GS tales were used (Figure 2.5). We suggest that these contrasting results are due to different states of chromosomal structure in episomal versus chromosomally located target sequences. It is conceivable that due to higher order chromatin structure, the effective length of the 16 bp-spacer between the two TALE -binding sites was shorter in a chromosomally located EGFP reporter, and that the CTR length of 62 amino acids was too long for favorable dimerization of FokI.

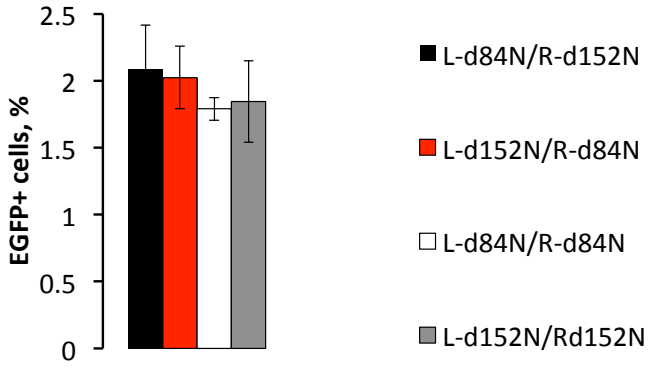


Figure 2.6 Analysis of TALEN performance with various NTR lengths, measured by episomal reporter correction. CTR configuration L-C31/R-C31-GS was used. Reporter correction in HeLa cells was estimated by quantifying EGFP-positive cells via flow cytometry.

We similarly checked if TALENs with the longer NTR (d84N) exhibited different correction efficacy compared with the d152N. For this purpose, we used the TALEN configuration that demonstrated the highest activity for chromosomal correction, L-C31/R-C31-GS. TALEN configuration with longer NTR did not result in significant difference in correction efficacy (Figure 2.6), once again confirming that the NTR between the amino acids 84 and 152 of the original TALE was not essential for the DNA-binding activity of the TALE.

2.2.3 Correction of Endogenous *CFTR* Gene in Airway Epithelial Cells

To test the gene correction efficacy of the designer TALEN in targeting endogenous *CFTR*, we utilized the human airway epithelial cell lines KKLEB[27], UNCCF3T[27], and CFT1[28]. All three cell lines were developed by immortalization from airway epithelial tissues of patients homozygous for the F508del mutation of *CFTR*[27-28]. By transfection of plasmids encoding GFP under CMV promoter using the Nucleofector Technology, we measured transfection efficiencies ranging between 20-40% for KKLEB, 8-13% for UNCCF3T, and 60-70 % for the CFT1 cell line (Table 2.1). Due to its highest efficiency of being transfected, CFT1 cells were used further for *CFTR* correction experiments.

Name	Cell type	Immortalization	Nucleofection efficiency
KKLEB[27]	Bronchial	hTERT/SV40LT	20-40%
UNCCF3T[27]	Bronchial	Bmi-1/hTERT	8-13%
CFT1[28]	Tracheobronchial	E6/E7	60-70%

Table 2.1 Characteristics of the human airway epithelia cell lines used in this study.

For the correction of endogenous *CFTR* with the F508del genotype, we constructed homology donors with various homology lengths and structure. Since efficient intrachromosomal recombination requires uninterrupted homology of 134-232 bp[29], and because of the positive correlation between the frequency of homologous recombination and homology length[30], we constructed plasmids containing 300 bp, 500 bp, and 1000 bp homology arms on each side of the mutation-containing mismatch. With recent reports of genome editing using single stranded oligodeoxynucleotides (ssODN) in combination with zinc-finger

nucleases[31-32], we also constructed ssODNs containing CFTR homology sequences 50 and 100 nt in length to use with our designer TALENs. All of the homology donors we constructed contained not only the F508del correcting insertion, but also TALEN-blocking mutations – 8 silent mutations to minimize re-binding of TALENs after the homologous recombination event (Figure 2.7a and b). Two of these silent mutations also created a unique new *EcoR*V restriction site, which we subsequently used for estimation of correction efficacy using restriction-digestion.

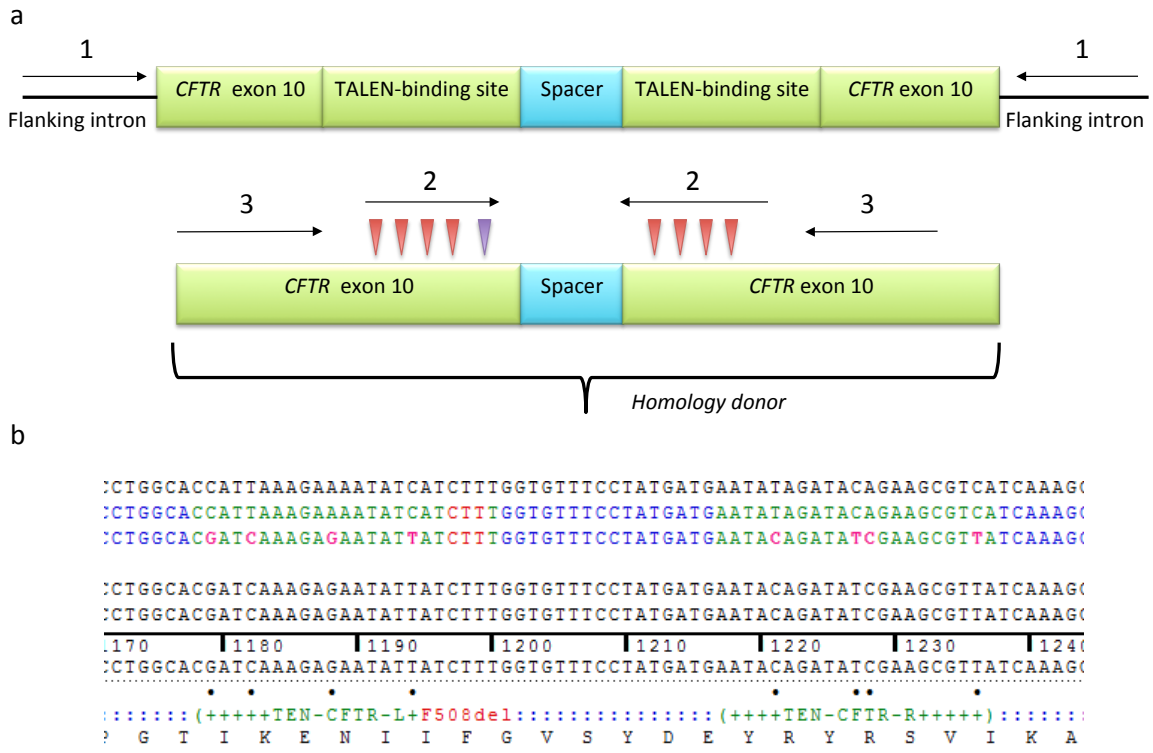


Figure 2.7 Illustration of endogenous *CFTR* exon 10 and homology donor. **(a)** Schematic of CFTR exon 10, including TALEN-binding sites and a 16-bp spacer between them (top) and schematic of homology donor, including TALEN-blocking mutations (red wedges) and the F508 insertion (purple wedge). Locations of primer pairs 1, 2, and 3 for quantitative PCR are indicated with arrows. **(b)** DNA sequence showing TALEN-binding sites, TALEN-blocking mutations, and the CTT trinucleotide (coding for F508).

We co-nucleofected different donor constructs separately with the plasmids encoding TALEN-CFTR with the CTR configuration L-C31/R-C31-GS and NTR configuration d152N, and allowed the cells to grow for 72 hours before analysis. Since correction efficacy estimation through PCR amplification and consecutive restriction digestion did not yield correction efficacy above the detection limit (~2 %, data not shown), we designed an alternative quantitative-PCR based method for detection of low correction efficacies. The method consists of amplifying the DNA sequence with the primers located outside the sequences covered by the homology donor (primer pair 1, Figure 2.7a) from genomic DNA isolated from TALEN/CFTR-donor treated cells. This is followed by gel-purification of the amplified product to remove any trace homology donor still present in cells. Control reactions are carried out without the DNA-polymerase, to confirm that the template for the next round of PCR is the PCR product amplified from genomic DNA, and not non-specific background amplification. Quantitative PCR is performed using the previous amplification product as a template, using primer pairs 2 or 3 (Figure 2.7a). Primer pair 2 binds to the sequence containing the TALEN-blocking mutations, and the 3′-end of one of the primers also ends with the sequence corresponding to the correcting mutation, absent in the uncorrected genomic *CFTR* sequence of CFT1 (F508 insertion, Figure 2.7a). Hence, primer pair 2 only amplifies from the corrected *CFTR* sequence. Primer pair 3 binds to sequences immediately outside the primer pair 2, and is used as a control for template loading. The same amount of template was used in all quantitative PCR reactions, which was confirmed by almost identical

Ct values in all amplifications using the primer pair 3. For construction of a standard curve for estimation of correction efficacy, we mixed PCR amplification products from F508del (amplified from genomic DNA of CFT1 cells) or donor sequence (amplified from the p2x1000 donor plasmid) in ratios ranging from 0.01 % to 100 % of donor sequence.

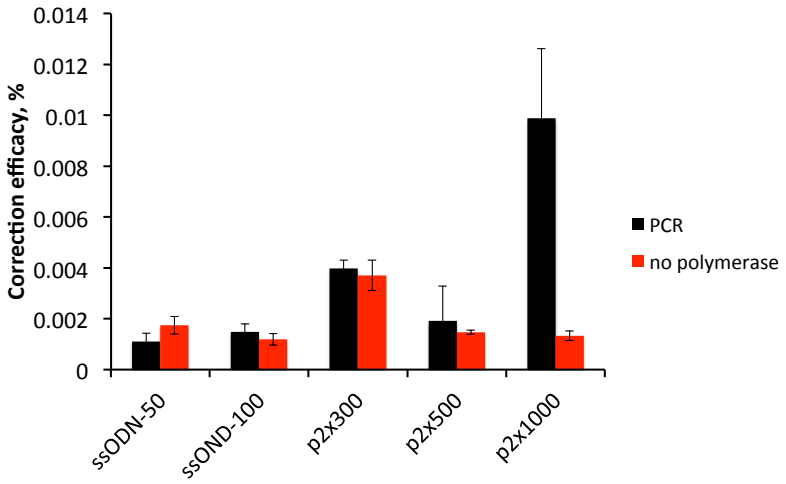


Figure 2.8 Correction efficiency of endogenous *CFTR* measured by quantitative PCR.

The quantitative PCR procedure described above allowed detection of polymorphisms smaller in quantities than what is detectable by PCR/restriction digestion. Only co-nucleofection of cells with the TALEN constructs and the donor construct with the 1,000 bp-long homology arms resulted in detectable correction efficacy, which was around 0.01 % (Figure 2.8). Other homology donors did not allow detection of correction efficacies above background amplification. We conclude that homology arms of at least 1 kb on each side of the insertion are required for detectable correction efficacy of *CFTR* using our TALEN system. Because of low correction efficacy, we did not follow with the phenotypical characterization of the TALEN-treated CFT1 cells.

2.3 Discussion

Designer nucleases, such as zinc-finger nucleases (ZFN), TALEN, and the CRISPR/Cas9 system (Cas9 nuclease in combination with the clustered regularly interspaced short palindromic repeats, CRISPR) have been developed as invaluable tools for genome editing applications. Especially promising is the application of these tools for treating genetic diseases, for they allow high-efficacy correction of an endogenous gene without the need of introducing an additional copy of a gene with unnatural expression regulation. In this work, we designed and optimized the TALEN system for the editing of the F508del mutation of *CFTR*, as an application of TALENs for tackling CF.

By comparison of various truncations, we investigated the optimal architecture of TALEN for the cleavage of a sequence from exon 10 of *CFTR*. The optimal NTR length was reported to be 147 residues N-terminally of TAL repeats in the context of TALE transcription factor[33], 155 residues[34] or 136 residues[35] in the context of TALENs, which is in agreement with our TALEN architecture with 140 residues. This region was reported to contain at least 126 residues that form a superhelical structure with four continuous pseudo-repeats resembling CRD repeats, which are important for the nonspecific binding of TALE to DNA[36].

Studies on ZFN demonstrated that the protein linker length that connects the DNA-binding domain to FokI affects the optimal spacer length between the two protein-binding sites[37-38]. Similarly, studies on TALEN showed that longer C-terminal configurations (~60 compared to ~20 residues) have a higher preference for longer spacer lengths between the two protein binding sites (12–30 bp

compared to 10–16 bp)[34-35, 39]. In accordance with these observations, we observed TALEN activity with both CTR linkers of 31 or 63 residues in the context of our target site with a 16 bp spacer. Surprisingly, the optimal CTR linker length of TALEN varied dramatically depending on whether the reporter gene was episomal or chromosomal. These results indicate that the local DNA conformation at the TALE target site plays as an important role in the effective distance between the two DNA-binding domains, as the length of the DNA backbone between them. We therefore conclude that for optimal results, the best TALE configuration with CTR ranging between 20 and 60 residues must be determined for each genomic target site.

Recently, correction of endogenous *CFTR* with the F508del using ZFN[40] and CRISPR/Cas9[41] was reported, however it is difficult to compare editing efficacies obtained by us with those reported previously. Using a 4.3 kb wild type *CFTR* donor sequence and an engineered ZFN cleaving 203 bp upstream of the F508del, correction of the mutation was observed in CF tracheal epithelial cells using a PCR assay[40]. However, the correction efficacy could not be determined due to it being below the detection limit of the assay based on PCR amplification/restriction digestion. The authors suggested that low editing efficacy (estimated to be < 1%) could be due to the long distance between the double strand break and the F508del mutation[40]. In another study, CRISPR/Cas9 genome-editing system was used to correct the F508del mutation in CF patient-derived cultured intestinal stem cells[41]. Since the gene correction procedure involved selection for puromycin-resistant cells[41], it is once again difficult to compare the correction efficacies between our investigations. However, their results imply that efficient nuclease-assisted editing

of *CFTR* is possible, provided that it is done in combination with site-specific co-integration of selection markers and subsequent selection for marker-resistant cells. Given these results, we conclude that the TALEN developed in this study for the correction of the F508del mutation of CFTR could potentially be used for editing of endogenous *CFTR* from CF patient-derived stem cells in combination with a homology donor sequence containing the exon 10 of *CFTR* and a selection marker.

2.4 Materials and Methods

2.4.1 Cloning of DNA Constructs

For chromosomal reporter correction, the enhanced green fluorescence protein (EGFP) gene was disrupted in the middle by the insertion of a sequence containing an in-frame stop codon and a 56 bp *CFTR* gene locus that is the target sequence for the engineered TALEN. This non-functional EGFP reporter sequence was cloned into the pLNCX2 retroviral vector (Clontech) downstream of the cytomegalovirus (CMV) promoter using the *Sal*I and *Hind*III restriction sites (plasmid pLNCX2-het-CFTR-EGFP*) and stably integrated into the HeLa cell line genome via retroviral transduction according to the manufacturer's protocol. Transfected Hela cells were selected in 500 mg/ml G418 for two weeks. A donor plasmid pCMV5-eGFP-donor containing the *EGFP* gene lacking a promoter and the first 37 bp of its sequence into the pNEB193 vector (New England Biolabs) was described elsewhere[39].

For episomal reporter construction, the plasmid pCMV5-Het-CFTR-EGFP1 containing both the donor and reporter EGFP sequences was constructed in the pCMV5 vector. For this purpose, the non-functional EGFP sequence with a stop

codon and the 56 bp *CFTR* gene locus insert (reporter EGFP) was sub-cloned using *Hind*III and *Sal*I restriction sites into the pCMV5-eGFP-donor plasmid.

TALEN-CFTR L and TALEN-CFTR R central repeat domain sequences were codon-optimized for expression in human cells and synthesized at DNA2.0 (Menlo Park, CA) in the pJ247 vector. The two constructs were each sub-cloned into the pCMV5 vector (a kind gift of Dr. David Russel, University of Texas Soutwestern), so that each contain a Kozac sequence, nuclear localization signal from Simian virus 40 (SV40), a FLAG tag at the N-terminus of the construct, NTR and CTR, or linker sequences of specified lengths, and FokI sequence at the C-terminus. The constructs were spliced using overlap extension PCR with primers bearing 100 bp overlaps, and inserted into the vector using *Not*I and *Hind*III restriction sites.

The sense sequence of the 50 bp olygodeoxynucleotide (ODN) donor sequence was 5′-CAGTTTTCCTGGATTATGCCTGGCACGATCAAAGAGAATATTATCTTTGGTGTTTCCTATGATGA ATACAGATATCGAAGCGTTATCAAAGCATGCCAACTAGAAGAGGT-3′, and the sense sequence of the 100 bp ODN donor sequence was 5′-CAGTGGAAGAATTTCATTCTGTTCTCAGTTTTCCTGGATTATGCCTGGCACGATCAAAGAGAATATTATCT TTGGTGTTTCCTATGATGAATACAGATATCGAAGCGTTATCAAAGCATGCCAACTAGAAGAGGTAAGAAAC TATGTGAAAACTTTTTG-3′. To construct donor plasmids containing *CFTR* gene exon 10 homology arms of 300 bp, 500 bp, and 1000 bp, genomic DNA from the airway epithelial cells CFT1 was purified using Wizard genomic DNA purification kit (Promega) according to manufacturer's instructions. The 3′ half-fragments of the donor constructs were amplified from the genomic DNA using the 50 bp sense ODN as a forward primer and the following reverse primers: 5′-

GCATGCCTGCAGAATGAGTGAACAAAATTCTCACCATTTCATAAAATGCATTTA-3′ for p2x300, 5′-GCATGCCTGCAGACAATTTTACCCCTCTAATTCTCTGCTGGC-3′ for p2x500, 5′-GCATGCCTGCAGCTGTTGAATGACTGAGTATATACATGGAAAGCCATTCAT-3′ for p2x1000. These amplicons were subsequently used as mega-primers for amplification of the entire fragments, with the following forward primers: 5′-TCTAGAGTCGACACTTTCCCTTGTATCTTTTGTGCATAGCAGAG-3′ for p2x300, 5′-TCTAGAGTCGACTGTGAAGATTAAATAAATTAATATAGTTAAAGCACATAGAACAGCA-3′ for p2x500, and 5′-TCTAGAGTCGACCAGATTTATCTTTGTATTGTTAAATCTGCTTATGCTTCTATTACTT-3′ for p2x1000. These three fragments were subsequently inserted into the pNEB-193 vector (New England Biolabs) using *Sal*I and *Pst*I restriction sites. All plasmid DNA used for airway epithelial cell nucleofection was purified using the Plasmid Mini Kit (Qiagen).

2.4.2 Transfection

Reporter HeLa cells were maintained in modified Eagle's Medium (MEM) supplemented with 10 % Fetal Bovine Serum (FBS). HeLa cells were plated a day before transfection at a density of $1x10^5$ cells/well in a 12-well plate in 1 ml of MEM with 10% FBS. The next day, transfection mixtures were prepared by mixing 50 µl OptiMEM (Life Technologies), 1 µg DNA, and 3 µl Fugene HD reagent (Promega). The DNA mixture contained 400 ng TALEN-CFTR L, 400 ng TALEN-CFTR R, and 200 ng donor plasmid (or the pCMV-donor-reporter plasmid, in the case of episomal correction). The transfection mixtures were incubated for 15 minutes at room

temperature and added to the wells with cells. The cells were allowed to recover and grow for 72 hours before flow cytometry.

2.4.3 Flow Cytometry

Cells were trypsinized from their culturing plates and resuspended in 200 µl phosphate buffered saline (PBS) supplemented with 10 mM EDTA. To determine the percentage of GFP-positive cells, 20 000 cells were analyzed by a BD LSRII flow cytometer (BD Biosciences). Transfection efficiency was determined by determining the percentage of GFP-positive cells that were transfected with functional GFP only. The transfection efficiency was determined to be around 80 %. The rate of gene targeting was determined by normalizing the percentage of GFP-positive cells in TALEN and donor co-transfected reporter cells to the transfection efficiency. The flow cytometry data were analyzed using the FCS Express 4 (BD Biosciences) software.

2.4.4 Airway Epithelial Cell Line Maintenance and Transfection

The CFT1 airway epithelial cell line (kind gift of James R. Yankaskas, University of North Carolina at Chapel Hill) was maintained on BD Biocoat Collagen I-coated 100 mm plates (BD Biosciiences) in the F12-7X medium containing 200 µg/ml gentamicin (Sigma), 10 µg/ml insulin (Sigma), 5 µg/ml transferrin (Sigma), 1 µM hydrocortisone (Sigma), 3.75 µg/ml endothelial cell growth supplement (Sigma), 30 nM triiodothyronine (Sigma), 25 ng/ml epidermal growth factor (Sigma), and 10 ng/ml Cholera toxin (Sigma). The cells were passaged at maximum expansion ratio

of 1:3 using 0.1 % trypsin/1 mM EDTA, and the trypsin was neutralized using twofold v/v excess soybean trypsin inhibitor (STI, Sigma). The UNCCF3T as well as the KKLEB airway epithelial cell lines (kind gift of Scott H. Randell, University of North Carolina at Chapel Hill) were maintained on BD Biocoat Collagen I-coated 100 mm plates (BD Biosciences) in the CnT-17 PCT Airway epithelium medium (Zenbio) and passaged at maximum expansion ratio of 1:8 using 0.1 % trypsin/1 mM EDTA, and the trypsin was neutralized using twofold v/v excess STI.

The Airway epithelial cells were transfected using Amaxa Epithelial, Bronchial (NHBE) Nucleofector Kit (Lonza) according to the manufacturer's instructions. Briefly, $5x10^5$ cells were nucleofected with 1 μg of each TALEN construct and 1 μg donor DNA. UNCN3T/UNCCF3T cells were nucleofected using the W-001 Nucleofector program, whereas HBE1/CFT1 and AALEB/KKLEB cells were nuclefected using T-020 Nucleofector program on the Nucleofector I Device. The cells were grown for 72 hours before assaying.

2.4.5 Quantitative PCR

Genomic DNA of CFT1 cells nucleofected with TALEN and donor constructs was purified using the Wizard Genomic DNA purification kit (Promega) according to the manufacturer's instructions. The *CFTR* exon 10 was amplified using primers 5′-CCTTCTACTCAGTTTTAGTCAGTAG-3′ and 5′-GTGTATAAAAAATAAGGATGGGGCTC-3′. In controls for nonspecific amplification, the Phusion High fidelity DNA polymerase (New England Biolabs) was omitted. Amplified CFTR *exon* 10 was gel-purified to remove any trace co-purified donor plasmid, and used as a template for quantitative PCR using (a) CFTR-TBM primers binding to the regions with TALEN blocking

mutations 5′-GCCTGGCACGATCAAAGAGAATATTATCT-3′ and 5′-GGCATGCTTTGATAACGCTTCGATATCTG-3′, and (b) CFTR-con primers binding to regions immediately outside of the regions with TALEN blocking mutations 5′-GTTTTCCTGGATTATGCCTGGCAC-3′ and 5′-ACCTCTTCTAGTTGGCATGCTTTG-3′. For construction of the standard curve, a 'modified reference' 2 kb fragment consisting entirely of the donor *CFTR* sequence was amplified from the pNEB-2x1000 plasmid and gel-purified. Likewise, an 'unmodified reference' 2 kb fragment consisting entirely of unmodified F508del *CFTR* sequence was amplified from genomic DNA purified from untransfected CFT1 cells, and gel-purified. Mixtures of these two references were prepared with the fraction of 'modified reference' ranging from 0.01 % to 100%, and the Ct values obtained using amplification with the TBM primers were used to generate the standard curve for the F508del *CFTR* sequence correction. Quantitative PCRs were run with 10 ng template, 500 nM primers, 0.02 U/µl Phusion High Fidelity DNA Polymerase in 1x Failsafe PCR Premix G (Epicentre). The correction efficacies were calculated by plotting the Ct values from unknown samples on the standard curve.

2.5 References

(1) Rosenstein, B. J.; Zeitlin, P. L., Cystic fibrosis. *Lancet* **1998,** *351*, 277-82.
(2) Knowles, M. R.; Durie, P. R., What is cystic fibrosis? *The New England journal of medicine* **2002,** *347*, 439-42.
(3) Riordan, J. R.; Rommens, J. M.; Kerem, B.; Alon, N.; Rozmahel, R.; Grzelczak, Z.; Zielenski, J.; Lok, S.; Plavsic, N.; Chou, J. L.; et al., Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **1989,** *245*, 1066-73.
(4) Rommens, J. M.; Iannuzzi, M. C.; Kerem, B.; Drumm, M. L.; Melmer, G.; Dean, M.; Rozmahel, R.; Cole, J. L.; Kennedy, D.; Hidaka, N.; et al., Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **1989,** *245*, 1059-65.

(5) Kerem, B.; Rommens, J. M.; Buchanan, J. A.; Markiewicz, D.; Cox, T. K.; Chakravarti, A.; Buchwald, M.; Tsui, L. C., Identification of the cystic fibrosis gene: genetic analysis. *Science* **1989,** *245*, 1073-80.
(6) Amaral, M. D., Therapy through chaperones: sense or antisense? Cystic fibrosis as a model disease. *Journal of inherited metabolic disease* **2006,** *29*, 477-87.
(7) Lukacs, G. L.; Durie, P. R., Pharmacologic approaches to correcting the basic defect in cystic fibrosis. *The New England journal of medicine* **2003,** *349*, 1401-4.
(8) Rich, D. P.; Anderson, M. P.; Gregory, R. J.; Cheng, S. H.; Paul, S.; Jefferson, D. M.; McCann, J. D.; Klinger, K. W.; Smith, A. E.; Welsh, M. J., Expression of cystic fibrosis transmembrane conductance regulator corrects defective chloride channel regulation in cystic fibrosis airway epithelial cells. *Nature* **1990,** *347*, 358-63.
(9) Zabner, J.; Couture, L. A.; Gregory, R. J.; Graham, S. M.; Smith, A. E.; Welsh, M. J., Adenovirus-mediated gene transfer transiently corrects the chloride transport defect in nasal epithelia of patients with cystic fibrosis. *Cell* **1993,** *75*, 207-16.
(10) Jabalameli, H.; Zahednasab, H.; Krimi, A.; Jabalameli, M., Zinc finger nuclease technology: Advances and obstacles in modelling and treating genetic disorders. *Gene* **2014,** *558*, 1-5.
(11) Scott, J. N.; Kupinski, A. P.; Boyes, J., Targeted genome regulation and modification using transcription activator-like effectors. *The FEBS journal* **2014,** *281*, 4583-97.
(12) Sander, J. D.; Joung, J. K., CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol* **2014,** *32*, 347-55.
(13) Cai, M.; Yang, Y., Targeted genome editing tools for disease modeling and gene therapy. *Current gene therapy* **2014,** *14*, 2-9.
(14) Rouet, P.; Smih, F.; Jasin, M., Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **1994,** *91*, 6064-8.
(15) Smih, F.; Rouet, P.; Romanienko, P. J.; Jasin, M., Double-strand breaks at the target locus stimulate gene targeting in embryonic stem cells. *Nucleic acids research* **1995,** *23*, 5012-9.
(16) Urnov, F. D.; Miller, J. C.; Lee, Y. L.; Beausejour, C. M.; Rock, J. M.; Augustus, S.; Jamieson, A. C.; Porteus, M. H.; Gregory, P. D.; Holmes, M. C., Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **2005,** *435*, 646-51.
(17) Kay, S.; Bonas, U., How Xanthomonas type III effectors manipulate the host plant. *Current opinion in microbiology* **2009,** *12*, 37-43.
(18) White, F. F.; Potnis, N.; Jones, J. B.; Koebnik, R., The type III effectors of Xanthomonas. *Molecular plant pathology* **2009,** *10*, 749-66.
(19) White, F. F.; Yang, B., Host and pathogen factors controlling the rice-Xanthomonas oryzae interaction. *Plant physiology* **2009,** *150*, 1677-86.
(20) Boch, J.; Bonas, U., Xanthomonas AvrBs3 family-type III effectors: discovery and function. *Annual review of phytopathology* **2010,** *48*, 419-36.
(21) 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* **2007,** *318*, 648-51.

(22) Boch, J.; Scholze, H.; Schornack, S.; Landgraf, A.; Hahn, S.; Kay, S.; Lahaye, T.; Nickstadt, A.; Bonas, U., Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **2009,** *326*, 1509-12.
(23) Moscou, M. J.; Bogdanove, A. J., A simple cipher governs DNA recognition by TAL effectors. *Science* **2009,** *326*, 1501.
(24) Bitinaite, J.; Wah, D. A.; Aggarwal, A. K.; Schildkraut, I., FokI dimerization is required for DNA cleavage. *Proceedings of the National Academy of Sciences of the United States of America* **1998,** *95*, 10570-5.
(25) Hopkins, C. M.; White, F. F.; Choi, S. H.; Guo, A.; Leach, J. E., Identification of a family of avirulence genes from *Xanthomonas oryzae* pv. *oryzae*. *Molecular plant-microbe interactions : MPMI* **1992,** *5*, 451-9.
(26) Szurek, B.; Rossier, O.; Hause, G.; Bonas, U., Type III-dependent translocation of the Xanthomonas AvrBs3 protein into the plant cell. *Molecular microbiology* **2002,** *46*, 13-23.
(27) Fulcher, M. L.; Gabriel, S. E.; Olsen, J. C.; Tatreau, J. R.; Gentzsch, M.; Livanos, E.; Saavedra, M. T.; Salmon, P.; Randell, S. H., Novel human bronchial epithelial cell lines for cystic fibrosis research. *American journal of physiology. Lung cellular and molecular physiology* **2009,** *296*, L82-91.
(28) Yankaskas, J. R.; Haizlip, J. E.; Conrad, M.; Koval, D.; Lazarowski, E.; Paradiso, A. M.; Rinehart, C. A., Jr.; Sarkadi, B.; Schlegel, R.; Boucher, R. C., Papilloma virus immortalized tracheal epithelial cells retain a well-differentiated phenotype. *The American journal of physiology* **1993,** *264*, C1219-30.
(29) Waldman, A. S.; Liskay, R. M., Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. *Molecular and cellular biology* **1988,** *8*, 5350-7.
(30) Fujitani, Y.; Yamamoto, K.; Kobayashi, I., Dependence of frequency of homologous recombination on the homology length. *Genetics* **1995,** *140*, 797-809.
(31) Chen, F.; Pruett-Miller, S. M.; Huang, Y.; Gjoka, M.; Duda, K.; Taunton, J.; Collingwood, T. N.; Frodin, M.; Davis, G. D., High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nature methods* **2011,** *8*, 753-5.
(32) Radecke, S.; Radecke, F.; Cathomen, T.; Schwarz, K., Zinc-finger nuclease-induced gene repair with oligodeoxynucleotides: wanted and unwanted target locus modifications. *Molecular therapy* **2010,** *18*, 743-53.
(33) Zhang, F.; Cong, L.; Lodato, S.; Kosuri, S.; Church, G. M.; Arlotta, P., Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **2011,** *29*, 149-53.
(34) Mussolino, C.; Morbitzer, R.; Lutge, F.; Dannemann, N.; Lahaye, T.; Cathomen, T., A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic acids research* **2011,** *39*, 9283-93.
(35) Miller, J. C.; Tan, S.; Qiao, G.; Barlow, K. A.; Wang, J.; Xia, D. F.; Meng, X.; Paschon, D. E.; Leung, E.; Hinkley, S. J.; Dulay, G. P.; Hua, K. L.; Ankoudinova, I.; Cost, G. J.; Urnov, F. D.; Zhang, H. S.; Holmes, M. C.; Zhang, L.; Gregory, P. D.; Rebar, E. J., A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **2011,** *29*, 143-8.
(36) Engler, C.; Kandzia, R.; Marillonnet, S., A one pot, one step, precision cloning method with high throughput capability. *PloS one* **2008,** *3*, e3647.

(37) Bibikova, M.; Carroll, D.; Segal, D. J.; Trautman, J. K.; Smith, J.; Kim, Y. G.; Chandrasegaran, S., Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Molecular and cellular biology* **2001,** *21*, 289-97.
(38) Handel, E. M.; Alwin, S.; Cathomen, T., Expanding or restricting the target site repertoire of zinc-finger nucleases: the inter-domain linker as a major determinant of target site selectivity. *Molecular therapy : the journal of the American Society of Gene Therapy* **2009,** *17*, 104-11.
(39) Sun, N.; Liang, J.; Abil, Z.; Zhao, H., Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Molecular bioSystems* **2012,** *8,* 1255-63.
(40) Lee, C. M.; Flynn, R.; Hollywood, J. A.; Scallan, M. F.; Harrison, P. T., Correction of the DeltaF508 Mutation in the cystic fibrosis transmembrane conductance regulator Gene by zinc-finger nuclease homology-directed repair. *BioResearch open access* **2012,** *1*, 99-108.
(41) Schwank, G.; Koo, B. K.; Sasselli, V.; Dekkers, J. F.; Heo, I.; Demircan, T.; Sasaki, N.; Boymans, S.; Cuppen, E.; van der Ent, C. K.; Nieuwenhuis, E. E.; Beekman, J. M.; Clevers, H., Functional repair of CFTR by CRISPR/Cas9 in intestinal stem cell organoids of cystic fibrosis patients. *Cell stem cell* **2013,** *13*, 653-8.

Chapter 3. Investigation of TALE DNA-Binding and Target Search Mechanism[2]

3.1 Introduction

The ability to engineer site-specific genome editing and modification tools is a powerful capability that paves the way for both important research and therapeutic applications. Transcription activator-like effectors (TALEs) have recently emerged as a scaffold of choice for engineering designer DNA-binding fusion proteins to be used as gene editing tools[1,2]. Naturally found in the bacterial pathogen *Xanthomonas*, TALE proteins are injected into host plant cells, where they transcriptionally activate genes required for bacterial colonization and spread[3,4]. TALE proteins are characterized by three conserved regions (Figure 1a): an N-terminal region (NTR) containing the type III translocation system required for secretion, the central repeat domain (CRD), and a C-terminal region (CTR) containing nuclear localization signals and an acidic activation domain[3,4]. The promoter specificity of the TALEs is conferred by the CRD, which harbors a series of tandem protein repeats that are typically 34 amino acids in length[3,5]. Sequence specific recognition is achieved by the repeat variable diresidues (RVD), amino acids at positions 12 and 13 of each repeat, that recognize one of the four nucleobases at each position of the target binding site[6,7]. Thus, this family of proteins is unique among other DNA-binding proteins in their 'single repeat to single base recognition mode'. It is therefore

[2] This work was done in collaboration with Dr. Charles M. Schroeder and Luke Cuculis. LC and ZA contributed equally to this work.

possible to design a TALE protein to recognize virtually any DNA sequence by simply reprogramming the RVD regions to bind the sequence of choice. Once the binding code of TALEs was revealed, fusion proteins constructed with TALEs as specific binding domains linked to nuclease domains (typically the dimeric nuclease FokI) were shown to function as highly specific designer nucleases. Following the first demonstration of TALEs as artificial nucleases[8,9], TALE proteins have also been used as artificial transcriptional activators[10,11,12], transcriptional repressors[11,13,14], demethylases[15], and fluorescent probes to study chromatin dynamics in live cells[16,17].

Understanding DNA recognition by TALEs will greatly improve the rational design of these proteins for biotechnological applications. A step towards rationalization of sequence-specific DNA recognition by TALEs has been achieved by solving the crystal structures of TALEs in both DNA-free and DNA-bound forms (Figure 1b)[18,19,20]. The repeats of the TALE CRD were shown to form a right-handed superhelical structure that wraps along the major groove of the B-form of DNA duplex. Each repeat comprises two α-helices that span residues 3-11 and 14(15)-33 and flank the loop region containing the RVDs. The residue at position 13 specifically interacts with a single base, while the residue at position 12 contributes to the stability of the RVD loop. The NTR also displays a right-handed superhelical structure with four continuous repeats that are strikingly similar in structure to those of the CRD[20]. In addition, comparison of free and DNA-bound forms of TALEs revealed that the protein undergoes a conformational change to a more compact

form, with the superhelical pitch being reduced from 60 Å in DNA-free form to 35 Å upon binding to target DNA[18,21].

While crystal structures of TALEs provide valuable insight as to the basis of TALE-DNA recognition, a central understanding of the molecular mechanism governing TALE sequence search is currently lacking. A critical need in the rational design of new, more efficient TALEN systems lies in the understanding of TALE binding dynamics and functions of TALE subdomains as they relate to target search. In this work, we utilize bulk binding assays as well as single molecule fluorescence microscopy in order to study the binding dynamics and sequence search mechanism of TALE proteins and TALE subdomains.

3.2 Results

3.2.1 Measurements of TALE-SCA2/DNA Binding Affinity in a Bulk Assay

To investigate the mechanism of TALE binding to DNA, we set to assay the difference in TALE binding to cognate and random DNA sequences. To ensure that the recombinant TALE we use in this study is functionally active, we utilized a TALE protein architecture previously reported for the correction of the human β–globin gene harboring a mutation associated with sickle cell disease[22]. The CRD of the protein, which contains 21.5 repeats, was engineered using the Golden Gate cloning method, and the flanking regions were truncated to 208 amino acids of NTR and 63 amino acids of CTR (Figure 3.1a and b). The TAL effector variable di-residues as well as the predicted binding site are provided in the Appendix A. The His-tagged

recombinant TALE protein was purified by affinity chromatography to homogeneity for subsequent *in vitro* binding assays.

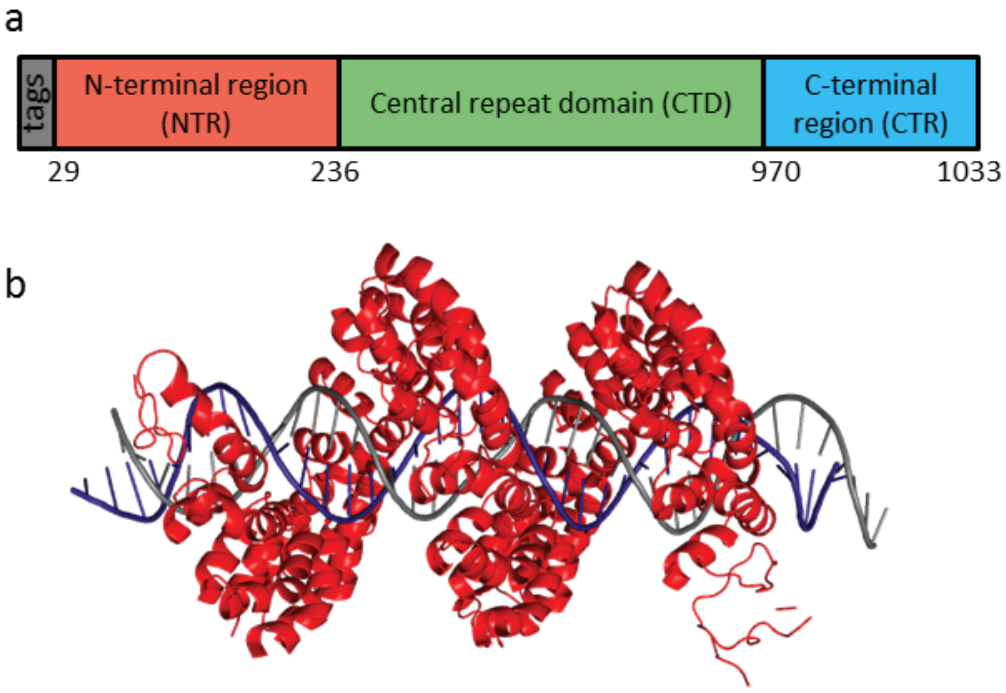


Figure 3.1 TALE protein structure and subdomain schematic. **(a)** General schematic of the TALE polypeptide chain. The N-terminal region (NTR) contains the type III translocation system necessary for secretion, the central repeat domain (CRD) contains the conserved 34 amino acid repeats for sequence-specific DNA binding, and the C-terminal region contains nuclear localization signals and an acidic activation domain. **(b)** Co-crystal structure of the PthXo1 TALE bound to its specific DNA target22, with only the CRD region shown for display. Cuculis *et al. Nat Commun* 2015, accepted

In vitro binding of the TALE to cognate and random DNA was investigated using fluorescence anisotropy. In our initial experiments, 50 bp-long fluorescein-labeled double-stranded DNA oligos were used as the ligand, and were designed to contain 22 nt sequence of interest flanked by unrelated 14 bp sequences. The 22 nt sequence of interest was either the cognate 22 nt sequence to the TALE ("50 bp

cognate sequence"), or an unrelated 22 nt sequence with the same sequence composition but with the order completely scrambled compared to the cognate sequence, ("50 bp random sequence"). These ds oligos were prepared by primer extension from two ss oligos that have an overlap region among them. One of these ss oligos was 5'-fluorescein-labeled for the fluorescence polarization measurements.

In order to investigate TALE binding to DNA, we sought to determine the conditions in which the greatest difference in binding affinities to cognate and random DNA would be observed. The initial buffer composition was 25 mM Tris-HCl pH7.5 buffer with 0.5 mM EDTA. We compared the effect of the additives including 0.1 mg/ml BSA, 1 mg/ml BSA, 6 mg/ml salmon sperm DNA (Figure 3.2), β-mercaptoethanol, and DTT (data not shown). In all cases, we did not observe significant difference between TALE binding to cognate and random DNA (Figure 3.2). We reasoned that high concentrations of salmon sperm DNA would screen nonspecific binding of TALE to DNA, thus allowing to observe a greater difference in binding affinities between cognate and random oligos in the presence of this additive. Salmon sperm DNA indeed significantly reduced the apparent binding affinity of TALE to ds oligos; however, it did not result in resolved specificity of TALE (Figure 3.2).

We reasoned that increasing ionic strength could also screen unspecific interactions of TALE with DNA, since the positively charged patches of the TALE CRD inner surface[18, 20] of the TALE solenoid would be screened from interacting with the negatively charged phosphate backbone of DNA. We therefore assayed TALE binding in various ionic strengths in the presence of 50-200 mM NaCl. Little

difference in binding affinity was observed for salt concentrations ranging between 0 and 100 mM, while a significant decrease in apparent binding affinity was observed in a buffer with 150 mM NaCl (Figure 3.3). In 200 mM NaCl, no binding was observed at 1 μ M protein concentration (data not shown). In neither of these conditions, however, did we observe the resolution of specific binding of the TALE.

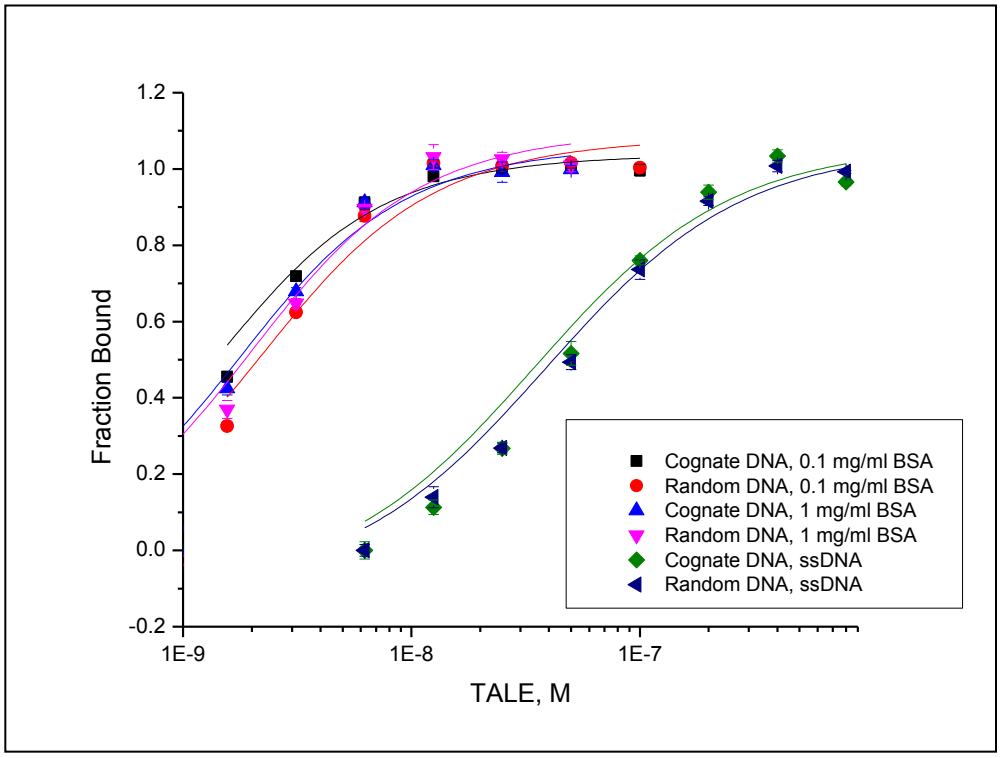


Figure 3.2 Fluorescence polarization assay of TALE21.5 binding to 50 bp-long ds oligodeoxynucleotide in 25 mM Tris-HCl pH7.5 buffer with different additives.

Since we found that the TALE binds to nonspecific DNA substantially in every condition that we tried, we reasoned that reduction of the oligo length should better resolve the slight difference in the apparent binding affinities between cognate and random DNA. We therefore reduced the ligand DNA length to 29 bp, with the 22 nt target sequence flanked by only 3 nt on each side (For the sequences of these oligos,

see Appendix A). As previously, we titrated the TALE with either cognate or random DNA in different NaCl concentrations. Interestingly, at 50 mM and 100 mM NaCl concentrations, we observed even tighter binding of TALE to 29 bp oligos compared with 50 bp oligos (Figure 3.4). We also observed a small, yet significant difference (~2-fold difference in dissociation constant K_D) in TALE binding affinities to cognate versus random DNA at 100 mM and 150 mM NaCl concentrations.

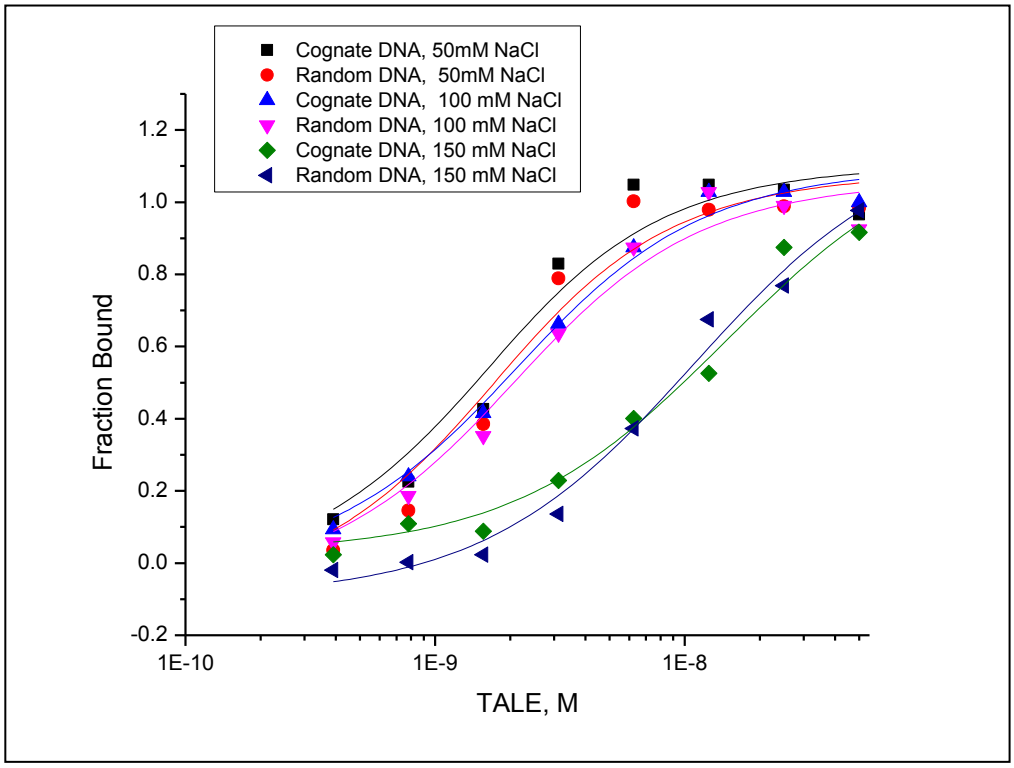


Figure 3.3 Fluorescence polarization assay of TALE21.5 binding to 50 bp-long ds oligodeoxynucleotide in 25 mM Tris-HCl pH7.5 at different NaCl concentrations.

Several studies have demonstrated the ability of TALEs to bind nonspecific DNA with varying levels of affinity[20,23,24]. However, the nonspecific binding affinity of TALE that was observed in our hands was unexpectedly high, and the binding affinities to random versus cognate DNA were surprisingly similar. It is particularly

perplexing given the enormity of the nonspecific sequence volume which the TALEs naturally have to screen in search of the few binding sites in plant genomes. These observations might indicate that the mechanism by which TALEs recognize the target DNA is governed by other factors than just DNA association and dissociation rates. As proposed previously, TALEs could bind DNA nonspecifically and search the target sites by 1D diffusion[20]. Hence, it is plausible that the residence times of the protein on the target site, as compared with other sites, is imperative for the TALE activity.

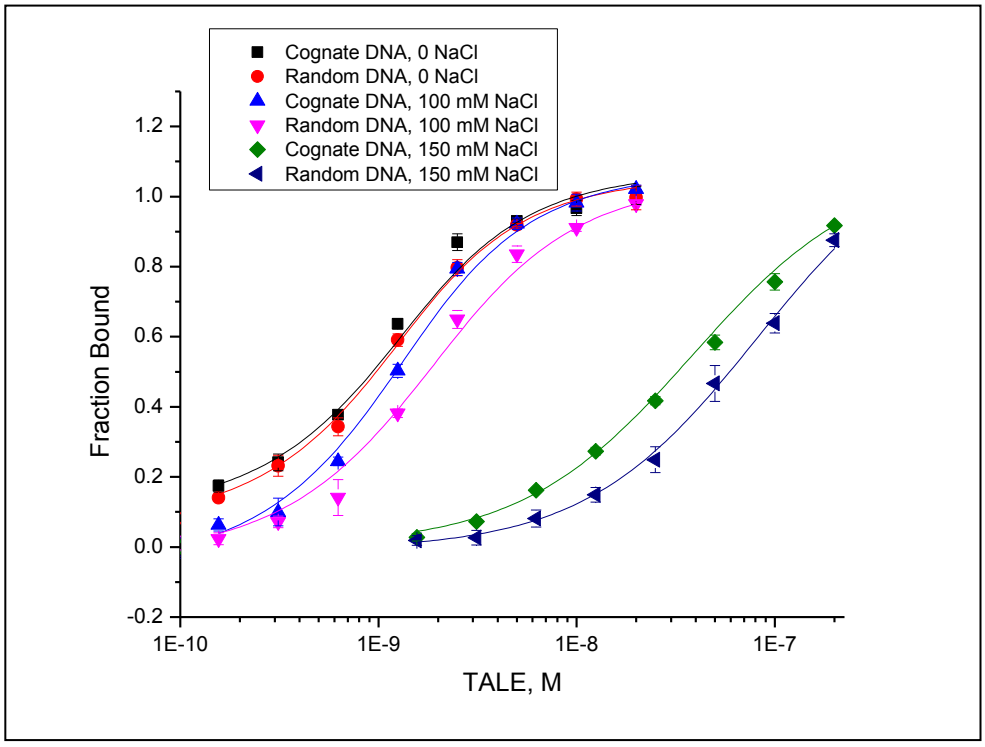


Figure 3.4 Fluorescence polarization assay of TALE21.5 binding to 29 bp-long ds oligodeoxynucleotide in Tris-HCl pH7.5 buffer at different salt concentrations.

3.2.2 Optimization of Protein and DNA Manipulation for Single Molecule Imaging

Although predicted[20], the 1D target search mechanism of TALEs has not yet been demonstrated. In order to directly observe the binding dynamics and sequence search mechanism of TALE proteins, we utilized single molecule imaging microscopy. To fluorescently label the TALE tSCA2, a bioorthogonal aldehyde labeling scheme[25-26] was employed, which allowed the site-specific and nonperturbative attachment of a hydrazide functionalized Cy3 organic dye. To record the binding dynamics of single TALEs, we utilized dual-tethered DNA substrates illuminated via total internal reflection (TIRF) in milli-fluidic flow cells.

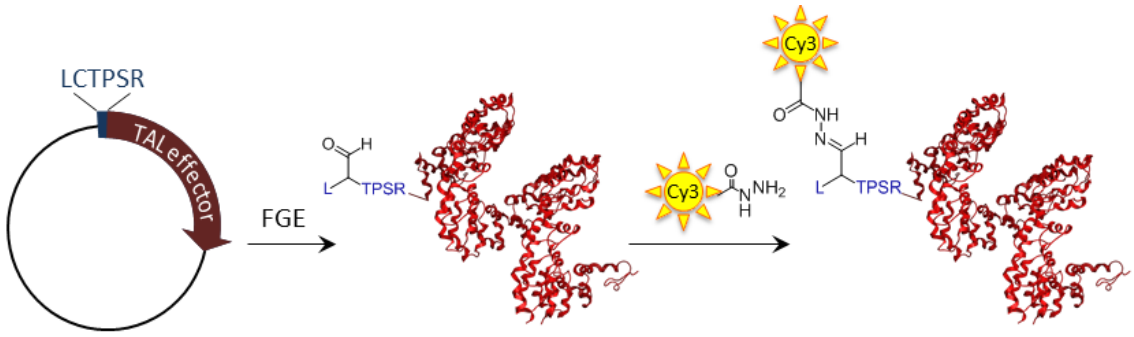


Figure 3.5 Aldehyde labeling scheme used to achieve non-perturbative one-to-one dye-to-protein stoichiometry. A six-amino acid motif (LCTPSR) is cloned into the N-terminus of the TALE protein and the construct is co-expressed with formylglycine generating enzyme (FGE) which then converts the cysteine to a formylglycine bearing an unnatural aldehyde. Aldehyde specific conjugation via hydrazide-functionalized Cy3 organic dyes then allows for non-perturbative, site-specific labeling.

Cuculis *et al. Nat Commun* 2015, accepted

To label the protein, an aldehyde tag (hexapeptide LCTPSR) was cloned upstream of the N-terminal His_6 tag and co-expressed with formylglycine-generating enzyme (FGE), which recognizes the 6-amino-acid motif and oxidizes

cysteine to formylglycine *in vivo*[25] (Figure 3.5). To this end, the BL21 (DE3) strain of *E. coli* was co-transformed with arabinose-inducible pBAD-FGE and IPTG-inducible pET28-Naldt-tSCA plasmids. Induction of both or either of the proteins did not result in noticeable TALE or FGE expression, as observed by SDS-PAGE gels (Figure 3.6a). Nevertheless, purification of the TALE protein from 500 ml of induced culture using affinity chromatography yielded TALE protein in > 90 % purity (Figure 3.6b).

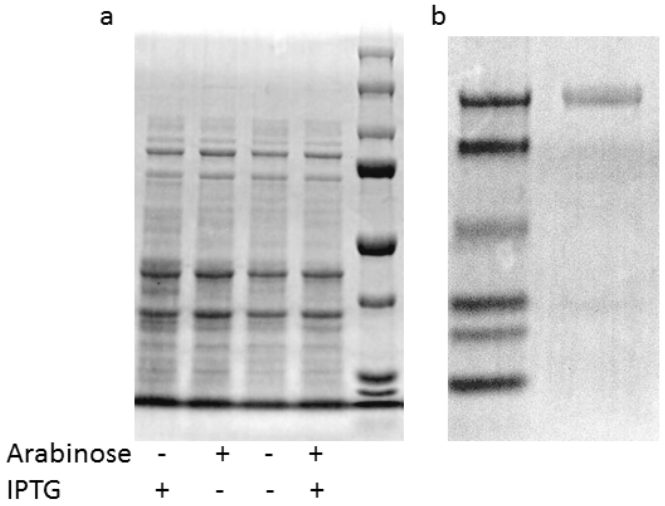


Figure 3.6 SDS-PAGE analysis of **(a)** aldehyde-tagged TALE expressed in BL21 (DE3) and **(b)** purified using affinity chromatography.

Analysis of the purified protein using LC-MS/MS indicated that > 82.7 % of the cysteines in the LCTPSR hexapeptide were modified to formylglycine (Figure 3.7). The aldehyde-tagged protein was next site-specifically modified with Cy3-hydrazide. After incubation of the protein with Cy3-hydrazide, the protein was purified from unreacted organic dye using micro bio-spin size exclusion columns (Bio-Rad). A time course of TALE labeling showed that the reaction starts to plateau at room temperature after approximately 48 hours of incubation, reaching the

efficiency 60 % of labeled protein at 48 hours of incubation (Figure 3.8a), as measured by absorption at 548 nm (Figure 3.8b) and Bradford assay. To check if unreacted Cy3 was still present in the solution, we imaged fluorescence of TALE ran through a SDS-PAGE gel, and observed that there was no significant amount of unreacted Cy3 (Figure 3.8c).

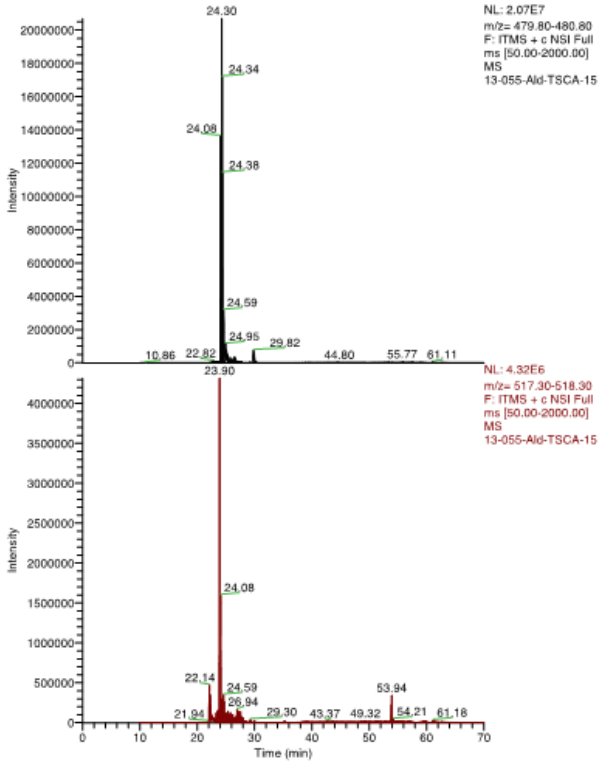


Figure 3.7 LC-MS/MS analysis of aldehyde-tagged TALE21.5. The cysteines of the purified protein were protected by alkylation with iodoacetamide and trypsin digested prior to LC-MS/MS. The percentage of the peak counts of peptides containing formylglycine ($2.07*10^7$) among all peptide peak counts containing alkylated cysteine ($2.50x10^7$) was calculated to be 82.8 %, which we assume to correspond to the aldehyde labeling efficiency.

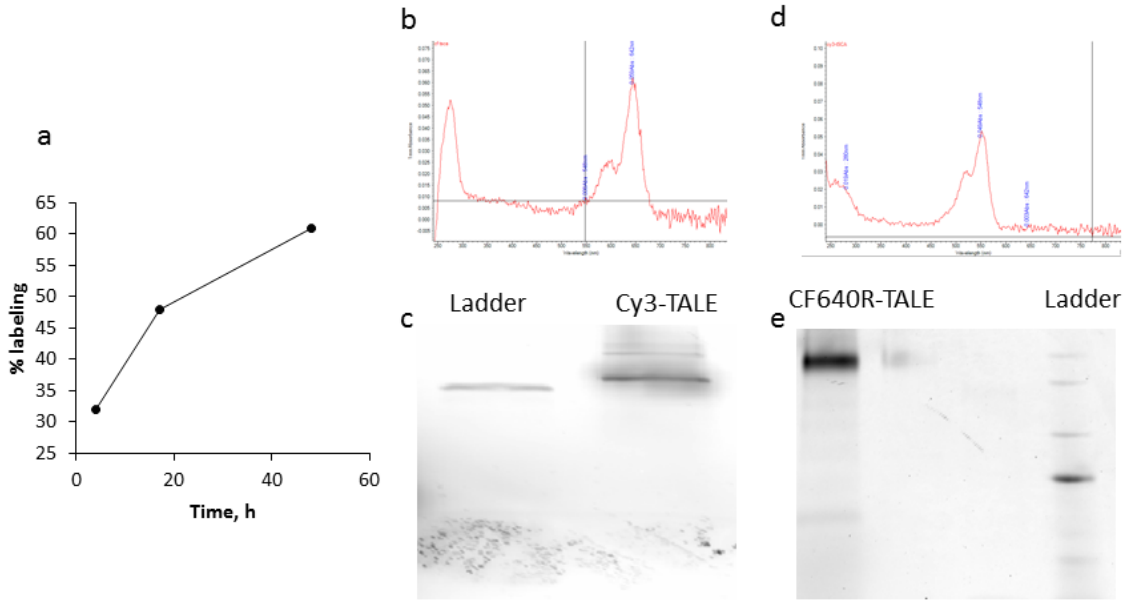


Figure 3.8 Labeling of TALE21.5 with Cy3 and CF640R. **(a)** Time course of labeling of Nald-TALE with Cy3-Hz at room temperature. **(b)** UV/Vis spectroscopy of Cy3-labeled TALE21.5. **(c)** Fluorescence imaging of Cy3-labeled TALE21.5 ran on an SDS-PAGE gel. **(d)** UV/Vis spectroscopy of CF640R-labeled TALE21.5. (**e)** Fluorescence imaging of CF640R-labeled TALE21.5.

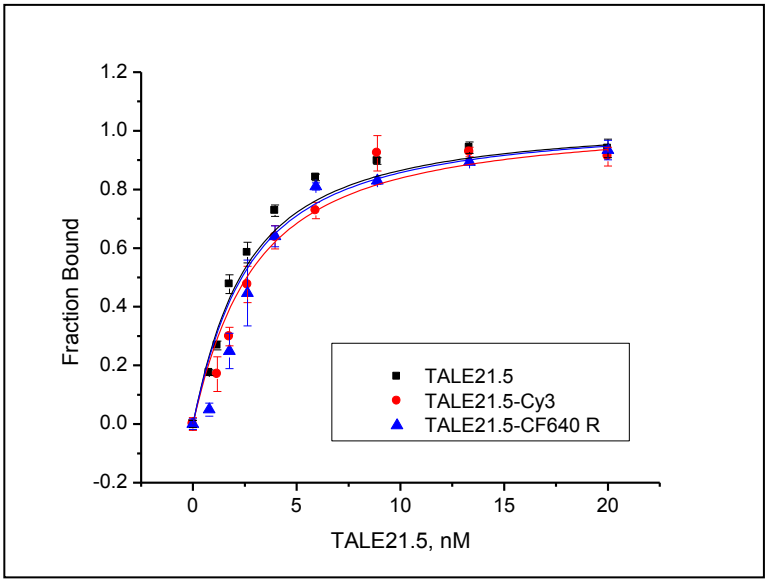


Figure 3.9 TALE21.5 labeled with different fluorophores binding to cognate DNA oligomer. Fluorescence polarization was performed with unlabeled, Cy3-labeled, and CF640 R-labeled TALE21.5.

Cuculis *et al. Nat Commun* 2015, accepted

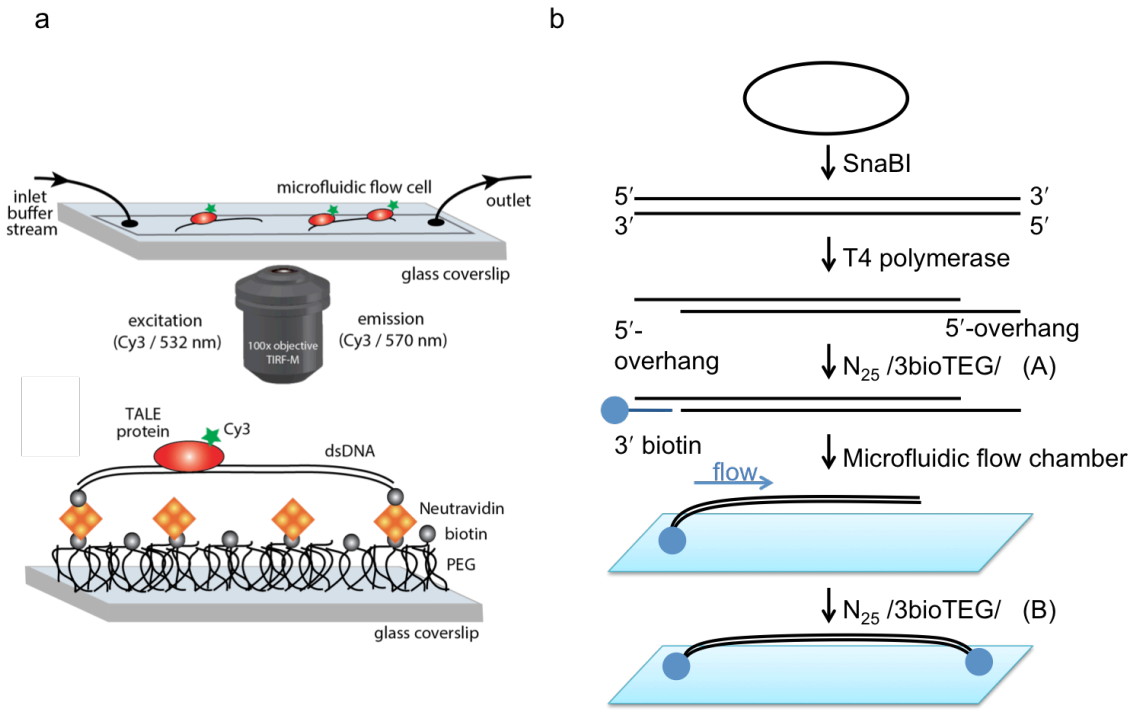


Figure 3.10 Schematic of the single molecule imaging microscopy set up for the tracking of TALE protein dynamics. **(a)** Schematic of experimental setup showing microfluidic flow cell and microscope objective lens for TIRF-M. **(b)** Schematic of dual-tethering of a DNA substrate. The 45 kb plasmid is linearized with the *Sna*BI restriction enzyme. The 3′-5′ exonuclease activity of T4 DNA polymerase is used to create the 5′ overhangs at both ends of the linearized DNA. The exposed 5′-overhangs are used to sequence-specifically anneal 3′-biotinylated oligonucleotides. First, oligonucleotide A is annealed to the substrate DNA. The substrate DNA is attached to the surface of the slide of the microfluidic flow chamber through the biotin-neutravidin interactions. Oligonucleotide B is flown through the chamber and is annealed to the second 5′-overhang, which results in the attachment of the second end of the substrate DNA.

Cuculis *et al. Nat Commun* 2015, accepted

In order to determine if the labeling affected the tSCA2 binding to DNA, we performed fluorescence polarization assay as previously, but did not observe

significant difference in binding affinity (Figure 3.9). To additionally check for Cy3 interference with the TALE binding activity, we also similarly labeled the TALE with CF640 R-hydrazide (Figure 3.8d and e), and similarly did not observe significant difference in binding to cognate DNA by fluorescence anisotropy (Figure 3.9).

For the construction of the milli-fluidic flow cells, we functionalized coverslips consecutively with PEG/PEG-biotin and neutravidin for surface attachment of DNA and reduction of nonspecific protein adsorption (Figure 3.10a). As a DNA substrate on which the TALE movement would be tracked, we chose a 44,898 bp plasmid in our lab that contains a neoaureothin synthesis pathway from *Streptomyces orinoci*. This plasmid does not contain binding sites for the engineered TALE, and contains a single *Sna*BI digestion site, which was used to linearize the plasmid. The ssDNA overhangs on both ends of the linearized plasmid were created by the 3'-5' exonuclease activity of T4 polymerase at room temperature in the absence of dNTPs, as described in SLIC cloning[27]. The DNA was double-tethered to the bottom of the flow cell via neutravidin-biotin interactions. For this purpose, DNA was biotinylated with 3'-biotinylated oligos complementary to the ssDNA overhangs on both ends of the substrate DNA, consecutively (Figure 3.10b). First, one oligo would be annealed to the substrate DNA, after which the DNA would be attached to the flow cell under flow. Second, the remaining oligo complementary to the second ssDNA overhang would be added to the flow cell, thus attaching the second end of the substrate DNA. The labeled TALE protein was passed through the flow cell last, and the single molecule TALE binding events on stretched DNA were imaged in the absence of the flow (Figure 3.11a and b). Note that all the experiments on single

molecule spectroscopy and corresponding data analysis were carried out by Luke Cuculis.

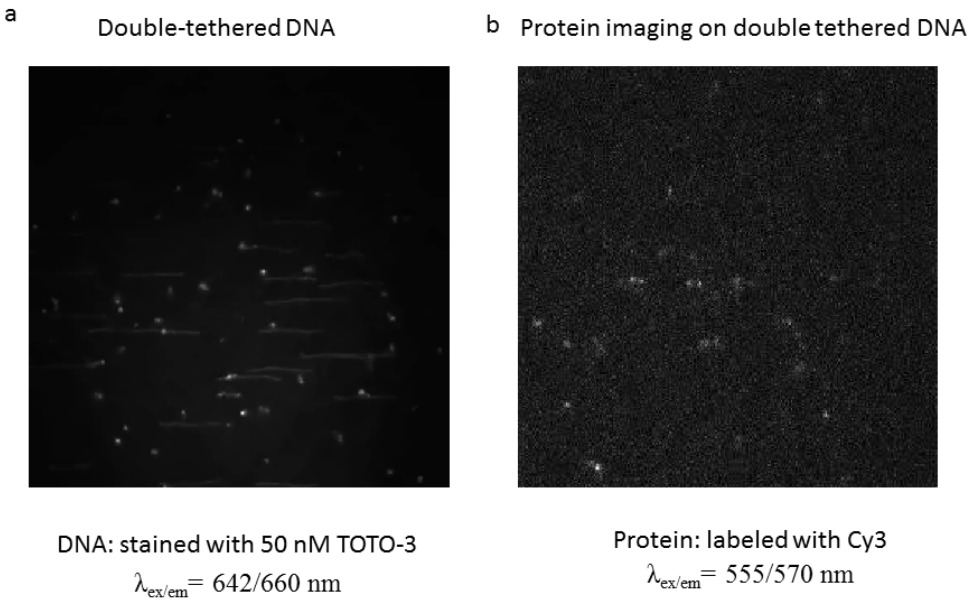


Figure 3.11 Visualization of single molecules of TALE and DNA substrate**. (a)** Double-tethered DNA substrate molecules were stained with SYTOX Green and excited using a 488 nm DPSS laser **(b)** Cy3-labeled TALE21.5 was excited using 532 nm diode-pumped solid state (DPSS) laser

3.2.3 One dimensional diffusion of TALE along DNA templates

The positions of single TALE molecules were recorded and tracked by fitting a two dimensional Gaussian point spread function to the diffraction-limited spots. We were able to directly observe TALE molecules diffusing one dimensionally along stretched DNA substrates (Figure 3.12a). The 1D mean-squared displacement (MSD) of the diffusing TALES increased linearly with time (Figure 3.12b), which is indicative of a Brownian motion. These observations corroborate the hypothesis that TALE proteins search for their targets via facilitated diffusion, where 1D

movement of TALEs along random sequences of DNA shortens the target search time.

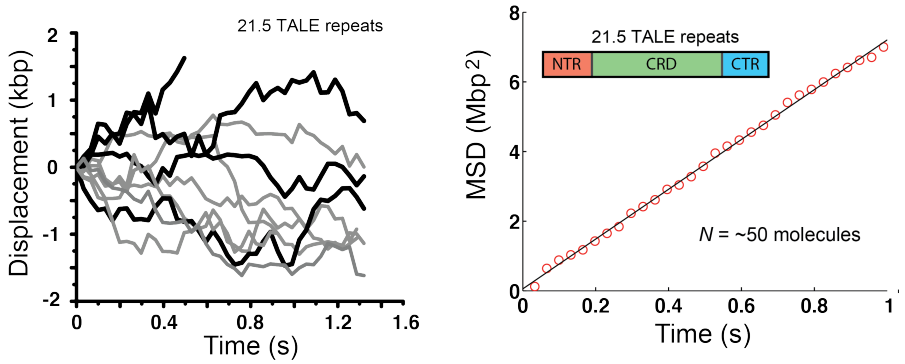


Figure 3.12 TALE proteins exhibit rapid 1-D diffusion along DNA templates. **(a)** Single molecule trajectories of TALE diffusion events over short time scales (~1 sec) at 90 mM KCl. **(b)** Mean-squared displacement of an ensemble of TALE proteins diffusing along non-specific DNA templates at 90 mM KCl.

Cuculis *et al. Nat Commun* 2015, accepted

We next sought to determine if the observed 1D diffusion is primarily sliding, where the protein remains in constant contact with the DNA, or hopping, where the protein briefly dissociates and re-associates at a distance shorter than the distance of persistence length of the DNA. To this end, we compared TALE diffusion rates along substrate DNA under different salt concentrations. Since an increase in ionic strength will stabilize the protein in solution, it will increase the energetic barrier to rebinding to DNA, and thus increase the fraction of the time the protein is unbound from DNA and mobile. Hopping diffusion is predicted to be faster than sliding diffusion due to reduced friction. Hence, the hopping behavior should correlate with increased apparent 1D diffusion speed at increased salt concentrations. Using the covariance based estimator (CVE) analysis[28], the apparent 1D diffusion coefficients

were determined to be $1.5*10^6$ bp^2/s and $5.9*10^6$ bp^2/s at 30 mM and 90 mM KCl concentrations, respectively (Figure 3.13a and b). Thus, increasing KCl concentration resulted in a nearly four-fold increase in 1D diffusion speed, which implies that the TALEs diffuse along their DNA substrates at least in part by a hopping mechanism.

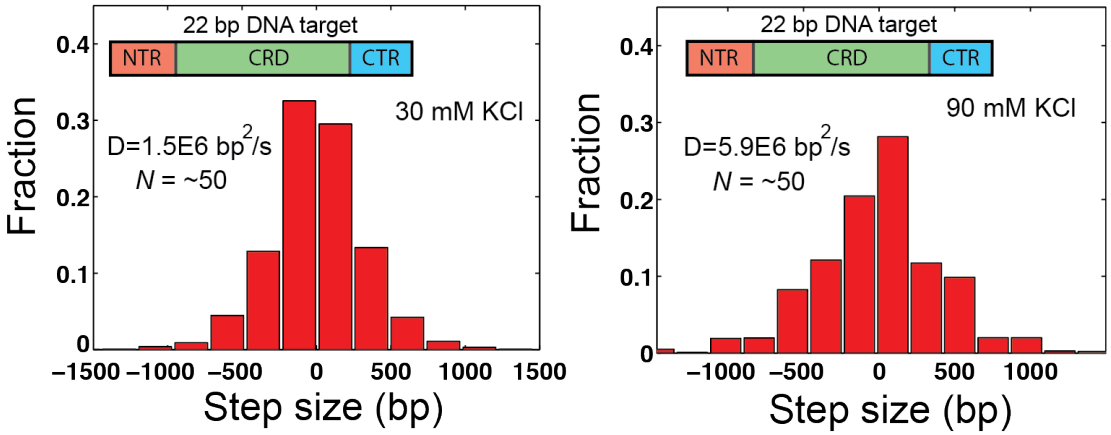


Figure 3.13 TALE proteins exhibit rapid 1-D diffusion along DNA templates. Histograms of TALE step sizes along non-specific DNA at 30 **(a)** and 90 mM KCl **(b)**, respectively.

Cuculis *et al. Nat Commun* 2015, accepted

Analysis of long (12-15 seconds) trajectories of TALE molecules along DNA substrates revealed at least two modes diffusion: rapid diffusion alternating with periods of relatively stagnant behavior, where diffusion slows and the protein remains constrained to a comparatively short distance (Figure 3.14a). We next analyzed the binding times of TALE proteins at 30 mM and 80 mM KCl concentrations. The binding times of the protein at both concentrations fit a double-exponential function (Figure 3.14b and inset), which is another indication of multiple modes of binding. These data imply that TALE proteins assume at least two

modes of DNA binding: a "search" mode, in which the protein diffuses rapidly in search of the target site, and a "binding" mode, in which the protein assumes the conformation described in the co-crystal structures with cognate DNA sequence[18], and is more restricted in diffusive ability.

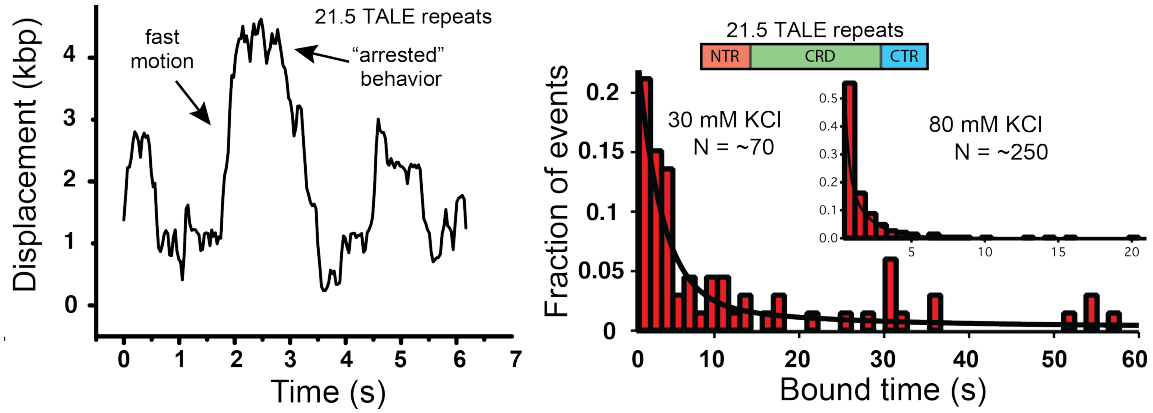


Figure 3.14 TALE proteins exhibit rapid 1-D diffusion along DNA templates. **(a)** Single molecule trajectory at 90 mM KCl over long time scales (>5 sec) shows periods of rapid diffusion interspersed with periods of arrested motion. **(b)** Distribution of characteristic TALE bound times at 30 mM KCl and (inset) 80 mM KCl with corresponding double-exponential fits.

Cuculis *et al. Nat Commun* 2015, accepted

3.2.4 1D-diffusion of TALE truncations along DNA templates

Although TALE specificity towards DNA is characterized by its CRD, it was consistently shown that the CRD alone is insufficient for the activity of the fusion proteins *in vivo*[8, 29-31], and that the NTR of TALEs immediately preceding the CRD was shown to be required for the activity of TALE fusions with other proteins[10, 29-31]. Importantly, a crystal structure revealed that the NTR immediately preceding the CRD contains four continuous repeats with a striking resemblance to the CRD repeat structure[20]. Moreover, the NTR alone was found to bind to random DNA sequences, although with lower affinity compared with the full length TALE, whereas the CRD

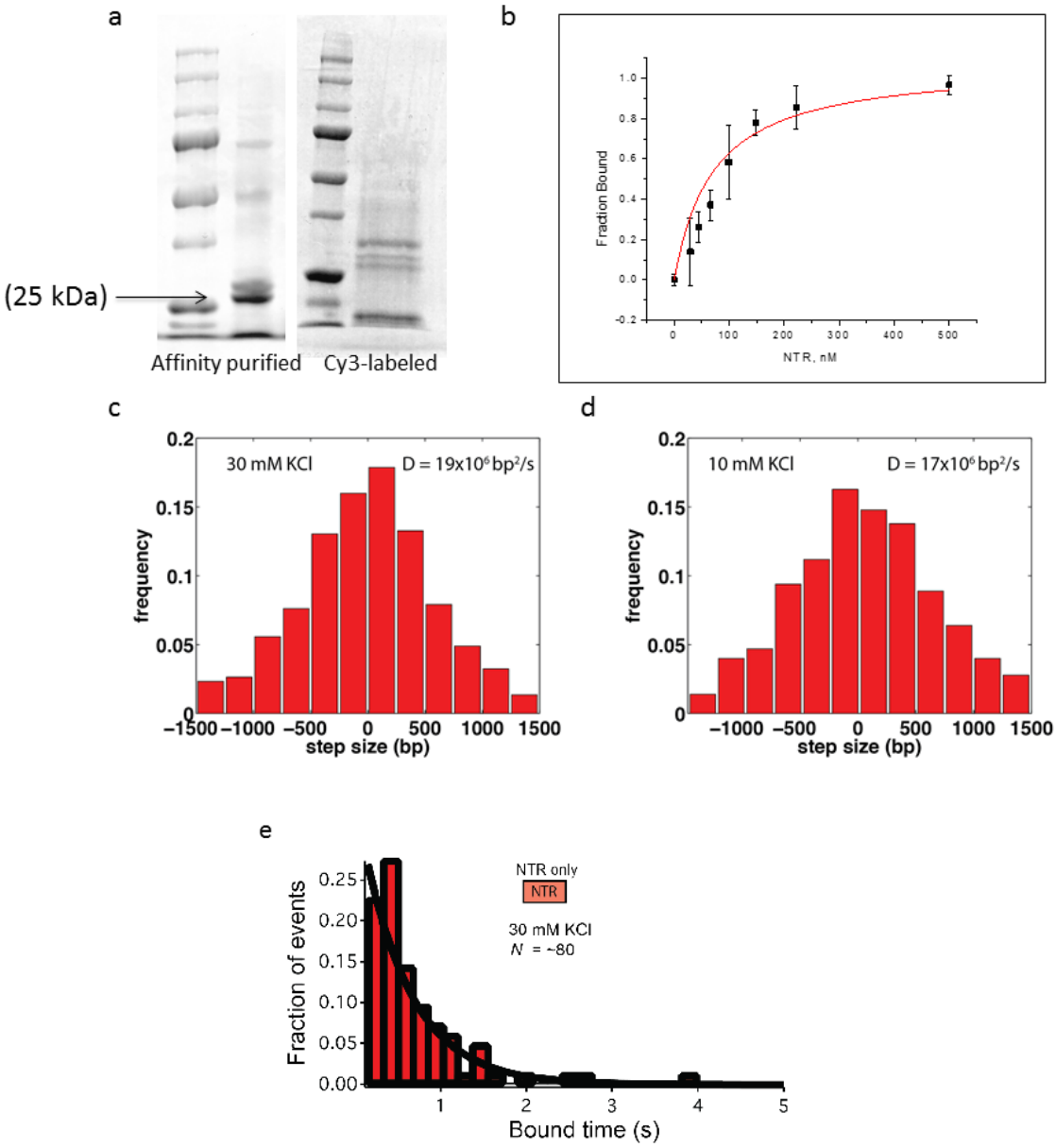


Figure 3.15 Bulk and single molecule analysis of NTR binding to random DNA. **(a)** Left panel: SDS-PAGE analysis of NTR purified using affinity chromatography. Right panel: SDS-PAGE analysis of Cy3-labeled NTR. **(b)** Fluorescence polarization aldehyde-tagged, unlabeled NTR and random DNA oligomer. **(c)** Histogram of step size distribution for NTR diffusion along DNA at 30 mM KCl **(d)** Histogram of step size distribution for NTR diffusion along DNA at 10 mM KCl. **(e)** Distribution of bound times for NTR truncation mutant on DNA templates, which is described by a single exponential decay.

Cuculis *et al. Nat Commun* 2015, accepted

alone was found not to bind DNA[20]. These findings suggest that the NTR serves as a nucleation site for TALE binding to DNA, allowing sliding of the TALE along DNA in search of the target sequence[20]. In this case, the NTR alone should be able to bind DNA and move along the DNA substrates. We therefore truncated the TALE protein sequence to NTR alone and studied its behavior on double-tethered DNA substrates.

The 208 amino acid-long NTR was aldehyde-tagged and purified using affinity chromatography. The ~25 kDa NTR was co-purified with proteins with the sizes of 50 and 75 kDa, as shown by SDS-PAGE analysis (Figure 3.15a left panel). Gel filtration allowed resolution of the three protein sizes into separate fractions (data not shown). However, the purified smallest size fraction again revealed three sizes of protein on SDS-PAGE (data not shown), which suggests that the observed three protein sizes are dimers and trimers of the purified NTR. Surprisingly, cy3-labeled NTR (~90 % labeled, as measured by absorbance at 548 nm in the Bradford assay), showed a much weaker tendency to aggregate into higher molecular structures (Figure 3.15 a right panel). Using fluorescence anisotropy, we found that unlabeled NTR binds to random DNA, although with reduced affinity compared with "full-length" TALE (Figure 3.15b, Table 3.1), which is consistent with available literature[20]. This reduced apparent affinity of NTR towards random DNA could at least partially be explained by the instability of NTR in solution due to high propensity of NTR towards aggregation.

Construct	K_d with cognate target, nM	K_d with random target, nM
TALE21.5	1.5 ± 0.3	2.3 ± 0.3
TALE15.5	16.7 ± 0.2	19.3 ± 0.3
TALE11.5	11.1 ± 2.0	13.2 ± 3.4
NTR	ND	70 ± 27

Table 3.1 Binding affinities of TALE truncations with cognate and random DNA oligomers, measured via fluorescence polarization assay. Dissociation constants, K_d, are represented as average ± STD (n=3-5). ND, not determined.

Cuculis *et al. Nat Commun* 2015, accepted

Via single-molecule imaging, one-dimensional diffusion of NTR on double-tethered DNA substrates was indeed observed. Analysis of the diffusion of single molecules of protein at 30 mM KCl concentration revealed a 1D diffusion coefficient of $19*10^6$ bp^2/s (Figure 3.15c), which is significantly higher than the diffusion coefficient of the "full length" TALE at the same salt concentration. We next sought to determine if the NTR translocates along the DNA by a sliding or hopping mechanism, as mentioned above. Since increasing the salt concentration to 90 mM decreased the binding time of NTR to an extent that an accurate diffusion coefficient could not be determined (data not shown), we compared the diffusion coefficient of NTR at 30 mM KCl to that at 10 mM KCl. The diffusion coefficient did not change significantly (Figure 3.15d), which may imply that the NTR moves along DNA primarily by a sliding mechanism, in contrast to the "full-length" TALE. Importantly, binding times of NTR were described by a single-exponential function (Figure 3.15e), which suggests that the NTR truncation exhibits only one mode of DNA binding.

We also generated a fluorescently labeled TALE truncation that lacked the NTR, containing only the CRD and CTR regions. Consistent with previous reports[20], the CRD-CTR truncation of TALE did not bind DNA in neither bulk or single molecule assays, further corroborating the importance of NTR as a nucleation site in TALE-DNA binding.

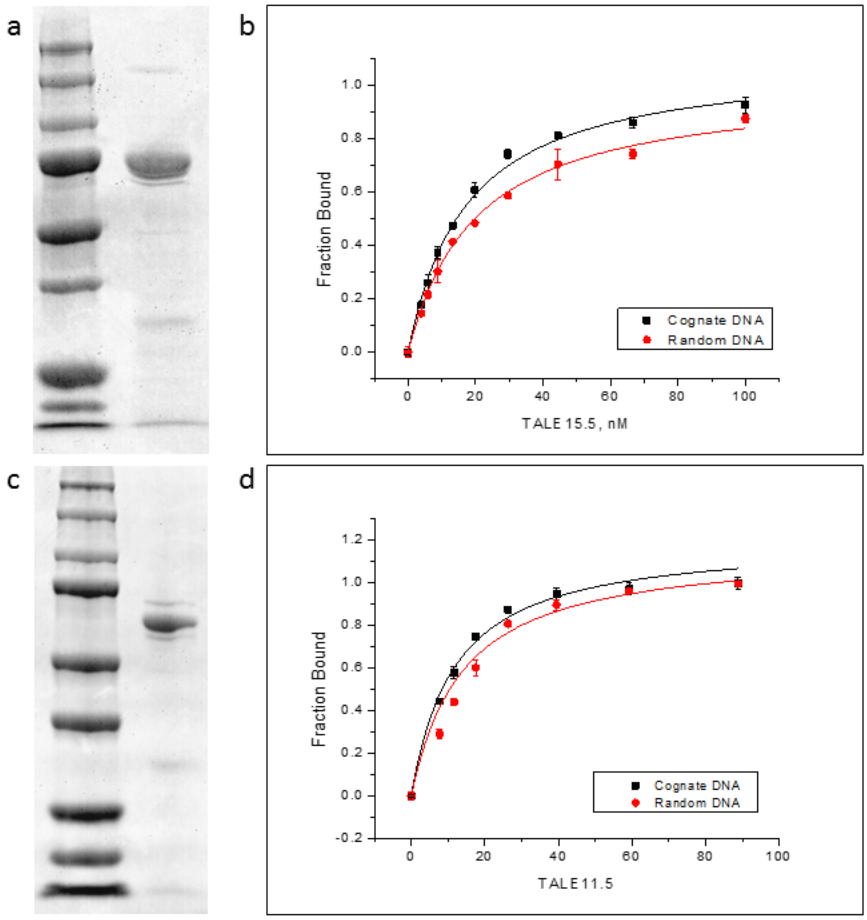


Figure 3.16 Bulk analyses of truncated TALEs. **(a)** and **(c)** SDS-PAGE analysis of affinity-purified TALE15.5 and TALE10.5, respectively. **(b)** and **(d)** Fluorescence anisotropy assay of TALE15.5 and TALE 11.5 binding to cognate and random DNA oligomers. Fluorescence polarization was performed with aldehyde tagged, fluorophore-free TALEs. Fluorescein labeled 29-mer DNA oligomers were individually assayed with the proteins.

Cuculis *et al. Nat Commun* 2015, accepted

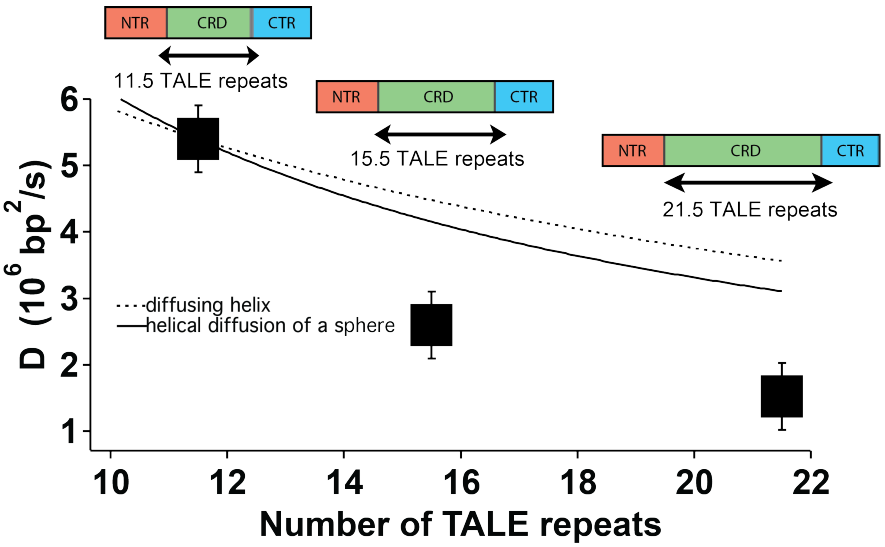


Figure 3.17 Effect of CRD size on 1-D diffusion coefficient of TALE proteins. TALEs with 11.5, 15.5, and 21.5 central repeats were generated and their apparent 1-D diffusion coefficients were determined at 30 mM KCl (square symbols). Experimental data is compared to two different hydrodynamic models of diffusion that consider protein conformation as a diffusing helix and a spherical protein tracking the DNA helix, shown by the dashed and solid lines, respectively.

Cuculis *et al. Nat Commun* 2015, accepted

Intrigued by this interplay of different diffusion modes of full-length TALE and the NTR, we further studied TALEs with fewer repeats compared to the 21.5-repeat "full-length" TALE. We predicted that these proteins with shorter CRD but with the NTR unchanged would demonstrate a behavior in between those observed for the "full-length" or NTR. We purified, labeled, and measured the bulk DNA-binding affinities of TALEs with the first 10.5 and 15.5 repeats (Figure 3.16a-d, Table 3.1). The binding affinities of these truncations to both random and cognate DNA were reduced compared to the "full-length" TALE, as expected. The diffusion coefficients also increased from 21.5-repeat TALE to 15.5- and 10.5-repeat TALEs (Figure 3.17). However, this increase could not be described solely by an additional hydrodynamic

drag given by the additional repeats as modeled for a rigid helical body diffusing one dimensionally[32] (Figure 3.17, dashed line). Hence, the extent of the diffusion coefficient decrease from TALES with 10.5 repeats to TALEs with 21.5 repeats is governed by more factors than simply the length of the TALE protein superhelix. We suggest that at least partially this change in the diffusion speeds is due to an altered equilibrium of the two modes of DNA binding among the TALEs with different lengths of the CRD.

3.3 Discussion

The ability of DNA-binding proteins (DBP) such as *lac* repressor to locate target sites orders of magnitude faster than that predicted by 3D diffusion has long been noted[33]. Resolution to this conundrum was offered by a now widely accepted concept of facilitated diffusion, where non-specific association of proteins to DNA can reduce the amount of search time of the target sequence[33-37]. It was later proposed that a most efficient search mechanism would be a combination of one-dimensional diffusion along DNA and three-dimensional diffusion, assuming that the search is conducted by just one protein molecule[38-39]. Nevertheless, whether in physiological conditions DBPs localize their targets primarily by 1-D diffusion, 3-D diffusion, or a combination thereof remains an open question. *Escherichia coli* RNA polymerase, for example, was reported to search for its target promoter primarily by a 3-D diffusion at physiological concentrations of the protein[40]. Similarly, the CRISPR RNA-guided endonuclease Cas9 was also reported to find targets primarily by 3-D diffusion until it finds the short trinucleotide sequence PAM required for Cas9 activity[41]. It is possible that different DBPs utilize different target search

mechanisms, which suggests that different classes of DBPs should be investigated on a case-by-case basis.

Given that TALE proteins stand out among other DBPs in their unique structure and DNA recognition mechanism, their mechanism of target search is particularly obscure. In this work, we investigated the DNA-binding dynamics and the target search mechanism of a TALE protein and the contribution of its subdomains in these processes. For the first time, we were able to observe single molecules of TALE moving by a 1-D diffusion on a double-tethered DNA substrate, which implies that TALEs can in principle utilize facilitated diffusion mechanism of target search. Whether TALE proteins search for their target sites primarily by 1-D diffusion or also rely heavily on 3-D diffusion in physiological conditions remains to be investigated. However, the conclusions from the *in vitro* experiments reported here encompass an important step in the elucidation of the complete mechanism.

The finding that the NTR demonstrates only a sliding behavior, whereas the "full-length" TALE exhibits a combination of hopping and sliding indicates that different domains of the TALE are responsible for distinct modes of motion. Thus, it could be speculated that the positively charged patches on the surface of NTR[20] are responsible for the initial interaction of the protein with the phosphate backbone of DNA, and subsequently are responsible for the continuous interaction with the phosphate backbone, allowing the movement of the full-length TALE by a sliding mechanism. The CRD, on the other hand, could be interacting with DNA with a much rougher energy landscape, thus destabilizing the complex and forcing the protein to dissociate more often compared with the NTR alone. The hopping behavior arising

as a result can be an advantageous translocation mechanism that would allow bypassing of a bound protein or sampling both strands of the DNA substrate[42].

Most DBPs, including TALEs, bind their target DNA sequences with large binding energies. For a rigid protein, this would cause the energy landscape of a sliding protein on nonspecific DNA so rough, that the target search time would be unreasonably slow, a phenomenon known as the search-speed stability paradox[43-44]. A solution to the paradox proposes a DBP capable of switching between two conformations separated by a free energy barrier[43-44]. In a "searching mode", the protein is loosely bound to DNA, and can rapidly slide along DNA due to a relatively smooth binding energy landscape. In a "recognition mode", the protein is bound to DNA more tightly and can be trapped in a deep energy well, which would reduce its diffusion speed dramatically. In this work, we have demonstrated that the TALE protein indeed exhibits a two-mode binding behavior.

Although no direct evidence of two conformations of TALEs bound to DNA substrates exists, conformational plasticity of TALEs in the DNA-free form as well as an almost two-fold decrease in the superhelical pitch of TALEs upon specific binding are widely reported both experimentally and computationally[18, 21, 45-47]. Thus, we suggest that the TALE conformation corresponding to the recognition mode closely resembles that of the observed in the co-crystal structure[18-19], while the TALE conformation corresponding to the searching mode has a larger superhelical pitch, either resembling that of the DNA-free structure of TALE[18], or anywhere in between (Figure 3.18).

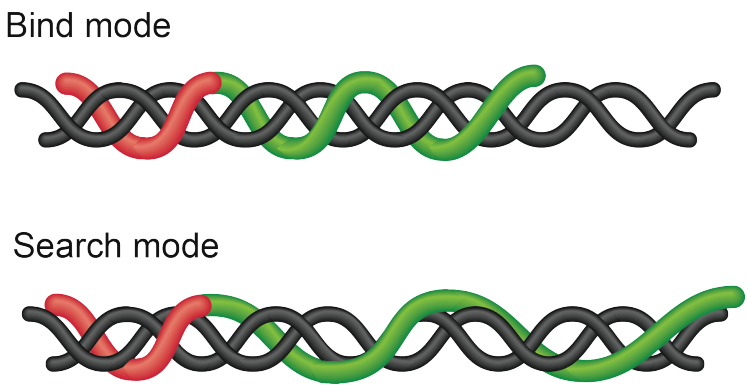


Figure 3.18 Schematic of two-state model for TALE protein diffusion along DNA. A TALE is in equilibrium between the 'searching mode' and 'recognition modes'. In the 'recognition mode', the TALE protein is tightly bound to the major groove of a DNA via electrostatic interactions with the phosphate backbone. In the 'searching mode', the protein assumes a conformation with a larger superhelical pitch, analogous to the DNA-free structures of TALEs.

Cuculis *et al. Nat Commun* 2015, accepted

It is agreed that a substantial contribution to nonspecific binding of TALEs to DNA is made by the positively charged patches along the inner surface of the TALE CRD[18, 20]. In fact, the consensus among crystallographic and computational studies is that residues G14, K16, and Q17 account for the most of the overall binding energy of TALE to DNA[18-19, 45, 47-48]. Moreover, a computational study revealed a low fluctuation of the RVD loop as well as the marginal contribution of RVDs to specific binding energetics[47]. These findings suggest that TALE sequence specificity is achieved though negative discrimination arising from prohibitive steric and electrostatic clashes, rather than positive recognition[47].

Thus, based on the observations of the large dominance of TALE non-specific energetic contributions, TALE conformational elasticity, as well as the two-mode binding described here, we propose the following model of TALE target search.

TALE proteins sample the DNA sequence by one-dimensional diffusion, which is achieved by a combination of hopping and sliding movements. While bound to DNA, the protein exhibits at least two conformations that are in an equilibrium: the conformation with a superhelical pitch matching that of the DNA helix pitch, the so called "recognition mode", and at least one more conformation with a larger superhelical pitch, or "searching mode" (Figure 3.18). In the searching mode, the superhelical pitch is larger than that of the DNA double helix, and the nonspecific interactions of the TALE CRD are mostly out of register to interact favorably with the phosphate backbone. This conformation would minimize the nonspecific electrostatic interactions of the CRD with the DNA, while the nonspecific interactions of NTR with DNA would allow a fast sliding movement along the DNA backbone. The stability of the recognition mode at a given DNA sequence depends on the matching of the DNA sequence to the RVD array. TALE in the recognition mode interacting with a noncognate DNA sequence would experience steric and electrostatic clashes that would exert a disruptive effect on the nonspecific interactions of G14, K16, and/or Q17 with the DNA backbone, thus destabilizing the complex. In the case of a match, the amino acids G14, K16, and Q17 favorably align and nonspecifically interact with the phosphate backbone of the DNA, thus stabilizing the complex. A co-crystal structure of a TALE with noncognate DNA would be helpful in testing the validity of this model, and would shed light on the nature of the searching mode of TALEs.

The fundamental understanding of TALE target search mechanism could enable the development of more specific and efficient TALE proteins for use in basic

science or genome engineering applications. This work provides a molecular-level view of the dynamic search process of TALEs, which is a step forward in gaining a fundamental understanding of the target search process.

3.4 Materials and Methods

3.4.1 Preparation of DNA constructs

Preparation of substrate DNA

The plasmid was purified using the Plasmid Midi Kit (Qiagen) and linearized with *Sna*BI enzyme (New England Biolabs (NEB)) for 3 hours. It was next purified 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 bovine serum albumin (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 later.

Cloning of TALE21.5 and Naldt-TALE21.5

The gene coding for the untagged TAL effector protein (TALE21.5) was assembled using the Golden Gate cloning method (Addgene TALEN Kit #1000000024, as described in[49]), 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 TAL effector as well as the *BsmB*I sites corresponding to the kit *BsmB*I sites. For fluorescent tagging of the TAL effector, we modified the original plasmid pET-tSCA2 with an oligonucleotide insert encoding N-terminal LCPTSR hexapeptide (aldehyde tag[25,26]) 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). The protein sequence of the tagged Naldt-TALE21.5 is provided in the Appendix A.

Cloning Naldt-TALE15.5 and Naldt-TALE11.5

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-TALE15.5 and Nald-TALE11.5 were assembled using the Golden Gate cloning method[49] into the destination vector pET-NaldT-GG-TALE.

Cloning of Naldt-NTR

To create the NTR truncation, we digested pET-Naldt-TALE21.5 with *Sac*II and *Hind*III, and inserted the CAAAGCGTGGTGGCGTGACCGCGGTGGAAGCGGTCCATGCCTGGCGTAATGCGTTGACGG GCGCCCCCCTGAACTAAGTCAGATAACCGGATACAGACAAGCTTGCGGCCGCACTCGAGC

ACCAC sequence synthesized as a gBlock Gene Fragment (Integrated DNA Technologies (IDT)), using the Gibson Assembly Master Mix (NEB). The sequence of the NTR is provided in the Appendix A.

Cloning of Naldt-CRD-CTR

Primers (IDT) containing *BsmB*I sites corresponding to *BsmB*I 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 described above. The protein sequence of Naldt-CRD-CTR is provided in the Appendix A. All primer and plasmid sequences are available upon request.

3.4.2 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.

3.4.3 Protein Purification

The 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 (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 5 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 dialyzed 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.

The *in vivo* aldehyde labeling efficiency was quantified using data from LC-MS/MS. Briefly, unmodified TALE cysteines were alkylated using iodoacetamide[50], the protein was ran on SDS-PAGE gel (BioRad), and the band corresponding to the TALE was in-gel trypsin digested and analyzed by LC-MS/MS using the VG Quattro Electrospray Mass Spectrometer at the Roy J. Carver Biotechnology Center at UIUC. The masses of the peptides containing the untagged and tagged hexapeptide motifs

were filtered from the raw chromatograms and their relative quantities were derived from their peak counts. Fluorescence gel images were taken using the Molecular Dynamics Typhoon 9400 Multilaser Scanner at the Roy J. Carver Biotechnology Center. UV/Vis spectroscopic measurements were taken on Nanodrop 200 UV-Vis Spectrophotometer (Thermo Scientific).

3.4.4 Protein Labeling

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) or CF640R (Sigma) and labeled for 24 hours at room temperature in the dark. The labeled proteins were diluted with 400 µl fluorescence anisotropy buffer (20 µM Tris-HCl pH 7.5 100 mM Nacl 0.5 mM ethylenediaminetetraacetic acid (EDTA) (Fisher)) and purified from unreacted Cy3 or CF640R by two consecutive passages through Micro Bio-spin 6 columns (Bio-Rad) following manufacturer's instructions. In the case of NTR labeling, the purification of labeled NTR from free dye was performed using a TALON Spin Column (Clontech Laboratories) as per manufacturer's instructions, except that the column was washed 10 times until no more dye was observed in the flow-through.

3.4.5 Fluorescence Anisotropy

Twenty-nine bp-long oligonucleotide containing the 23 nt-long TAL effector 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-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 (sequences provided in the Appendix A).

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 1, where A is anisotropy and P is polarization. The K_D value was calculated by curve fitting on Origin 8.5 (OriginLab) using Equation 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.

$$A = \frac{2P}{3-P} \qquad (1)$$

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

3.4.6 Flow Cell Preparation and DNA Substrate Attachment (performed by Luke Cuculis, Dr. Charles M. Schroeder Lab)

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. 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. Polyethylene tubing was affixed to the ports drilled in each end of the flow cell in order to facilitate rapid exchange of buffer solutions and allow for stretching of tethered DNA.

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 μL/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[51]. Chloroquine is subsequently removed by washing the flow cell with blocking solution containing 40 mM $MgCl_2$ and 200 mM NaCl for 5 minutes.

3.4.7 Single Molecule Imaging (performed by Luke Cuculis, Dr. Charles M. Schroeder Lab)

Single molecule imaging experiments were carried out on an inverted total internal fluorescence (TIRF) microscope (Nikon IX70) coupled to an EMCCD camera (Andor iXon Ultra 897). Cy3 labeled proteins were excited using a 532 nm diode-pumped solid state (DPSS) laser (CrystaLaser) and SYTOX Green labeled DNA was excited using a 488 nm DPSS laser (SpectraPhysics Excelsior). Image sequences were acquired at a rate of 30-50Hz. 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. A triplet state quencher (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 blocking buffer, along with the protein.

3.4.8 Data Analysis (performed by Luke Cuculis, Dr. Charles M. Schroeder Lab)

Image sequences were recorded as TIF stacks using the Andor Solis software. Regions of interest containing single diffusing TALE proteins were isolated using ImageJ and the centroid locations of single proteins were determined using RapidStorm super resolution fitting software[52]. Single molecule trajectories were then further analyzed using custom MATLAB scripts.

3.5 Acknowledgements

We are thankful to Dr. Taekjip Ha for the generous gift of the pBAD-FGE plasmid and Dr. Xingua Shi for advice on the aldehyde-tagging procedure. We gratefully acknowledge Dr. Brian Imai and Dr Peter Yau at Roy J. Carver Biotechnology Center at UIUC for their assistance with LC-MS experiments of the proteins. We also thank Stella Xinzi Wu for her help with substrate DNA preparation.

3.7 References

(1) Bogdanove, A. J.; Voytas, D. F., TAL effectors: customizable proteins for DNA targeting. *Science* **2011,** *333*, 1843-6.
(2) Mussolino, C.; Cathomen, T., TALE nucleases: tailored genome engineering made easy. *Current opinion biotechnology* **2012,** *23*, 644-50.
(3) Boch, J.; Bonas, U., *Xanthomonas* AvrBs3 family-type III effectors: discovery and function. *Annual review of phytopathology* **2010,** *48*, 419-36.
(4) Bogdanove, A. J.; Schornack, S.; Lahaye, T., TAL effectors: finding plant genes for disease and defense. *Current opinion in plant biology* **2010,** *13*, 394-401.
(5) White, F. F.; Potnis, N.; Jones, J. B.; Koebnik, R., The type III effectors of Xanthomonas. *Molecular plant pathology* **2009,** *10*, 749-66.
(6) Boch, J.; Scholze, H.; Schornack, S.; Landgraf, A.; Hahn, S.; Kay, S.; Lahaye, T.; Nickstadt, A.; Bonas, U., Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **2009,** *326*, 1509-12.
(7) Moscou, M. J.; Bogdanove, A. J., A simple cipher governs DNA recognition by TAL effectors. *Science* **2009,** *326*, 1501.
(8) Christian, M.; Cermak, T.; Doyle, E. L.; Schmidt, C.; Zhang, F.; Hummel, A.; Bogdanove, A. J.; Voytas, D. F., Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics* **2010,** *186*, 757-61.
(9) Li, T.; Huang, S.; Jiang, W. Z.; Wright, D.; Spalding, M. H.; Weeks, D. P.; Yang, B., TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic acids research* **2011,** *39*, 359-72.
(10) Zhang, F.; Cong, L.; Lodato, S.; Kosuri, S.; Church, G. M.; Arlotta, P., Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nature biotechnology* **2011,** *29*, 149-53.
(11) Li, Y.; Moore, R.; Guinn, M.; Bleris, L., Transcription activator-like effector hybrids for conditional control and rewiring of chromosomal transgene expression. *Scientific reports* **2012,** *2*, 897.
(12) Geissler, R.; Scholze, H.; Hahn, S.; Streubel, J.; Bonas, U.; Behrens, S. E.; Boch, J., Transcriptional activators of human genes with programmable DNA-specificity. *PloS one* **2011,** *6*, DOI: 10.1371/journal.pone.0019509.

(13) Cong, L.; Zhou, R.; Kuo, Y. C.; Cunniff, M.; Zhang, F., Comprehensive interrogation of natural TALE DNA-binding modules and transcriptional repressor domains. *Nature communications* **2012,** *3*, 968.
(14) Mahfouz, M. M.; Li, L.; Piatek, M.; Fang, X.; Mansour, H.; Bangarusamy, D. K.; Zhu, J. K., Targeted transcriptional repression using a chimeric TALE-SRDX repressor protein. *Plant molecular biology* **2012,** *78*, 311-21.
(15) Maeder, M. L.; Angstman, J. F.; Richardson, M. E.; Linder, S. J.; Cascio, V. M.; Tsai, S. Q.; Ho, Q. H.; Sander, J. D.; Reyon, D.; Bernstein, B. E.; Costello, J. F.; Wilkinson, M. F.; Joung, J. K., Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. *Nature biotechnology* **2013,** *31*, 1137-42.
(16) Miyanari, Y.; Ziegler-Birling, C.; Torres-Padilla, M. E., Live visualization of chromatin dynamics with fluorescent TALEs. *Nature structural & molecular biology* **2013,** *20*, 1321-4.
(17) Ma, H.; Reyes-Gutierrez, P.; Pederson, T., Visualization of repetitive DNA sequences in human chromosomes with transcription activator-like effectors. *Proceedings of the national academy of sciences of the United States of America* **2013,** *110*, 21048-53.
(18) Deng, D.; Yan, C.; Pan, X.; Mahfouz, M.; Wang, J.; Zhu, J. K.; Shi, Y.; Yan, N., Structural basis for sequence-specific recognition of DNA by TAL effectors. *Science* **2012,** *335*, 720-3.
(19) Mak, A. N.; Bradley, P.; Cernadas, R. A.; Bogdanove, A. J.; Stoddard, B. L., The crystal structure of TAL effector PthXo1 bound to its DNA target. *Science* **2012,** *335*, 716-9.
(20) Gao, H.; Wu, X.; Chai, J.; Han, Z., Crystal structure of a TALE protein reveals an extended N-terminal DNA binding region. *Cell research* **2012,** *22*, 1716-20.
(21) Murakami, M. T.; Sforca, M. L.; Neves, J. L.; Paiva, J. H.; Domingues, M. N.; Pereira, A. L.; Zeri, A. C.; Benedetti, C. E., The repeat domain of the type III effector protein PthA shows a TPR-like structure and undergoes conformational changes upon DNA interaction. *Proteins* **2010,** *78*, 3386-95.
(22) Sun, N.; Liang, J.; Abil, Z.; Zhao, H., Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Molecular Biosystems* **2012,** *8*, 1255-63.
(23) Meckler, J. F.; Bhakta, M. S.; Kim, M. S.; Ovadia, R.; Habrian, C. H.; Zykovich, A.; Yu, A.; Lockwood, S. H.; Morbitzer, R.; Elsaesser, J.; Lahaye, T.; Segal, D. J.; Baldwin, E. P., Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic acids research* **2013,** *41*, 4118-28.
(24) Juillerat, A.; Dubois, G.; Valton, J.; Thomas, S.; Stella, S.; Marechal, A.; Langevin, S.; Benomari, N.; Bertonati, C.; Silva, G. H.; Daboussi, F.; Epinat, J. C.; Montoya, G.; Duclert, A.; Duchateau, P., Comprehensive analysis of the specificity of transcription activator-like effector nucleases. *Nucleic acids research* **2014,** *42*, 5390-402.
(25) Carrico, I. S.; Carlson, B. L.; Bertozzi, C. R., Introducing genetically encoded aldehydes into proteins. *Nature chemical biology* **2007,** *3*, 321-2.
(26) Shi, X.; Jung, Y.; Lin, L. J.; Liu, C.; Wu, C.; Cann, I. K.; Ha, T., Quantitative fluorescence labeling of aldehyde-tagged proteins for single-molecule imaging. *Nature methods* **2012,** *9*, 499-503.

(27) Li, M. Z.; Elledge, S. J., SLIC: a method for sequence- and ligation-independent cloning. *Methods in molecular biology* **2012,** *852*, 51-9.
(28) Vestergaard, C. L.; Blainey, P. C.; Flyvbjerg, H., Optimal estimation of diffusion coefficients from single-particle trajectories. *Physical review. E, Statistical, nonlinear, and soft matter physics* **2014,** *89*, DOI:10.1103/PhysRevE.89.022726.
(29) Miller, J. C.; Tan, S.; Qiao, G.; Barlow, K. A.; Wang, J.; Xia, D. F.; Meng, X.; Paschon, D. E.; Leung, E.; Hinkley, S. J.; Dulay, G. P.; Hua, K. L.; Ankoudinova, I.; Cost, G. J.; Urnov, F. D.; Zhang, H. S.; Holmes, M. C.; Zhang, L.; Gregory, P. D.; Rebar, E. J., A TALE nuclease architecture for efficient genome editing. *Nature biotechnology* **2011,** *29*, 143-8.
(30) Mussolino, C.; Morbitzer, R.; Lutge, F.; Dannemann, N.; Lahaye, T.; Cathomen, T., A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic acids research* **2011,** *39*, 9283-93.
(31) Sun, N.; Liang, J.; Abil, Z.; Zhao, H., Optimized TAL effector nucleases (TALENs) for use in treatment of sickle cell disease. *Molecular bioSystems* **2012***8,* 1255-63.
(32) Butenko, A. V.; Mogilko, E.; Amitai, L.; Pokroy, B.; Sloutskin, E., Coiled to diffuse: Brownian motion of a helical bacterium. *Langmuir* **2012,** *28*, 12941-7.
(33) Riggs, A. D.; Bourgeois, S.; Cohn, M., The lac repressor-operator interaction. 3. Kinetic studies. *Journal of molecular biology* **1970,** *53*, 401-17.
(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* **1981,** *20*, 6929-48.
(35) 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* **1981,** *20*, 6961-77.
(36) Berg, O. G.; Ehrenberg, M., Association kinetics with coupled three- and one-dimensional diffusion. Chain-length dependence of the association rate of specific DNA sites. *Biophysical chemistry* **1982,** *15*, 41-51.
(37) von Hippel, P. H.; Berg, O. G., Facilitated target location in biological systems. *The Journal of biological chemistry* **1989,** *264*, 675-8.
(38) Halford, S. E.; Marko, J. F., How do site-specific DNA-binding proteins find their targets? *Nucleic acids research* **2004,** *32*, 3040-52.
(39) Kampmann, M., Facilitated diffusion in chromatin lattices: mechanistic diversity and regulatory potential. *Molecular microbiology* **2005,** *57*, 889-99.
(40) Wang, F.; Redding, S.; Finkelstein, I. J.; Gorman, J.; Reichman, D. R.; Greene, E. C., The promoter-search mechanism of *Escherichia coli* RNA polymerase is dominated by three-dimensional diffusion. *Nature structural & molecular biology* **2013,** *20*, 174-81.
(41) Sternberg, S. H.; Redding, S.; Jinek, M.; Greene, E. C.; Doudna, J. A., DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **2014,** *507*, 62-7.
(42) Hedglin, M.; O'Brien, P. J., Hopping enables a DNA repair glycosylase to search both strands and bypass a bound protein. *ACS chemical biology* **2010,** *5*, 427-36.
(43) Slutsky, M.; Mirny, L. A., Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. *Biophysical journal* **2004,** *87*, 4021-35.

(44) Benichou, O.; Kafri, Y.; Sheinman, M.; Voituriez, R., Searching fast for a target on DNA without falling to traps. *Physical review letters* **2009,** *103*, 138102.
(45) Wan, H.; Hu, J. P.; Li, K. S.; Tian, X. H.; Chang, S., Molecular dynamics simulations of DNA-free and DNA-bound TAL effectors. *PloS one* **2013,** *8*, e76045.
(46) Flechsig, H., TALEs from a spring - superelasticity of Tal effector protein structures. *PloS one* **2014,** *9*, DOI: 10.1371/journal.pone.0109919.
(47) Wicky, B. I.; Stenta, M.; Dal Peraro, M., TAL effectors specificity stems from negative discrimination. *PloS one* **2013,** *8*, DOI: 10.1371/journal.pone.0080261.
(48) Meckler, J. F.; Bhakta, M. S.; Kim, M. S.; Ovadia, R.; Habrian, C. H.; Zykovich, A.; Yu, A.; Lockwood, S. H.; Morbitzer, R.; Elsaesser, J.; Lahaye, T.; Segal, D. J.; Baldwin, E. P., Quantitative analysis of TALE-DNA interactions suggests polarity effects. *Nucleic acids research* **2013**, *41,* 4118-28.
(49) Cermak, T.; Doyle, E. L.; Christian, M.; Wang, L.; Zhang, Y.; Schmidt, C.; Baller, J. A.; Somia, N. V.; Bogdanove, A. J.; Voytas, D. F., Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research* **2011,** *39*, DOI: 10.1093/nar/gkr218.
(50) Sechi, S.; Chait, B. T., Modification of cysteine residues by alkylation. A tool in peptide mapping and protein identification. *Analytical chemistry* **1998,** *70*, 5150-8.
(51) Yardimci, H.; Loveland, A. B.; van Oijen, A. M.; Walter, J. C., Single-molecule analysis of DNA replication in *Xenopus* egg extracts. *Methods* **2012,** *57*, 179-86.
(52) Wolter, S.; Löschberger, A.; Holm, T.; Aufmkolk, S.; Dabauvalle, M. C.; van de Linde, S.; Sauer, M., rapidSTORM: accurate, fast open-source software for localization microscopy. *Nature Methods* **2012,** *9*, 1040-1.

Chapter 4. PUF Module Library Creation and PUF Assembly

4.1 Introduction

The presence of RNA-binding proteins (RBPs) in RNA biology is ubiquitous. Hundreds to thousands of eukaryotic proteins are estimated to function as RBPs[1] and govern many aspects of RNA biology including translation, turnover, processing, and cellular localization[2-4]. Despite their great diversity in function, only a few types of RNA-binding domains are known, which are combined in different structural arrangements with a variety of functional domains[5]. This modular architecture makes RBPs an attractive tool for studying the vast complexity of eukaryotic transcriptomes as well as manipulating RNA for therapeutic purposes[6-7].

The function of many RBPs can be studied[8-11] by tethering them to a reporter RNA through a well-characterized RNA-binding peptide with a fixed specificity[12]. However, this approach can only be applied to manipulate heterologous RNA because prior tagging of the RNA is required. In order to manipulate endogenous RNA in its native expression conditions, one could envision a designer RBP with an RNA-binding scaffold that could be easily engineered for sequence specificity. To date, only pentatricopeptide repeat[13-14] and Pumilio/fem-3 mRNA binding factor (PUF)[15-18] proteins have been demonstrated to have the potential to be rationally modified for predictable and specific RNA recognition.

PUF proteins are eukaryotic RBPs that are involved in post-transcriptional gene regulation[19]. The crystal structure of Pumilio homology domain (PUM-HD), the RNA-binding region of the human Pumilio 1 (PUM1) protein (Figure 4.1a), reveals 8

structural repeats, each containing ~36 amino acids (aa), and flanking N-and C-terminal regions[20-21]. The structure also suggests that recognition of the target RNA sequence is highly modular since each repeat binds to a single RNA base[15]. The N-terminal repeat (R1) binds to the 3′-nucleotide residue (N8) of the target sequence (Figure 4.1a and b), while the C-terminal repeat (R8) binds to the 5′-nucleotide residue (N1). Residues at positions 12 and 16 in each repeat directly interact with a Watson-Crick edge of a base, whereas the residue at position 13 is involved in a stacking interaction between two adjacent bases[15]. The structure suggests a recognition "code", where residues at positions 12 and 16 in each repeat contribute to specific recognition of a base, with $N_{12}Q_{16}$ recognizing uracil, $C_{12}Q_{16}$ adenine, and $S_{12}E_{16}$ guanine[15]. The residue combination $S_{12}R_{16}$ was engineered to recognize cytosine[17-18]. By swapping the key residues at these positions, it was shown that designed PUF proteins with altered specificity could be engineered[16, 22]. In the past several years, engineered PUF domains were successfully fused to different effector domains for polyadenylation of an endogenous gene or repression of a reporter gene in *Xenopus*[23], cleavage of a mitochondrial-encoded gene in human cells[24], splicing regulation of endogenous *Bcl-X* pre-mRNA in human cells[25], and imaging endogenous RNA[26-28]. These advancements demonstrate the growing potential for the RBPs with various functional domains and engineered specificity.

However, PUF-based RBPs are still far from widespread implementation. One of the limitations in engineering PUF domains with novel specificities is the lack of a cloning platform capable of rapid and efficient introduction of multiple mutations in separate repeats simultaneously. In this study, we report the implementation of the

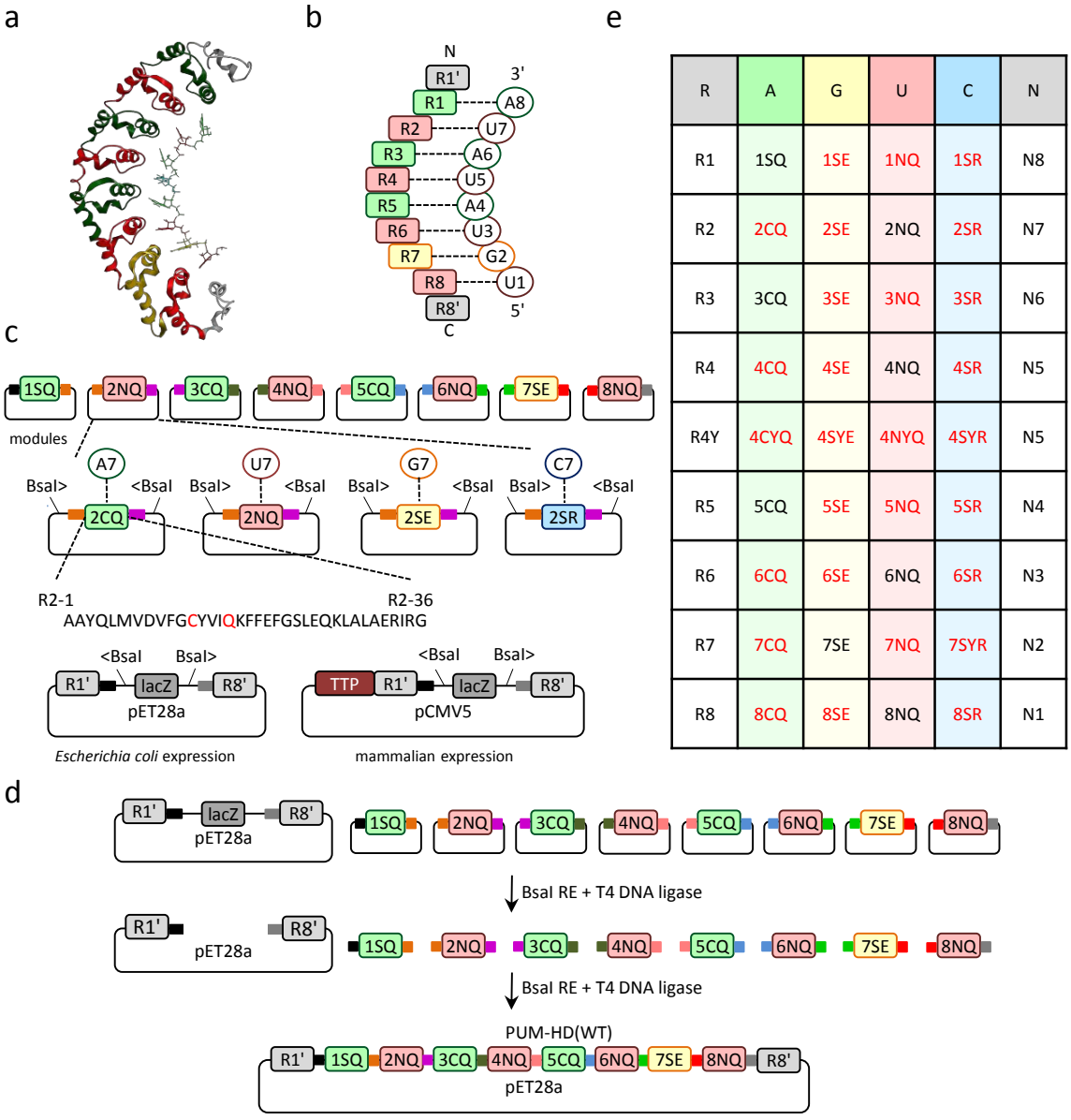
Golden Gate (GG) cloning, Type IIS restriction endonuclease-based approach[29] for engineering of PUF-based RBPs. To demonstrate the efficiency of this approach, we first used this cloning method for construction of several mutant PUF domains with novel specificities and assayed their binding affinities *in vitro*. Second, we tested the binding activity of multiple PUF mutants in yeast and compared the effect of various mutations at different positions on PUF functionality *in vivo*.

4.2 Results and Discussion

4.2.1 Assembly of the PUF Gene from Modules Using the Golden Gate Method

The GG cloning method, which is implemented here for the assembly of custom PUF domains, is based on the ability of Type IIS restriction enzymes to cleave outside of their non-palindromic recognition sequence[29], thus creating overhangs unrelated to the recognition sequence. This polarity and flexibility in the overhang sequence allows for a seamless removal of the original restriction site as well as a ligation of multiple fragments in one step.

As a scaffold for the development of the assembly toolkit, we used the human PUM-HD, which consists of the amino acids 828-1176 of the full-length PUM1[21]. Each of the 8 structural repeats of PUM-HD was cloned individually into a pNEB193-based "intermediate vector" and was used as a separate assembly module (Figure 4.1c). We designed all 8 modules as well as the *lacZ*α gene in the "receiving vector" to be flanked by two *Bsa*I sites in such a way that would allow creation of 9 unique overhangs (Figure 4.1c). In a one-pot reaction, the 8 modules and the receiving vector can be efficiently cut and re-ligated in a predefined order (Figure 4.1d).



R	A	G	U	C	N
R1	1SQ	1SE	1NQ	1SR	N8
R2	2CQ	2SE	2NQ	2SR	N7
R3	3CQ	3SE	3NQ	3SR	N6
R4	4CQ	4SE	4NQ	4SR	N5
R4Y	4CYQ	4SYE	4NYQ	4SYR	N5
R5	5CQ	5SE	5NQ	5SR	N4
R6	6CQ	6SE	6NQ	6SR	N3
R7	7CQ	7SE	7NQ	7SYR	N2
R8	8CQ	8SE	8NQ	8SR	N1

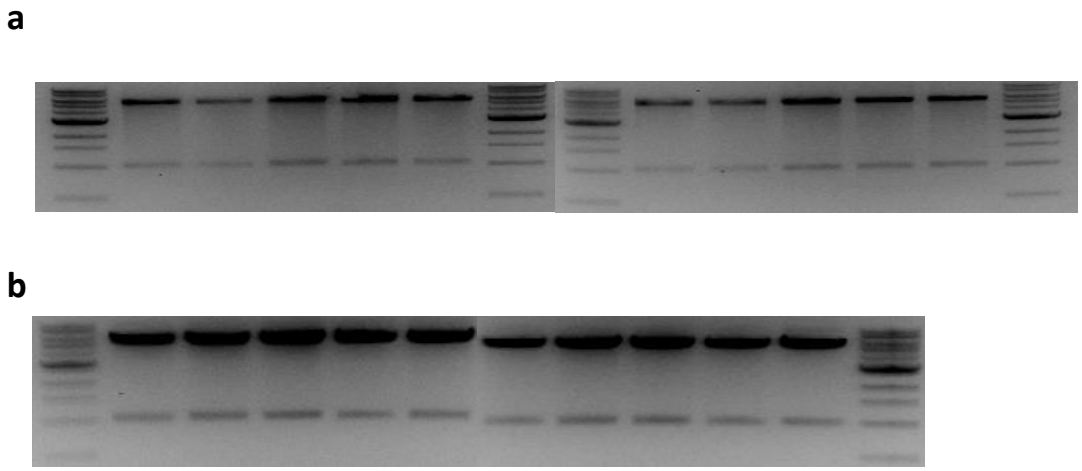
Abil *et al. J Biol Eng* 2014, 8:7

Figure 4.1 The GG library and assembly schematic. **(a)** Crystal structure of PUM-HD bound to RNA, adapted from reference [15], GenBank ID code 1M8Y. **(b)** Schematic of PUM-HD bound to RNA. Filled boxes, PUF modules. Circles, RNA bases. **(c)** Schematic of the main library components: 8 repeat modules, with matching overhangs colored in identical colors; 4 variations of module 2, with corresponding recognition nucleotide indicated above; the aa sequence of module 2, with mutant aa indicated in red; and two receiving vectors. **(d)** The GG assembly schematic. A one-pot reaction that contains 8 modules of choice, a receiving vector, and enzymes allows the creation of 9 unique overhangs. The exact matching of the overhangs results in the predetermined repeat order

assembled in the receiving vector. **(e)** Schematic of the GG library. R, module; N, nucleotide; recognized nucleotides indicated in the top row. First and last letters in the module names represent aa residues 12 and 16, in each module, respectively. Middle letter, if present, represents the "stacking" aa 13. Black font, WT modules. Red font, mutant modules. Green, yellow, pink, and blue fillings for modules recognizing A, G, U, and C, respectively.

We constructed two receiving vectors, pET28-GG-PUF for *Escherichia coli* expression and pCMV-TTP(C147R)-GG-PUF for mammalian expression. They both contain a *lacZ*α gene for easy identification of assembled clones using blue-white screening, as well as the flanking N- and C-terminal regions of PUM-HD, which we refer to as R1′ and R8′, respectively.

In order to test the efficiency of the method, we re-assembled the original PUM-HD from the WT modules into both receiving vectors. We analyzed 10 randomly chosen clones from each assembly by restriction digestion and found that all produced the anticipated digestion pattern (Figure 4.2a and b). Next, we sequenced one clone from each assembly and found that both sequences were correct, indicating that the assembly process is highly efficient regardless of a receiving vector. Assembly of the consecutive mutant PUF domains was as efficient as the assembly of the WT PUF domain (data not shown). The entire process takes 3 days, with the GG reaction and *E. coli* transformation on day 1, colony picking on day 2, and plasmid purification and digestion-confirmation of the clones on day 3. The procedure is therefore ideal for the rapid introduction of multiple mutations in a PUF domain with high efficiency.



Abil *et al. J Biol Eng* 2014, 8:7

Figure 4.2 Confirmation through restriction enzyme digestion and gel-electrophoresis of GG assembled plasmids from randomly picked clones. First and last lanes, 1 kb DNA ladder (NEB). **(a)** *Kpn*I and *Hind*III digestion of PUF(WT) clones assembled into pET28-GG-PUF receiving vector. 1 kb fragment contains the full length of the assembled PUF domain. **(b)** *Sal*I and *Kpn*I digestion of PUF(WT) clones assembled into pCMV-TTP-GG-PUF receiving vector. 1 kb fragment contains the assembled PUF domain region.

4.2.2 Development of the PUF Module Library

The 8 modules were further expanded into a library where each module has 4 variations for the recognition of any of the 4 nucleotides, consistent with the PUF recognition "code"[15, 17-18]. The variants of the same module position have the same overhangs and the same amino acids as the wild type (WT) module except at positions 12 and 16 (Figure 4.1c, Table 4.1). The introduced mutations are uniform across modules (Figure 4.1e, Table 4.1), except for module 7 for the recognition of cytosine, where the "stacking" residue was also substituted with tyrosine, in

Repeat	Recognition	AA sequence
1′		MGRSRLLEDFRNNRYPNLQLREIAG
1	A	HIMEFSQDQHGSRFIQLKLERATPAERQLVFNEILQ
1	G	HIMEFSQDQHGSRFIELKLERATPAERQLVFNEILQ
1	U	HIMEFSQDQHGNRFIQLKLERATPAERQLVFNEILQ
1	C	HIMEFSQDQHGSRFIRLKLERATPAERQLVFNEILQ
2	A	AAYQLMVDVFGCYVIQKFFEFGSLEQKLALAERIRG
2	G	AAYQLMVDVFGSYVIEKFFEFGSLEQKLALAERIRG
2	U	AAYQLMVDVFGNYVIQKFFEFGSLEQKLALAERIRG
2	C	AAYQLMVDVFGSYVIRKFFEFGSLEQKLALAERIRG
3	A	HVLSLALQMYGCRVIQKALEFIPSDQQNEMVRELDG
3	G	HVLSLALQMYGSRVIEKALEFIPSDQQNEMVRELDG
3	U	HVLSLALQMYGNRVIQKALEFIPSDQQNEMVRELDG
3	C	HVLSLALQMYGSRVIRKALEFIPSDQQNEMVRELDG
4	A	HVLKCVKDQNGCHVVQKCIECVQPQSLQFIIDAFKG
4	G	HVLKCVKDQNGSHVVEKCIECVQPQSLQFIIDAFKG
4	U	HVLKCVKDQNGNHVVQKCIECVQPQSLQFIIDAFKG
4	C	HVLKCVKDQNGSHVVRKCIECVQPQSLQFIIDAFKG
4	A	HVLKCVKDQNGCYVVQKCIECVQPQSLQFIIDAFKG
4	G	HVLKCVKDQNGSYVVEKCIECVQPQSLQFIIDAFKG
4	U	HVLKCVKDQNGNYVVQKCIECVQPQSLQFIIDAFKG
4	C	HVLKCVKDQNGSYVVRKCIECVQPQSLQFIIDAFKG
5	A	QVFALSTHPYGCRVIQRILEHCLPDQTLPILEELHQ
5	G	QVFALSTHPYGSRVIERILEHCLPDQTLPILEELHQ
5	U	QVFALSTHPYGNRVIQRILEHCLPDQTLPILEELHQ
5	C	QVFALSTHPYGSRVIRRILEHCLPDQTLPILEELHQ
6	A	HTEQLVQDQYGCYVIQHVLEHGRPEDKSKIVAEIRG
6	G	HTEQLVQDQYGSYVIEHVLEHGRPEDKSKIVAEIRG
6	U	HTEQLVQDQYGNYVIQHVLEHGRPEDKSKIVAEIRG
6	C	HTEQLVQDQYGSYVIRHVLEHGRPEDKSKIVAEIRG
7	A	NVLVLSQHKFACNVVQKCVTHASRTERAVLIDEVCTMNDGPHS
7	G	NVLVLSQHKFASNVVEKCVTHASRTERAVLIDEVCTMNDGPHS
7	U	NVLVLSQHKFANNVVQKCVTHASRTERAVLIDEVCTMNDGPHS
7	C	NVLVLSQHKFASYVVRKCVTHASRTERAVLIDEVCTMNDGPHS
8	A	ALYTMMKDQYACYVVQKMIDVAEPGQRKIVMHKIRP
8	G	ALYTMMKDQYASYVVEKMIDVAEPGQRKIVMHKIRP
8	U	ALYTMMKDQYANYVVQKMIDVAEPGQRKIVMHKIRP
8	C	ALYTMMKDQYASYVVRKMIDVAEPGQRKIVMHKIRP
8′		HIATLRKYTYGKHILAKLEKYYMKNGVDLG

Table 4.1 (Cont. on next page)

Table 4.1 GG library sequences **(a)** Aa sequences of WT and mutant modules. Black, WT aa. Red, mutant aa. **(b)** DNA sequences of WT and mutant modules. Black, WT sequence. Red, mutant nucleotides.

accordance with a previous report by Dong *et al.* [18]. Since module 3S $_{12}$R$_{16}$ requires a tyrosine as a "stacking" residue in the adjacent module 4 for *in vivo* activity[18], there are four additional module 4 variants, where the "stacking" residue at position 13 was mutated from histidine to tyrosine. The entire library hence consists of 36 intermediate vectors and 2 receiving vectors. Thus, our library of PUF repeat modules is potentially capable of a one-step assembly of PUF domains with specificity for any RNA sequence of 8 nt, given that they can be expressed in a soluble manner.

4.2.3 Analysis of PUF RNA-Binding *in vitro*

With the combinatorial assembly tool in hand, we set to determine if increasing the number of mutations affects the activity or specificity of a PUF domain. We assembled four variant PUF domains with 2, 4, 6, and 8 mutant modules that were named PUF(S2), PUF(S4), PUF(S6), and PUF(S8), respectively (Table 4.2). To test their *in vitro* activity, we assayed the binding affinity of the proteins to WT and their cognate RNA sequences using a fluorescence polarization assay (Figure 4.3b-f). We found that the WT PUF as well as the PUF variants all bound to their cognate RNA sequences with high affinity (Table 4.2). As predicted, all of the PUF proteins exhibited lower affinity to noncognate RNA, which contained between 2 and 8 nucleotide mismatches with the cognate RNA (Table 4.2, Figure 4.3b-f). The binding

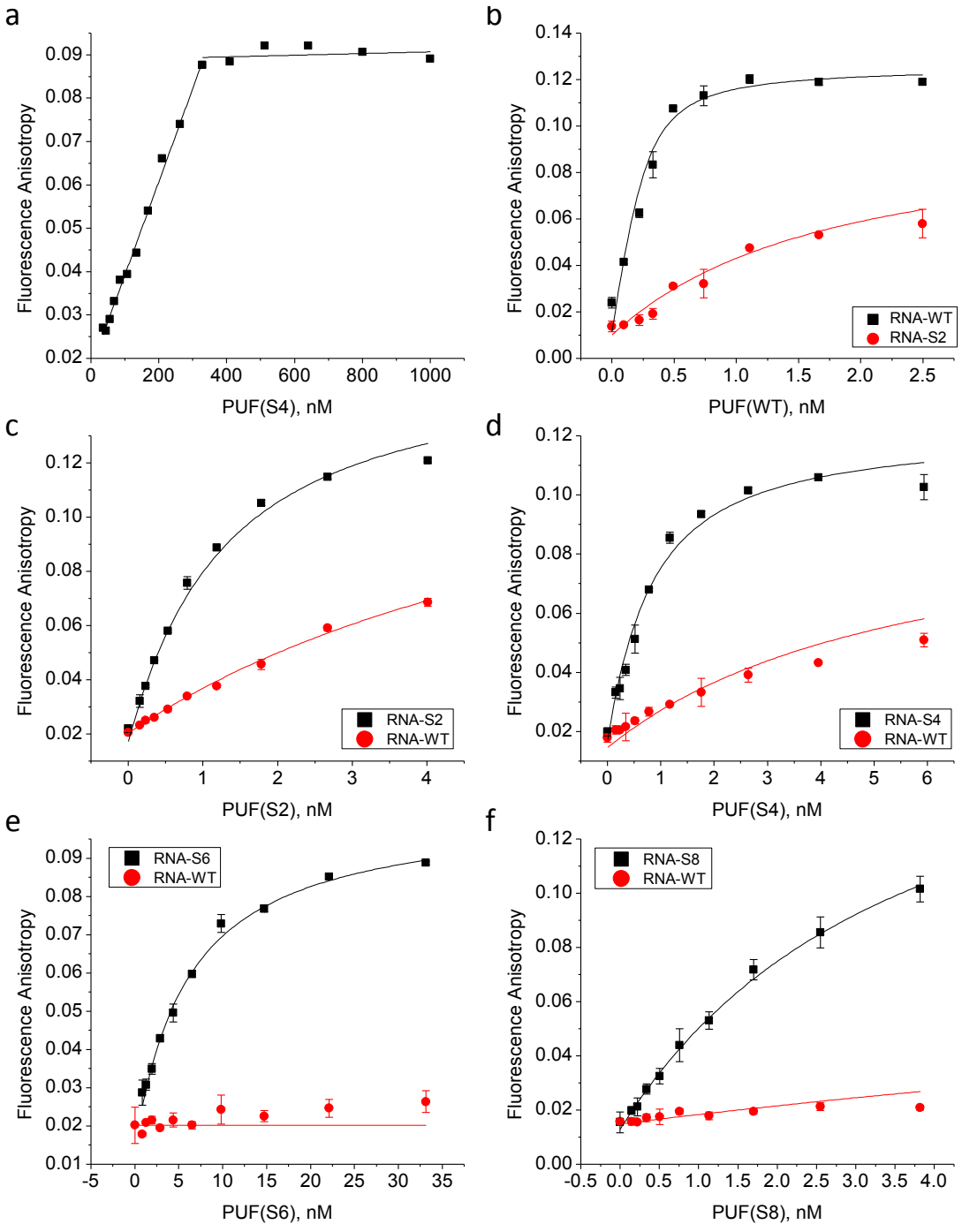


Figure 4.3 Representative fluorescence anisotropy data for RNA binding to various PUF proteins. **(a)** Representative saturation curve of PUF(S4). **(b)** Binding curves of PUF(WT) **(c)** Binding curves of PUF(S2) **(d)** Binding curves of PUF(S4) **(e)** Binding curves of PUF(S6) **(f)** Binding curves of PUF(S8). Black, binding to cognate RNA. Red, binding to non-cognate RNA. Each data point is represented by the mean ± SD. K_D values were calculated from nonlinear curve fitting.

Abil *et al. J Biol Eng* 2014, 8:7

affinities to cognate sequences decreased from PUF(WT) to PUF(S8), indicating that there is a weak negative correlation between the number of mutations and binding affinity to cognate RNA, although the correlation is not strictly linear (Table 4.2). However, even PUF(S8), which has 8/8 mutant modules, binds specifically and with

Protein	Protein modules	Cognate RNA sequence	K_D, nM	Noncognate RNA sequence	K_D, nM
	Ct 8 7 6 5 4 3 2 1 Nt	5' 12345678 3'		5' 12345678 3'	
PUF(WT)	NQ-SE-NQ-CQ-NQ-CQ-NQ-SQ	CCAGAAUUGUAUAUAUUCG	0.043 ± 0.023	CCAGAAUAGUAUAUUUUCG	3.56 ± 0.90
PUF(S2)	CQ-SE-NQ-CQ-NQ-CQ-NQ-NQ	CCAGAAUAGUAUAUUUUCG	0.76 ± 0.11	CCAGAAUUGUAUAUAUUCG	6.29 ± 2.80
PUF(S4)	NQ-SE-CQ-NQ-CYQ-NQ-NQ-SQ	CCAGAAUUGAUAUUAUUCG	0.59 ± 0.10	CCAGAAUUGUAUAUAUUCG	5.77 ± 1.33
PUF(S6)	CQ-SE-CQ-NQ-CYQ-NQ-NQ-NQ	CCAGAAUAGAUAUUUUUCG	6.05 ± 0.25	CCAGAAUUGUAUAUAUUCG	89.6 ± 16.5
PUF(S8)	CQ-NQ-CQ-NQ-CYQ-NQ-SE-NQ	CCAGAAUAUAUAUGUUUCG	2.79 ± 0.69	CCAGAAUUGUAUAUAUUCG	ND

Abil *et al. J Biol Eng* 2014, 8:7

Table 4.2 Mutations, cognate and non-cognate RNA oligonucleotide sequences, and binding affinities of PUF(WT), and PUFs S2-S8. Black, WT modules and corresponding RNA bases. Red, mutant modules and corresponding RNA bases. Bold, mismatching RNA nucleotides. Ct, C-terminus, and Nt, N-terminus of the protein. K_D values were determined using nonlinear curve fitting and represent the mean ± SD (n=3). ND, K_D not determined due to little binding at 300 nM PUF(S8) to WT RNA.

high affinity to its cognate RNA. These results further corroborate the study by Cheong and Hall, who demonstrated the specificity and modularity of PUF protein target recognition[16].

4.2.4 Analysis of PUF RNA-Binding in *Saccharomyces cerevisiae*[3]

The modular nature of RNA recognition by PUF proteins has been previously reported[15-18], and suggests that the specificity of the protein can be predictably

[3] This section was done in collaboration with Stella Xinzi Wu, University of Illinois.

modified. However, certain natural or engineered PUF proteins behave unpredictably. Thus, bases in the target RNA sequence can be flipped out in the protein:RNA complexes *in vitro,* the specificity of the engineered repeat is sometimes broadened, or engineered repeats can affect the protein's recognition of bases adjacent to the target base[22, 30-31]. In addition, we have observed the drop in soluble expression of certain PUF mutants in mammalian or *E. coli* cells (data not shown), which can also be an obstacle for PUF engineering. For the development of an optimal PUF backbone for highly modular, predictable RNA recognition, systematic investigation of PUF RNA recognition, specificity, and mutant stability is required. In this section, we sought to investigate how the position of introduced mutations along PUM1 structural repeats affects the RNA binding activity of the protein.

To measure the *in vivo* protein-RNA interactions between predicted cognate RNA and PUF domains, we implemented the yeast three hybrid system[32]. This approach utilizes the bacterial repressor LexA fusion with the MS2 coat protein, which is recruited to the LexA operator upstream of the reporter gene. The RNA element of the system consists of the MS2-binding site attached to the RNA sequence of interest, and thus is recruited to the reporter gene together with LexA-MS2 fusion. The protein of interest, in turn, is fused to the Gal4 activation domain, which, upon the interaction with the RNA sequence of interest, is likewise recruited to the reporter gene and activates its transcription[32-33]. Here, we utilized the *LacZ* gene as a reporter.

Act2-PUF	Cognate RNA sequence (5′-3′)	PUF modules							
		8	7	6	5	4	3	2	1
T	UAUAUAUA	NQ	CQ	NQ	CQ	NQ	CQ	NQ	CQ
S1	AAUAUAUA	CQ	CQ	NQ	CQ	NQ	CQ	NQ	CQ
S2	UUUAUAUA	NQ	NQ	NQ	CQ	NQ	CQ	NQ	CQ
S3	UAAAUAUA	NQ	CQ	CQ	CQ	NQ	CQ	NQ	CQ
S4	UAUUUAUA	NQ	CQ	NQ	NQ	NQ	CQ	NQ	CQ
S5	UAUAAAUA	NQ	CQ	NQ	CQ	CQ	CQ	NQ	CQ
S6	UAUAUUUA	NQ	CQ	NQ	CQ	NQ	NQ	NQ	CQ
S7	UAUAUAAA	NQ	CQ	NQ	CQ	NQ	CQ	CQ	CQ
S8	UAUAUAUU	NQ	CQ	NQ	CQ	NQ	CQ	NQ	NQ
M1	AUAUAUAU	CQ	NQ	CQ	NQ	CQ	NQ	CQ	NQ
M2	AUAUAUAA	CQ	NQ	CQ	NQ	CQ	NQ	CQ	CQ
M3	AUAUAUUA	CQ	NQ	CQ	NQ	CQ	NQ	NQ	CQ
M4	AUAUAAUA	CQ	NQ	CQ	NQ	CQ	CQ	NQ	CQ
M5	AUAUUAUA	CQ	NQ	CQ	NQ	NQ	CQ	NQ	CQ
M6	AUAAUAUA	CQ	NQ	CQ	CQ	NQ	CQ	NQ	CQ
M7	AUUAUAUA	CQ	NQ	NQ	CQ	NQ	CQ	NQ	CQ
M8	AAUAUAUA	CQ	CQ	NQ	CQ	NQ	CQ	NQ	CQ
G1	GAUAUAUA	SE	CQ	NQ	CQ	NQ	CQ	NQ	CQ
WT	UGUAUAUA	NQ	SE	NQ	CQ	NQ	CQ	NQ	CQ
G3	UAGAUAUA	NQ	CQ	SE	CQ	NQ	CQ	NQ	CQ
G4	UAUGUAUA	NQ	CQ	NQ	SE	NQ	CQ	NQ	CQ
G5	UAUAGAUA	NQ	CQ	NQ	CQ	SE	CQ	NQ	CQ
G6	UAUAUGUA	NQ	CQ	NQ	CQ	NQ	SE	NQ	CQ
G7	UAUAUAGA	NQ	CQ	NQ	CQ	NQ	CQ	SE	CQ
G8	UAUAUAUG	NQ	CQ	NQ	CQ	NQ	CQ	NQ	SE
C1	CAUAUAUA	SR	CQ	NQ	CQ	NQ	CQ	NQ	CQ
C2	UCUAUAUA	NQ	SR	NQ	CQ	NQ	CQ	NQ	CQ
C3	UACAUAUA	NQ	CQ	SR	CQ	NQ	CQ	NQ	CQ
C4	UAUCUAUA	NQ	CQ	NQ	SR	NQ	CQ	NQ	CQ
C5	UAUACAUA	NQ	CQ	NQ	CQ	SR	CQ	NQ	CQ
C6	UAUAUCUA	NQ	CQ	NQ	CQ	NQ	SR	NQ	CQ
C7	UAUAUACA	NQ	CQ	NQ	CQ	NQ	CQ	SR	CQ
C8	UAUAUAUC	NQ	CQ	NQ	CQ	NQ	CQ	NQ	SR

Table 4.3 Mutations, cognate RNA oligonucleotide sequences, and *in vivo* activities of Act2-PUF fusion proteins. Black, WT modules and corresponding RNA bases. Red, mutant modules and corresponding RNA bases.

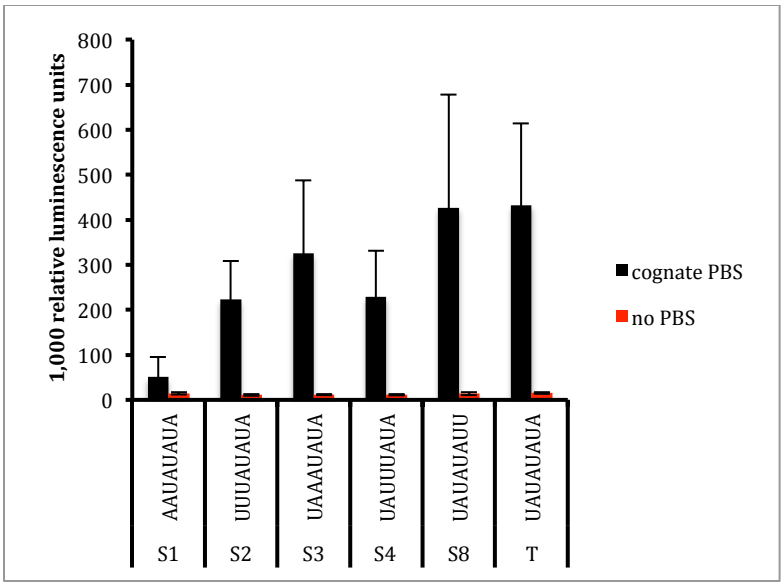


Figure 4.4 Relative luciferase activity of Act-PUF proteins (S1-S8), measured by yeast three hybrid assay. Predicted cognate RNA sequences are indicated (bottom-top, 5'-3'). Data represented as average ± SD.

To investigate if the nature and position of mutations affect PUF activity differently than others, we mutagenized each PUF repeat for an altered specificity one repeat at a time, starting from a PUF mutant that recognizes an RNA target with lowest complexity: the 5'-UAUAUAUA-3' sequence. In this mutant, which we termed Act2-PUF (T), $7S_{12}E_{16}$ was mutagenized to $7N_{12}Q_{16}$, for the change in recognition of PBS from 2G to 2A (Table 4.3). PUF mutants included proteins with: single modules replaced from $N_{12}Q_{16}$ to $C_{12}Q_{16}$ or vice versa (S1-S8), various numbers of modules replaced from $N_{12}Q_{16}$ to $C_{12}Q_{16}$ or vice versa (M1-M8), a module recognizing a G ($S_{12}E_{16}$) introduced individually in each module position (G1-G8), or a module recognizing a C ($S_{12}R_{16}$) introduced individually in each module position (C1-C8, see Table 4.3). Despite several attempts, some of the plasmid combinations encoding Act2-PUF fusions and corresponding transcripts with PUF-binding sites (PBS) did

not yield *S. cerevisiae* transformants (S5-S7, M6-M8, and G1), possibly due to chance mutations in plasmid backbones. Due to time constraints, these plasmids were not investigated further. The remaining Act2-PUF fusions were investigated using the Beta-Glo luciferase assay (Figure 4.4a-d).

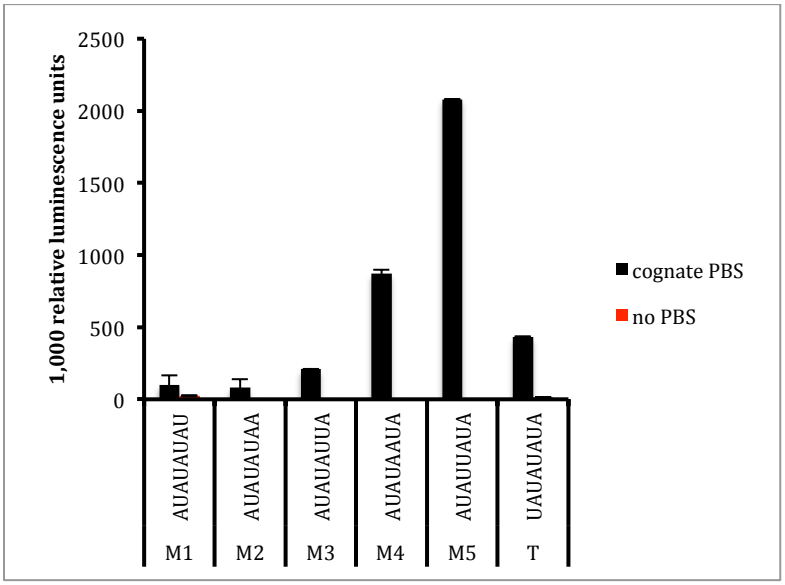


Figure 4.5 Relative luciferase activity of Act-PUF proteins (M1-M8), measured by yeast three hybrid assay. Predicted cognate RNA sequences are indicated (bottom-top, 5'-3'). Data represented as average ± SD.

Relative luciferase activities, indicative of *in vivo* PUF-RNA interaction, were measured for each Act2-PUF fusion either in combination with RNA with cognate PBS or RNA lacking PBS. We observed that Act2-PUF (T) exhibited ~4-fold decrease in activity compared to the Act2-PUF (WT) (Figure 4.6), indicating the importance of $S_{12}E_{16}$ module in repeat 7 for the RNA-binding activity of PUF. However, additional single module replacements of $N_{12}Q_{16}$ to $C_{12}Q_{16}$ or vice versa in other module positions did not result in considerable decrease in activity (Figure 4.4). The

exception was the mutagenesis of repeat 8 in the Act2-PUF mutant S1 (Figure 4.4), which demonstrated a significant decrease in activity compared to other positions. The triplet UGU, which is recognized by repeats 6-8 of PUF domains, is conserved among recognition sequences of many PUFs[15]. Our results suggest that PUF repeats 7 and 8, and to a smaller extent, repeat 6, may also be the most sensitive towards mutagenesis, probably due to a greater energetic contribution of these repeats to the overall PUF RNA-binding activity.

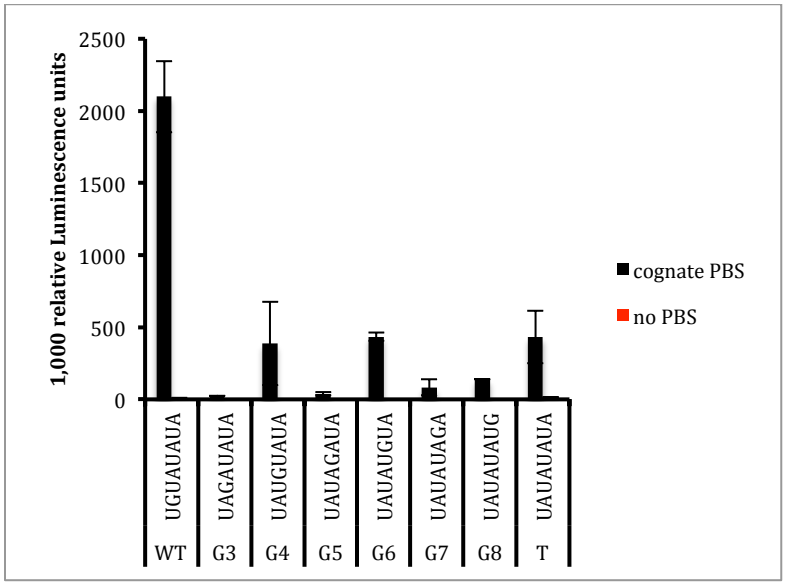


Figure 4.6 Relative luciferase activity of Act-PUF proteins (G3-G8), measured by yeast three hybrid assay. Predicted cognate RNA sequences are indicated (bottom-top, 5'-3'). Data represented as average ± SD.

Decreasing activities of Act2-PUF fusions from M5 to M1 (which correspond to PUF mutants with 4 replaced modules to those with 8 replaced modules, Table 4.3) suggest that increasing the number of mutations reduces either PUF stability and/or RNA-binding affinity (Figure 4.5). It is likely that RNA-binding affinity of PUF proteins indeed generally decreases with increased number of mutations in the

protein, as was shown in Table 4.2 above. We suggest that this could be due to the role of the residues at positions 12 and 16 in each repeat not only in binding to specific RNA base, but in the structural stability of the protein as well.

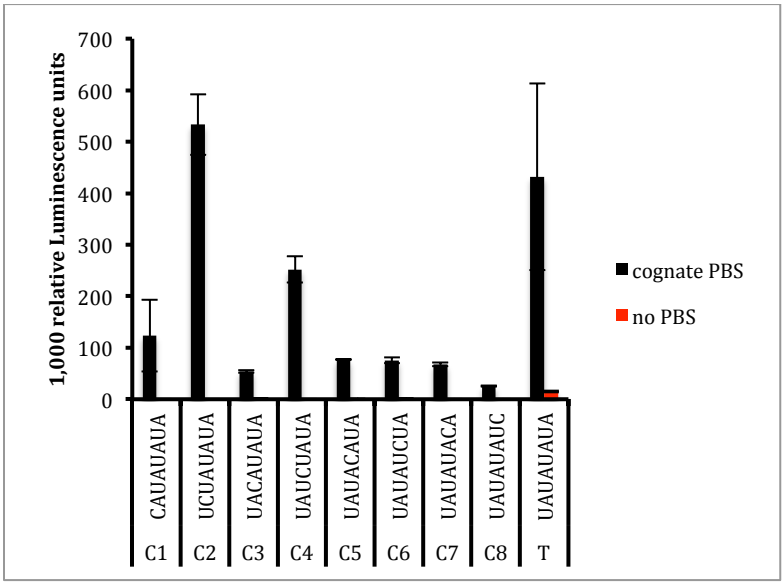


Figure 4.7 Relative luciferase activity of Act-PUF proteins (C1-C8), measured by yeast three hybrid assay. Predicted cognate RNA sequences are indicated (bottom-top, 5'-3'). Data represented as average ± SD.

Among Act2-PUF fusions containing the $S_{12}E_{16}$ mutations, proteins G3, G5, and G7 demonstrated the lowest activities (Figure 4.6), which may suggest that PUF module replacements from $N_{12}Q_{16}$ to $S_{12}R_{16}$ contribute greater to the destabilization of the PUF domain, compared to $C_{12}Q_{16}$ to $S_{12}R_{16}$ mutagenesis, due to yet unknown reasons. Finally, we observed that mutagenesis of most PUF repeats to contain $S_{12}R_{16}$ leads to a considerable decrease in activity compared to a WT repeat in the corresponding position (Figure 4.7). Protein C2 exhibited the highest activity among others; however, this protein has a single module replaced, in contrast to others, which have two modules replaced (Table 4.3). These results imply that mutagenesis

of PUF proteins to contain $S_{12}R_{16}$ modules destabilize the protein or reduce the affinity to RNA more compared to mutagenesis to other residues.

4.3 Conclusions

In this study, we established a toolkit for rapid engineering of designer RBPs that can be used for manipulation of endogenous genes. This approach should allow greater flexibility and speed at creating PUF domains with user-defined specificities and thus facilitate the use of PUF-based designer RBPs as a tool in research and therapeutics. We demonstrated that PUF mutants with as many as 8/8 mutant modules can be cloned with high efficiency and that the resulting proteins retain their specificity and high affinity to their cognate RNA *in vitro*. This result is consistent with the study by Cheong and Hall[16] and confirms the RNA recognition code and modularity of PUF domains. We envision the future development of PUF-based RBPs with various functionalities that could range from endogenous RNA splicing, imaging, and localization to various base modifications. The development of rapid assembly tools for PUF specificity engineering, as presented in this work, could play a critical role in facilitating and enhancing these endeavors.

4.4 Materials and Methods

All the chemicals and solutions were purchased from Fisher Scientific (Pittsburgh, PA), unless noted otherwise. Oligonucleotides were purchased from Integrated DNA Technologies. All the enzymes were purchased from New England Biolabs (NEB), unless noted otherwise.

4.4.1 Cloning of the PUF Module Library

The intermediate vector, pChlr-pNEB193, was created by replacing the original *amp* resistance gene in pNEB193 (NEB) plasmid with the *cam* resistance gene from pACYC (New England Biolabs) plasmid by Gibson Assembly (GA). Plasmid pTYB3-PUM1-HD [GenBank:D43951] was a gift of Dr. Traci M. Tanaka Hall (Addgene plasmid 17543). The WT GG assembly modules were amplified from pTYB3-PUM1-HD and inserted in the SacI and HindIII sites of the intermediate vector. The amplification primers also contained BsaI sites for subsequent GG cloning. Some of the modules' 5′ and 3′ ends were modified with silent mutations for the creation of non-overlapping BsaI overhangs. The mutant GG modules were created by GA from the corresponding WT modules.

The receiving vector pET28-GG-PUF was GA-cloned from the following fragments: pET28a (Novagen) digested with NdeI and SalI, flanking PUM-HD repeats R1′ and R8′ amplified from pTYB3-PUM1-HD, and *lacZ*α amplified from pNEB193. BsaI sites were introduced next to R1′ and R8′ for GG cloning of PUM-HD. The receiving vector pCMV-TTP(C147R)-GG-PUF was cloned in two steps. First, pCMV-TTP-GG-PUF was created by replacing PUM-HD with a *lacZ*α gene flanked by BsaI sites and removal of 3 existing BsaI sites in the pCMV-TTP(WT)-PUM-HD (see the effector plasmids section). Next, pCMV-TTP(C147R)-GG-PUF was GA-cloned by replacing Flag with 3xFlag and mutating the C147R of TTP in the plasmid pCMV-TTP-GG-PUF.

All the plasmids that comprise the PUF Module library have been deposited to Addgene plasmid repository, and the DNA sequences of all the modules and receiving vectors can be found there.

4.4.2 Golden Gate Assembly of Mutant PUFs

His-tagged PUF constructs for *E. coli* expression were assembled in pET28-GG-PUF. Receiving vector (50 ng) and 8 modules of choice (75 ng each) were combined with 1 µl T4 DNA ligase and 1 µl BsaI in 10 µl 1x T4 DNA ligase buffer. The reactions were cycled 10 times for 5 min at 37°C and 10 min at 16°C, and a final incubation of 15 min at 37°C. TOP10 *E. coli* cells (Invitrogen) were then transformed with the cloning reactions and plated on LB plates (Cell Media Facility, UIUC) with *kan* selection, and supplemented with 10 µl 0.4 M IPTG (GoldBio) and 40 µl 20 mg/ml Bluo-Gal (Invitrogen) for blue-white screening. All the plasmids for *E. coli* expression were purified using Qiagen Qiaprep Spin Miniprep kit.

Receiving vector pACT2-GG-PUF was modified from the plasmid pACT2 (kind gift of Dr. Marvin P. Wickens, University of Wisconsin-Madison) to include two *Bsa*I sites flanked by 1' and 8' regions of the PUF sequence. The Act2-PUF fusion constructs were cloned by the Golden Gate cloning method as previously using pACT2-GG-PUF as a receiving vector and 8 PUF modules of interest from the library.

4.4.3 Cloning of transcription vectors for Yeast 3 hybrid Assay

For the preparation of transcription vectors for Act2-PUF fusion proteins in groups S and M, the DNA fragments encoding target RNA sequences were designed to contain *Xma*I-compatible overhangs at both ends. The fragments were synthesized

at IDT DNA, the two strands annealed in TE buffer with 50 mM NaCl by heating up the mixture to 90C for 5 min and slowly cooling down to 4C. The ds oligos were then phosphorylated using T4 polynucleotide kinase (NEB) as per manufacturer's instructions. The plasmid p3HR2 (kind gift of Dr. Marvin P. Wickens, University of Wisconsin-Madison) was digested with *Xma*I and ligated with the DNA inserts containing the sequences encoding RNA sequences. For the preparation of transcription vectors for Act2-PUF fusion proteins in groups S and M, the receiving vector p3HR2-insert was constructed to contain an additional restriction site, *Hind*III, for one-directional insertion of fragments. To this end, the insert was designed to contain *Xma*I, *Hind*III and *Age*I sites. The p3HR2 plasmid was digested with *Xma*I, while the insert was digested with *Xma*I and *Age*I. The *Age*I restriction site was destroyed by ligation. Next, the PUF-binding sites were cloned into the p3HR2-insert plasmid using *Xma*I and *Hind*III restriction sites.

4.4.4 Protein Expression and Purification

His-tagged recombinant PUF proteins were expressed in *E. coli* strain BL21(DE3) (Novagen). The transformed BL21 cultures were grown in 100 ml LB until they reached an OD_{600} of 0.8, induced with 0.4 mM IPTG and expressed at 18°C, 250 RPM overnight. Bacterial pellets were resuspended in lysis buffer (25mM Tris-HCl pH 7.5, 0.3 M NaCl, 0.5% Triton (Bio-Rad), 5% glycerol (Sigma), 1 mg/ml lysozyme (Sigma), and 0.002 U/µl DNase I) and lysed by sonication. The proteins were purified using Talon Spin Columns (Clontech), according to manufacturer's instructions. The eluted protein was flash-frozen in 25% glycerol in dry ice and stored in aliquots at -80°C.

4.4.5 Bradford Assay

The protein concentrations were measured by mixing 20 µl of appropriately diluted protein solution with 980 µl of Coomassie Plus Protein Assay Reagent (Thermo Scientific) in 1 ml cuvette. Quick Start Bovine Serum Albumine Standard Set (Bio-Rad) was used to build a protein standard curve. A_{595} was measured 5 min later using Nanodrop.

4.4.6 Fluorescence Polarization

RNA oligomers were modified with 6-carboxyfluorescein (IDT) at the 5′-end. To determine active protein fractions, we performed saturation assays for PUF proteins against their cognate RNA (a representative saturation curve is shown in Fig. S2a). High concentrations (100 nM) of RNA oligomers in fluorescence anisotropy buffer (20 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 50 mM KCl, 0.1 mg/ml BSA) were mixed with various protein concentrations (determined by Bradford assay), and 200 µl protein-RNA mixtures were assayed (for fluorescence polarization measurements, see below) in black 96-well plates (Corning). The stoichiometric points were used to estimate the active protein fractions, which were determined to be 31% for PUF(WT), 30% for PUF(S2), 30% for PUF(S4), 33% for PUF(S6), and 29% for PUF(S6). Corrected active protein concentrations were used in the subsequent binding curves for the determination of the dissociation constants K_D, where RNA oligomers (250 pM RNA for PUF(WT) and 1 nM RNA for PUF(S2)-PUF(S8)) in the fluorescence anisotropy buffer were mixed with various protein concentrations, and duplicates of 200 µl protein-RNA mixtures were assayed.

Fluorescence polarization measurements were taken on Tecan Infinite 200Pro using excitation and emission wavelengths of 485 nm and 535 nm, respectively. The fluorescence polarization values were converted to fluorescence anisotropy values using Equation 1, where A is anisotropy and P is polarization. The K_D was calculated by curve fitting on Origin 8.5 using Equation 2, where A is observed anisotropy value, A_f is anisotropy of free RNA, A_b is anisotropy of bound RNA, L_T is total ligand (RNA) concentration, and R_T is total receptor (protein) concentration.

$$A = \frac{2P}{3-P} \qquad (1)$$

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

4.4.7 Yeast Heat Shock Transformation

The YBZ-1 *Saccharomyces cerevisiae* strain (kind gift of Dr. Marvin P. Wickens, University of Wisconsin-Madison) was transformed with plasmids pACT2-PUF and p3HR2-target plasmids containing cognate PUF-binding sites using the heat shock protocol. To this end, 50 ml YPAD medium was inoculated with a 0.5 ml overnight yeast culture and shaken at 30 °C and 250 rpm for 4–5 h until OD_{600} reached 0.8–1.0. Yeast cells were harvested by centrifugation at 4°C and 4000 rpm for 5 min. The supernatant was discarded and the cell pellet was washed with 25 ml ice-cold Milli-Q double-deionized water, followed by another wash with cold 1 ml water, and finally resuspended in 500 µl water, aliquoted to 50 µl samples, after which the cells were centrifuged again and the supernatant removed. The transformation mixture was prepared by adding 240 µl PEG (50% w/v), 50 µl single stranded salmon sperm

DNA (prepared by boiling and immediately chilling on ice), 1 μg of each plasmid DNA in a total of 34 μl aqueous solution, and 36 μl 1.0 M LiAc in the order listed on top of the cell pellets. The mixtures were vortexed and heat-shocked at 42 °C for 20 minutes. The cells were centrifuged at 8,000 rpm for 15 seconds, the transformation mix was removed and replaced with 100 μl double-distilled water. The cells were spread on SC-Ura plates for the selection of the transformants, and the plates were incubated at 30 °C for 2–4 days until colonies appeared.

4.4.8 Yeast 3-Hybrid Assay (with Stella Xinzi Wu at Zhao Lab)

Three colonies were randomly picked into 2 ml SC-Ura liquid medium and grown for 1 day at 30 °C and 250 rpm, after which the β-galactosidase activity was measured using the Beta-Glo Assay System (Promega) as per manufacturer's instructions. Briefly, cells were diluted 1/40 in the medium and grown for 4 hours at 30C until they reached an A600 of 0.1. The cell samples (50 ul) were mixed with an equal volume of Beta-Glo Reagent in white 96-well plates (Greiner Bio One), and the mixtures were incubated at room temperature in the dark for 60 minutes. The samples were then read in a luminometer (Tecan Infinite 200Pro) for one second. The luminescent signal values were normalized with optical density (OD_{600}) values (Nanodrop 2000).

4.5 References

(1) Glisovic, T.; Bachorik, J. L.; Yong, J.; Dreyfuss, G., RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters* **2008,** *582*, 1977-86.
(2) Pichon, X.; Wilson, L. A.; Stoneley, M.; Bastide, A.; King, H. A.; Somers, J.; Willis, A. E., RNA-binding protein/RNA element interactions and the control of translation. *Current protein and peptide science* **2012,** *13*, 294-304.

(3) Hoskins, A. A.; Moore, M. J., The spliceosome: A flexible, reversible macromolecular machine. *Trends in biochemical sciences* **2012,** *37*, 179-88.
(4) Jansen, R. P.; Niessing, D., Assembly of mRNA-protein complexes for directional mRNA transport in eukaryotes - an overview. *Current protein and peptide science* **2012,** *13*, 284-93.
(5) Lunde, B. M.; Moore, C.; Varani, G., RNA-binding proteins: Modular design for efficient function. *Nature reviews.* **2007,** *8*, 479-90.
(6) Filipovska, A.; Rackham, O., Designer RNA-binding proteins: New tools for manipulating the transcriptome. *RNA biology* **2011,** *8*, 978-83.
(7) Mackay, J. P.; Font, J.; Segal, D. J., The prospects for designer single-stranded rna-binding proteins. *Nature structural and molecular biology* **2011,** *18*, 256-61.
(8) De Gregorio, E.; Preiss, T.; Hentze, M. W., Translation driven by an eif4g core domain in vivo. *EMBO journal* **1999,** *18*, 4865-74.
(9) Coller, J. M.; Gray, N. K.; Wickens, M. P., mRNA stabilization by poly(a) binding protein is independent of poly(a) and requires translation. *Genes and development* **1998,** *12*, 3226-35.
(10) Graveley, B. R.; Maniatis, T., Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. *Molecular cell* **1998,** *1*, 765-71.
(11) Long, R. M.; Gu, W.; Lorimer, E.; Singer, R. H.; Chartrand, P., She2p is a novel RNA-binding protein that recruits the Myo4p-She3p complex to Ash1 mRNA. *EMBO journal* **2000,** *19*, 6592-601.
(12) Coller, J.; Wickens, M., Tethered function assays: An adaptable approach to study RNA regulatory proteins. *Methods in enzymology* **2007,** *429*, 299-321.
(13) Barkan, A.; Rojas, M.; Fujii, S.; Yap, A.; Chong, Y. S.; Bond, C. S.; Small, I., A combinatorial amino acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS genetics* **2012,** *8*, DOI: 10.1371/journal.pgen.1002910.
(14) Yagi, Y.; Hayashi, S.; Kobayashi, K.; Hirayama, T.; Nakamura, T., Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. *PLoS ONE* **2013,** *8*, DOI: 10.1371/journal.pone.0057286.
(15) Wang, X.; McLachlan, J.; Zamore, P. D.; Hall, T. M., Modular recognition of RNA by a human pumilio-homology domain. *Cell* **2002,** *110*, 501-12.
(16) Cheong, C. G.; Hall, T. M., Engineering RNA sequence specificity of pumilio repeats. *Proceedings of the national academy of sciences of the United States of America* **2006,** *103*, 13635-9.
(17) Filipovska, A.; Razif, M. F.; Nygard, K. K.; Rackham, O., A universal code for RNA recognition by PUF proteins. *Nature chemical biology* **2011,** *7*, 425-7.
(18) Dong, S.; Wang, Y.; Cassidy-Amstutz, C.; Lu, G.; Bigler, R.; Jezyk, M. R.; Li, C.; Hall, T. M.; Wang, Z., Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. *Journal of biological chemistry* **2011,** *286*, 26732-42.
(19) Quenault, T.; Lithgow, T.; Traven, A., PUF proteins: Repression, activation and mRNA localization. *Trends in cell biology* **2011,** *21*, 104-12.
(20) Edwards, T. A.; Pyle, S. E.; Wharton, R. P.; Aggarwal, A. K., Structure of Pumilio reveals similarity between RNA and peptide binding motifs. *Cell* **2001,** *105*, 281-9.

(21) Wang, X.; Zamore, P. D.; Hall, T. M., Crystal structure of a Pumilio homology domain. *Molecular cell* **2001,** *7*, 855-65.
(22) Opperman, L.; Hook, B.; DeFino, M.; Bernstein, D. S.; Wickens, M., A single spacer nucleotide determines the specificities of two mRNA regulatory proteins. *Nature Structural and molecular biology* **2005,** *12*, 945-51.
(23) Cooke, A.; Prigge, A.; Opperman, L.; Wickens, M., Targeted translational regulation using the PUF protein family scaffold. *Proceedings of the national academy of sciences of the United States of America* **2011,** *108*, 15870-5.
(24) Choudhury, R.; Tsai, Y. S.; Dominguez, D.; Wang, Y.; Wang, Z., Engineering RNA endonucleases with customized sequence specificities. *Nature communications* **2012,** *3*, DOI: 10.1038/ncomms2154.
(25) Wang, Y.; Cheong, C. G.; Hall, T. M.; Wang, Z., Engineering splicing factors with designed specificities. *Nature methods* **2009,** *6*, 825-30.
(26) Ozawa, T.; Natori, Y.; Sato, M.; Umezawa, Y., Imaging dynamics of endogenous mitochondrial RNA in single living cells. *Nature methods* **2007,** *4*, 413-9.
(27) Tilsner, J.; Linnik, O.; Christensen, N. M.; Bell, K.; Roberts, I. M.; Lacomme, C.; Oparka, K. J., Live-cell imaging of viral RNA genomes using a Pumilio-based reporter. *Plant journal* **2009,** *57*, 758-70.
(28) Yamada, T.; Yoshimura, H.; Inaguma, A.; Ozawa, T., Visualization of nonengineered single mRNAs in living cells using genetically encoded fluorescent probes. *Analytical chemistry* **2011,** *83*, 5708-14.
(29) Engler, C.; Kandzia, R.; Marillonnet, S., A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **2008,** *3*, DOI: 10.1371/journal.pone.0003647.
(30) Koh, Y. Y.; Wang, Y.; Qiu, C.; Opperman, L.; Gross, L.; Tanaka Hall, T. M.; Wickens, M., Stacking interactions in PUF-RNA complexes. *RNA* **2011,** *17*, 718-27.
(31) Campbell, Z. T.; Valley, C. T.; Wickens, M., A protein-RNA specificity code enables targeted activation of an endogenous human transcript. *Nature structural and molecular biology* **2014,** *21*, 732-8.
(32) SenGupta, D. J.; Zhang, B.; Kraemer, B.; Pochart, P.; Fields, S.; Wickens, M., A three-hybrid system to detect RNA-protein interactions *in vivo*. *Proceedings of the national academy of sciences of the United States of America* **1996,** *93*, 8496-501.
(33) Bernstein, D. S.; Buter, N.; Stumpf, C.; Wickens, M., Analyzing mRNA-protein complexes using a yeast three-hybrid system. *Methods* **2002,** *26*, 123-41.

Chapter 5. Engineering of TPUF, or PUF-Repressor

5.1 Introduction

The ability to specifically and potently knock down the expression of specific genes could be a powerful tool both in basic research and therapeutics. RNA interference (RNAi) technology has been a remarkable advancement of the past decade that has enabled such selective gene silencing. The technology has been successfully used to silence numerous disease-causing genes for the treatment of viral diseases, neurodegenerative disorders, and cancers[2-3]. However, for the safe and effective application in the clinic, the technology still has a few hurdles to overcome, which include off-target effects, toxicity due to saturation of endogenous RNAi pathways, rapid degradation by nucleases, and induction of interferon 1 responses. Although novel strategies are being developed to overcome these limitations[2-3], developing alternative tools that could be used in lieu of or in combination with the RNAi technology would allow greater flexibility in the field.

RNA-binding proteins with programmable specificity for post-transcriptional regulation of gene expression could become a powerful alternative or complementation to the RNAi technology. Targeted translation regulation using the PUF domain as a scaffold was originally reported by Cooke *et al.*, who showed that *Xenopus* translational repressor CAF1 linked to a PUF domain could sequence-specifically promote mRNA deadenylation and corresponding reporter gene repression in *Xenopus* oocytes[4]; however, repression of an endogenous gene was not reported. During the engineering work of our PUF-repressor, Choudhury *et al.*

reported a synthetic endoribonuclease by fusing the PUF domain to non-specific endonuclease PIN[5], and showed that they were able to sequence-specifically cut target mRNA and silence the corresponding endogenous gene in *E. coli* and a mitochondrially-encoded gene in human cells. While comparable in its ability to sequence-specifically knock down a gene, the PUF-based repressor engineered in our lab functions via a different mechanism (see later). Thus, ours and previously reported engineering examples cross-corroborate the feasibility and efficacy of using PUF as a scaffold for modular construction of post-transcriptional gene regulation devices.

As a proof-of-concept engineering of PUF-based chimeric proteins, we fused the RNA-binding PUM-HD of PUM1 to known post-transcriptional repressor domains. For the candidates of repression domains to be fused to the PUF domain, we considered human tristetraprolin (TTP) and human up-frameshift protein 1 (UPF1). TTP is known to bind to the AU-rich elements (AREs) in the 3′-untranslated region (UTR) of the target genes and promote mRNA degradation by recruiting enzymes involved in deadenylation[6-8] and decapping[8-9]. TTP is also known to promote ARE-dependent translation repression through cooperation with a general translation inhibitor RCK/P54[10], although the details of this mechanism remain unknown. It was reported that the C-terminal domain (CTD) of TTP (residues 176-326 of full-length TTP), which lacks the RNA-binding CCCH-type zinc-finger domain of the full-length TTP, alone could activate mRNA decay[8]. UPF1, on the other hand, targets mRNA for nonsense-mediated decay when bound downstream of a

termination codon[11]. In addition, it is involved in staufen1-mediated mRNA decay[12-13] and replication-dependent histone mRNA decay[14].

In this chapter, we developed a PUF-based repressor for post-transcriptional regulation of gene expression. To this end, we fused the PUM-HD domain to a translational repressor, optimized the construct as well as the reporter gene assay to measure the effectiveness of the chimeric protein, and tested the effectiveness of the engineered protein for the repression of an endogenously expressed gene.

5.2 Results

5.2.1 Choice of a Repression Domain

As an example of engineering a PUF-based post-transcriptional repressor, we fused human TTP and UPF1 to the N-terminus of the WT RNA-binding PUM-HD through a flexible $(G_4S)_3G_4$ linker. We reasoned that utilization of the minimal repression domain of TTP could be advantageous due to the smaller size of the construct and minimization of the interference of the RNA-binding region of TTP. We therefore also tested a fusion construct containing the CTD truncation of TTP. These DNA constructs were co-transfected with firefly luciferase (FL) containing either 5 PUF-binding sites (PBSs) in the 3′ UTR or no PBSs, as well as *Renilla* luciferase (RL) as transfection control. Repression activities of fusion proteins were estimated from the dual luciferase assay, where FL_{PBS5x}/RL activity was normalized to FL/RL. As shown in Figure 5.1, only the TTP-PUF fusion construct exhibited some repression activity and was chosen for further optimization.

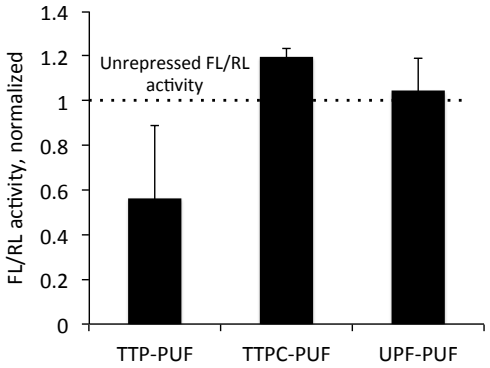


Figure 5.1 Comparison of repression domain activities. Repression domains were N-terminally fused to the PUF domain and the effect on FL activity was measured by dual luciferase assay in HeLa cells. TTPC, C-terminal domain of TTP.

5.2.2 Optimization of the Luciferase Assay

To optimize the dual luciferase assay conditions, 1-10 (PBSs) separated by 6-18 nt were cloned (Figure 5.2a) to the 3′ UTR of FL. As described previously, RL lacking any PBSs was co-transfected with FL as a transfection control. In addition, FL_{Random}, was constructed, where 10 PBSs were replaced with a random sequence of identical length, since it was found that the length of the 3′ UTR also affects apparent FL activity (data not shown). Values of FL_{PBS}/RL normalized to FL_{Random}/RL were reported as "relative FL/RL activity." For reproducibility, we chose to assess the reporter activity in the linear correlation range between transfected FL DNA amount and measured FL activity. We determined that the correlation is linear in the range of 50-400 ng FL DNA in a 24-well plate format (Figure 5.2b). However, when co-transfected with 75 ng of TTP-PUM-HD DNA construct, the upper limit of the linear correlation range was reduced to 200 ng FL DNA (Figure 5.2c). Thus, we used 150 ng of FL and 75 ng of TTP-PUF fusion in our subsequent assays.

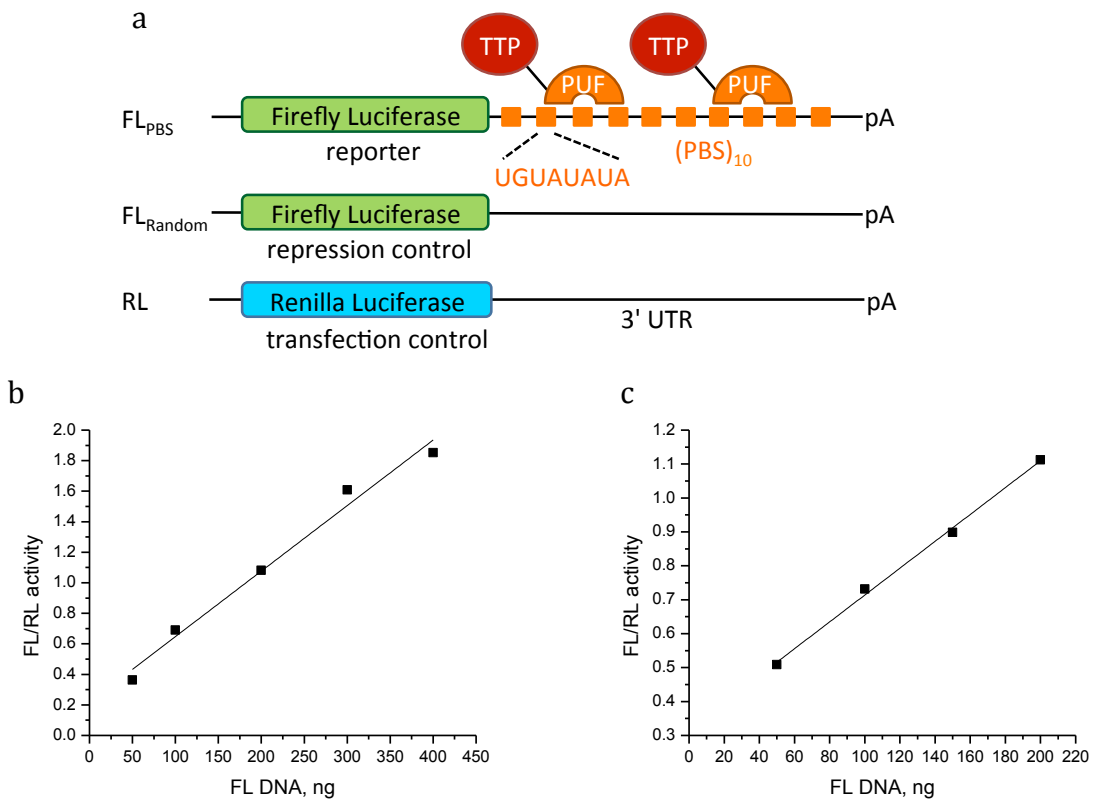


Figure 5.2 Double luciferase assay optimization. **(a)** Schematic of luciferase reporters. Orange boxes, PUF binding sites. **(b and c)** Linearity ranges between FL DNA and FL/RL activity. **(b)** Linearity range when HeLa cells were transfected with indicated amounts of FL DNA, 2 ng RL DNA, and pCMV5 to 500 ng. R^2= 0.995 **(c)** Linearity range when transfected with indicated amounts of FL DNA, 2ng RL DNA, 75 ng TTP(WT)-PUM-HD, and pCMV5 to 500 ng. R^2= 0.997

Abil *et al. J Biol Eng* 2014, 8:7

To determine the optimal amount of PBS in the 3′ UTR of the reporter gene, we tested repression levels of FL containing 1, 3, 5, and 10 PBS in the 3′ UTR of FL (Figure 5.3) and observed repression activities of TPUF(WT) ranging between 31-73%. However, to obtain the greatest dynamic range of our reporter assay, we conducted consequent experiments using 10 PBSs. Although using 10 PBSs is in contrast compared with those using 1 PBS in previous assays[4], it is comparable to

using 5-6 binding sites in tethering assays of TTP and other ARE-mediated decay activation domains[8, 10].

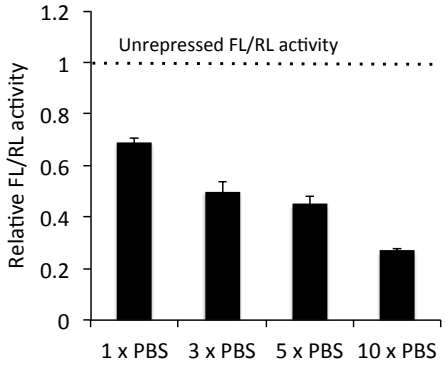


Figure 5.3 Dual luciferase assay showing TTP-PUF$_{WT}$ repression of FL$_{PBS(WT)}$ with increasing number of PBSs in the 3′UTR of the reporter gene. Data are represented as mean fold change relative to cells transfected with FL with no PBS ± SD

Abil *et al. J Biol Eng* 2014, 8:7

5.2.3 Optimization and Analysis of TPUF Activity

We observed that untethered TTP did not significantly repress the FL$_{PBS(WT)}$ activity, whereas PUM-HD$_{WT}$ alone repressed FL$_{PBS(WT)}$ by 20% (Figure 5.4a). The observed weak activity of PUM-HD$_{WT}$ RNA-binding motif alone can be explained by previous findings that the Pumilio RNA-binding domain is also a translational regulator that is capable of recruiting deadenylases to the concave surface of repeats 7 and 8[15]. However, the TTP-PUM-HD$_{WT}$ fusion construct repressed the FL$_{PBS(WT)}$ activity by 80%, thus demonstrating that the fusion construct exhibits both specific RNA binding and high repression activity.

Since TTP is an RBP that binds to the ARE (UUAUUUAUU minimal consensus motif)[16], we further optimized the TTP-PUF construct for decreased ARE binding in order to reduce the interference of TTP towards PUF binding to PBSs. The ARE-binding activity of TTP was shown to be dependent on the presence of two CCCH-type zinc finger motifs in the central region of the protein, and a single cysteine to arginine substitution in either of these motifs abolished binding[16]. We designed another variation of the fusion construct, TTP(C147R)-PUM-HD$_{WT}$ (Figure 5.4b), in

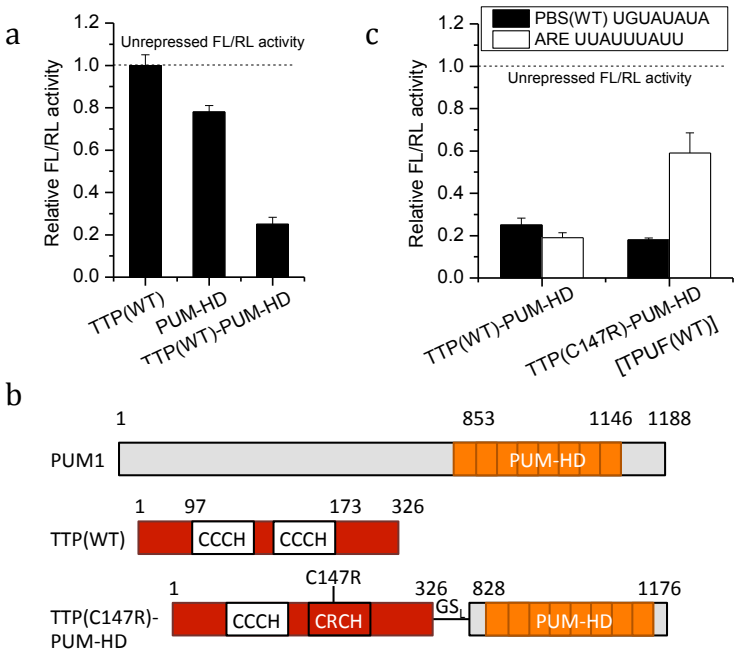


Figure 5.4 Analysis and optimization of TPUF. **(a)** Dual luciferase assay shows that TTP-PUM-HD$_{WT}$ down-regulates FL$_{PBS(WT)}$ expression the most compared with TTP or PUM-HD$_{WT}$ alone. Relative FL$_{Random}$/RL activity is indicated with a dashed line. **(b)** Schematic of full-length PUM1, TTP(WT), and TPUF constructs. CCCH, zinc finger domain; GS$_L$, glycine-serine linker. **(c)** TPUF(WT) binding to ARE is reduced.

Abil *et al. J Biol Eng* 2014, 8:7

which one of the two zinc fingers was disrupted, and compared its activity with TTP(WT)-PUM-HD$_{WT}$ at repressing FL$_{PBS(WT)}$ and FL$_{ARE}$ (10x ARE in the 3′ UTR). As expected, PUM-HD fused to TTP(WT) repressed both FL$_{PBS(WT)}$ and FL$_{ARE}$ (Figure 5.4c). In contrast, fusion to TTP(C147R) decreased repression of FL$_{ARE}$ from 80% to 40%. Thus, TTP(C147R)-PUM-HD$_{WT}$ fusion scaffold, which showed lower interference towards PBS binding, was used in further experiments and was referred to as TPUF.

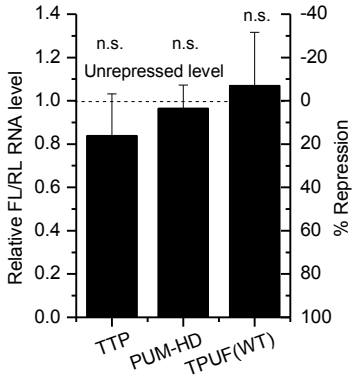


Figure 5.5 Relative levels of FL/RL mRNA, normalized to FL$_{Random}$/RL mRNA in the presence of effectors. Fluorescence RT-PCR data were analyzed by $\Delta\Delta C_T$ method. Data are represented as mean fold change relative to cells transfected with FL$_{Random}$ (dashed line, unrepressed level) ± SD: n.s., not significant (n=3, t test).

Abil *et al. J Biol Eng* 2014, 8:7

To determine if the TPUF(WT) construct functions by promoting degradation of target RNA, we performed real-time PCR (RT-PCR) analysis on the FL reporter, and used RL as the internal control. We did not observe decrease in RNA levels in FL$_{PBS(WT)}$ compared with FL$_{Random}$ in the presence of TPUF(WT) (Figure 5.5). Several FL RT-PCR primer pairs were used, and consistently no RNA destabilization was observed (data not shown). These results are in accordance with a similar TTP

tethering assay[10], where luciferase activity was knocked down despite little RNA destabilization. We suggest that the TPUF constructs function similarly, by promoting translational repression rather than RNA degradation[10].

5.2.4 Validation of the TPUF and Luciferase Assay

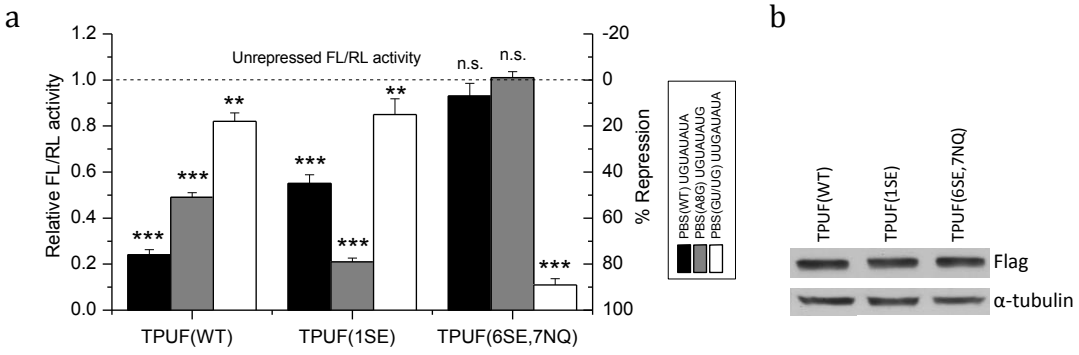


Figure 5.6 Validation of the TPUF and Luciferase Assay. **(a)** Luciferase assay shows predicted specificity of previously reported PUF mutants[1]. Data represent means ± SD: n.s., not significant, **P ≤ 0.01, ***P ≤ 0.001 (n=3, t test). **(b)** Western blot of effector proteins using anti-Flag antibody shows no major difference in the expression. Anti-α–tubulin antibody was used as a loading control.

Abil *et al. J Biol Eng* 2014, 8:7

To further test the design concept, we used our GG toolkit to assemble two previously reported PUF mutants[1]. TPUF(1SE) has repeat 1 replaced for recognition of G8 [PBS(A8G)], and TPUF(6SE,7NQ) has repeats 6 and 7 replaced for the recognition of G and U at positions 2 and 3, respectively. TPUF(WT) and the two PUF mutants exhibited highest repression activities towards their cognate PBSs, with repression levels of 76%, 79%, and 88%, respectively. Compared with cognate PBSs,

we observe a diminished activity towards PBSs with 1 or 2 mismatches, though some cross-reactivity was evident (Figure 5.6). These observations are consistent with cross-reactivity between WT and PUF(1SE) in *in vitro* assays[1] and similar cross-reactivity between WT and other mutant PUF proteins[1, 17] that differ by 1-2 repeats. Western Blotting showed that TPUF(WT) and mutants TPUF(1SE) and TPUF(6SE,7NQ) are expressed at similar levels, thus excluding the possibility of protein abundance variability distorting the observed repression activities. Overall, luciferase repression by the TPUF constructs was sequence-specific, corroborating the validity of the TPUF-reporter system.

5.2.5 Analysis of Mutant TPUF Activities

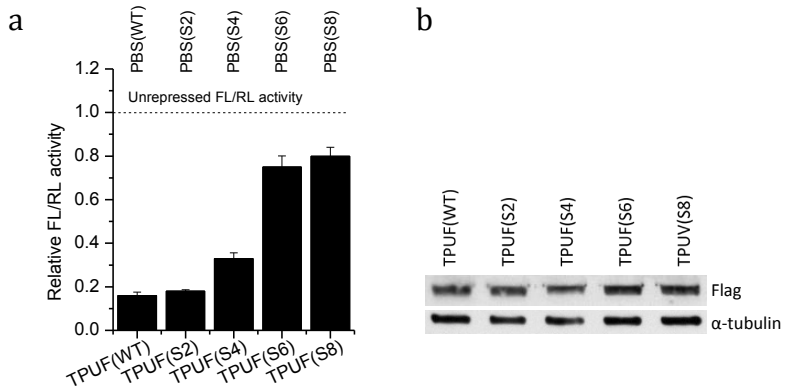


Figure 5.7 Analysis of TPUFs S2-S8. **(a)** Repression by TPUFs S2-S8 in HeLa cells decreases with increased number of mutant modules used. **(b)** Western blotting of 3xFlag-tagged TPUF constructs. α-tubulin was imaged as a loading control.

To further analyze our repressor construct, we used PUF modifications S2-S8 (Table 4.2). TPUF(S2) and TPUF(S4) demonstrated repression levels of 82% and

67% respectively (Figure 5.7a). However, TPUF(S6) and TPUF(S8) showed a significantly lower repression activity (25% and 20%, respectively), despite their high binding affinities *in vitro*. These data could indicate a correlation between *in vitro* binding affinity and protein activity in cell cultures since PUF(S6) and PUF(S8) showed lower *in vitro* RNA binding affinities compared with PUF(S2) and PUF(S4) (Table 4.2). In addition, lower repression activities of TPUFs S6 and S8 could in part be due to decreased solubility. To test this, we analyzed the soluble expression by immunoblotting of the transfected HeLa cell lysates, but did not observe a significant TPUF expression difference (Figure 5.7b). Thus, we observed a correlation between PUF mutation number, *in vitro* binding affinity, and specific TPUF activity, as well as over-sensitivity of TPUF constructs to more than 4 PUF repeat replacements in our luciferase assay.

In order to further verify the functionality of the cloning method and the TPUF platform, we assembled more TPUF constructs with mutant repeats randomly introduced throughout the PUF domain (Figure 5.8a), denoted as TPUFs A-E. Out of 5 TPUF mutants, only TPUF(B) showed low (17%) repression activity towards a cognate PBS, whereas TPUFs A, C-E showed repression activities ranging from 43% to 75% (Figure 5.8b). TPUFs with all 8 replaced modules demonstrated poor repression activities (data not shown), indicating that accumulation of mutations in the PUF domain does not always result in active TPUF proteins *in vivo*.

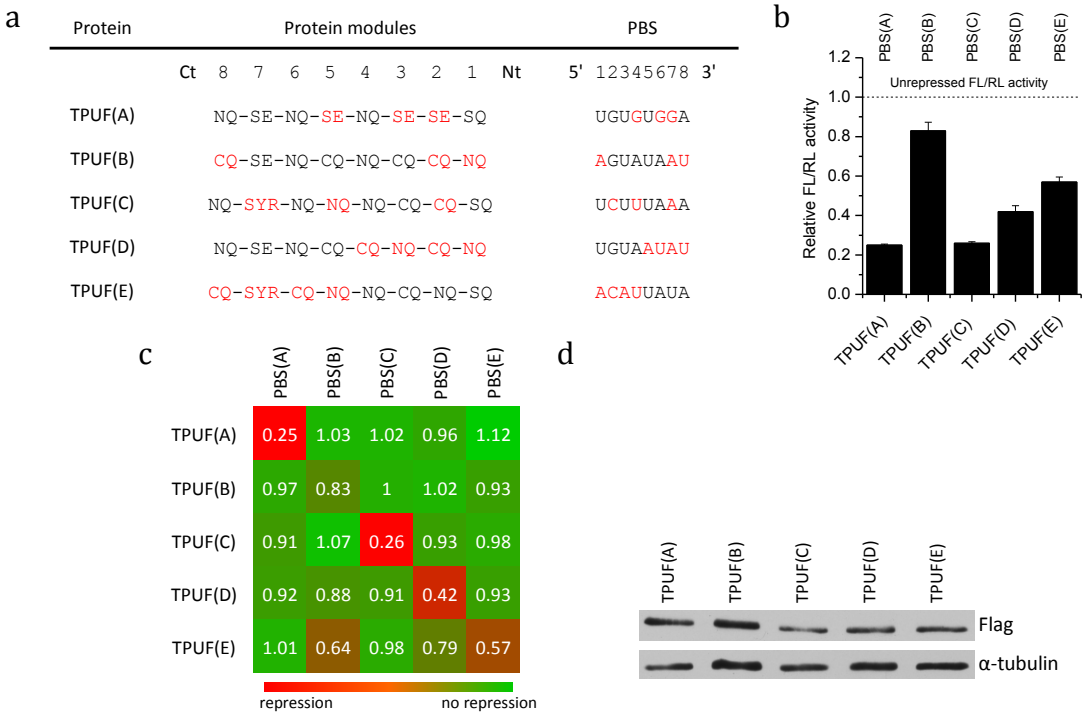


Figure 5.8 Analysis of TPUFs A-E with mutant repeats randomly introduced throughout the PUF domain. **(a)** Mutations and PBSs of TPUFs A-E. Black, WT modules. Red, mutant modules. **(b)** TPUFs A-E repress FL with cognate PBSs. **(c)** Cross-activity of TPUFs A-E. **(d)** Western Blot of TPUFs A-E.

Abil *et al. J Biol Eng* 2014, 8:7

We next analyzed the orthogonality of these TPUF variants to each other by analyzing cross-repression of PBSs among all five protein constructs. We observed that TPUFs A-D had highest repression activities towards their cognate PBSs, and insignificant repression activity towards noncognate PBSs (Figure 5.8c). Only TPUF(E) showed significant repression activity towards noncognate PBS(B). By comparing and contrasting the architecture of TPUFs(A-E), as well as TPUFs (S2-S8), we noticed that TPUFs with the worst performance *in vivo*, such as repression activity or specificity, all had the first module mutated to $8C_{12}Q_{16}$. This finding is not unexpected, since most, if not all, natural PUF proteins recognize RNA containing UGU at the 5 end of the target sequence, corresponding to modules 6, 7, and 8, implying that these modules could be the most important in the recognition of the

consensus and least amenable to mutagenesis. Our data indeed suggest that the modification of the 8th module can often result in the drop of the PUF binding activity.

The difference between TPUF activities in a cell line could also be dependent on the expression and solubility levels of the fusion proteins. We therefore investigated soluble expression levels of the effector proteins mentioned above using Western blotting (Figure 5.8d). We noticed some variability among soluble expression levels of TPUFs(A-E); however, this variability does not seem to be consistent with the observed TPUF activities. In particular, it does not explain the apparent lower activities of TPUF(B) and TPUF(E).

5.2.6 Analysis of Module 4 of TPUF

Most stacking amino acids in the PUM-HD are either tyrosine or arginine, with two exceptions being asparagine in repeat 7 and histidine in repeat 4 (Table 4.1). Although Dong *et al.* reported that module $3S_{12}R_{16}$ requires a tyrosine as a stacking residue in the adjacent module 4 for *in vivo* activity[18], the effect of this mutation on the recognition of the 5th base recognized by module 4 was not investigated. In order to test if all four variants of module $4Y_{13}$ possess similar activity and base preference as the WT $4H_{13}$ modules, we compared TPUF constructs containing $4Y_{13}$ with TPUF constructs containing $4H_{13}$ against all possible 5th base variations in the target sequence. All other modules were unchanged.

The dual luciferase assay was used as described previously to test the activities of TPUF variants. We observed that all TPUFs containing the WT $4H_{13}$ preferred the

5G base. Thus, the $4N_{12}Q_{16}$ and $4C_{12}Q_{16}$ modules wrongly preferred the 5G base, instead of the predicted 5U and 5A, respectively (Figure 5.9a). The $4S_{12}E_{16}$ module correctly recognized 5G, whereas the $4S_{12}R_{16}$ module preferred 5G and 5C equally. In contrast, the $4N_{12}Y_{13}Q_{16}$ module preferred the 5U as predicted by the general code, and module $4C_{12}Y_{13}Q_{16}$ did not prefer any specific base in the position 5 (Figure 5.9b). The bases preferences of modules $4S_{12}Y_{13}E_{16}$ and $4S_{12}Y_{13}R_{16}$ resembled the counterparts with the WT stacking amino acids. We therefore conclude that module 4 with the $4Y_{13}$ mutation has a comparable, if not better (at least for the module $4N_{12}Y_{13}Q_{16}$), specificity compared to the WT histidine at position 13.

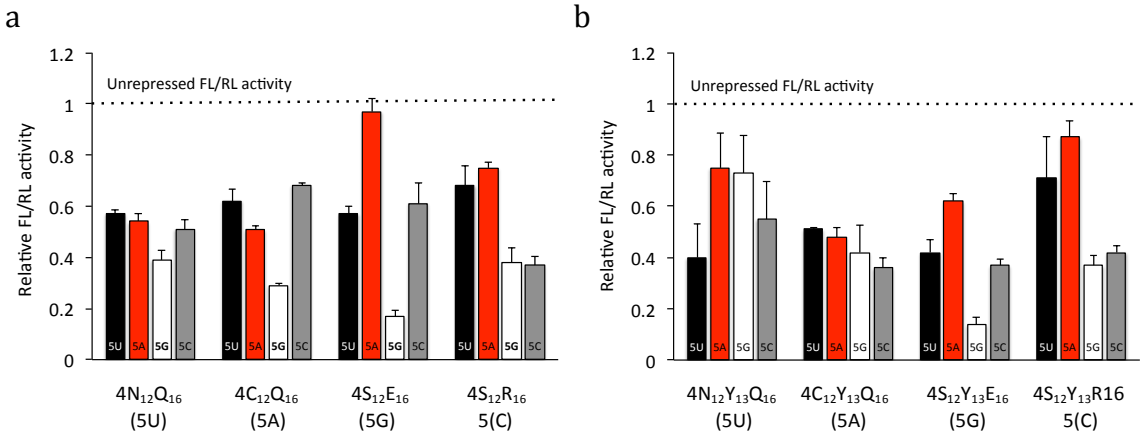


Figure 5.9 Comparison of TPUFs containing modules WT and 4H13Y as the "stacking" amino acid. (a) Base preferences of the $4H_{13}$ modules. (b) Base preferences of the 4H13Y modules.

5.2.7 Endogenous Gene Repression by Designer TPUFs

As a proof of concept for implementing designer RBP for endogenous gene regulation, we designed TPUFs that bind to the 3′ UTR of human vascular endothelial growth factor-A (*VEGFA*) mRNA. VEGFA is one of the central mediators of angiogenesis, and was shown to be overexpressed in many human tumors. VEGFA is up-regulated in these tumors by hypoxic growth conditions that many tumors create[19]. As such, VEGFA is an attractive target for the development of therapeutics to inhibit pathological angiogenesis. We reasoned that for our TPUF repression assay, elevated VEGFA levels are more therapeutically relevant than physiological lower levels, and therefore employed two different strategies to up-regulate endogenous VEGFA expression. In the first strategy, we incubated HEK293 cells with 500 μM $CoCl_2$, an agent that causes conditions mimicking hypoxia[20], and achieved an 8-fold VEGFA induction compared to cells grown in the absence of the agent (Figure 5.10b). In the second strategy, we used a HEK293 cell line in which a small molecule-responsive gene switch for VEGFA expression was stably integrated [21]. Upon induction with 4,4′-dihydroxybenzyl (DHB), endogenous VEGFA was up-regulated 19-fold (Figure 5.10c) [21].

For the post-transcriptional down-regulation of VEGFA, we assembled TPUFs VEGF1, VEGF3, and VEGF7 (Figure 5.10a), which have, respectively, 1, 3, and 7 PBSs in the 3′ UTR of all known transcript variants of *VEGFA* gene. PUF domains in the TPUF VEGF3 and VEGF7 constructs each carry one mutant repeat and were previously reported to be active *in vitro*[1]. TPUF(VEGF1) has repeats 6 and 7 replaced for the recognition of 2A and 3C in the PBS. WT repeat $4N_{12}Q_{16}$ was left

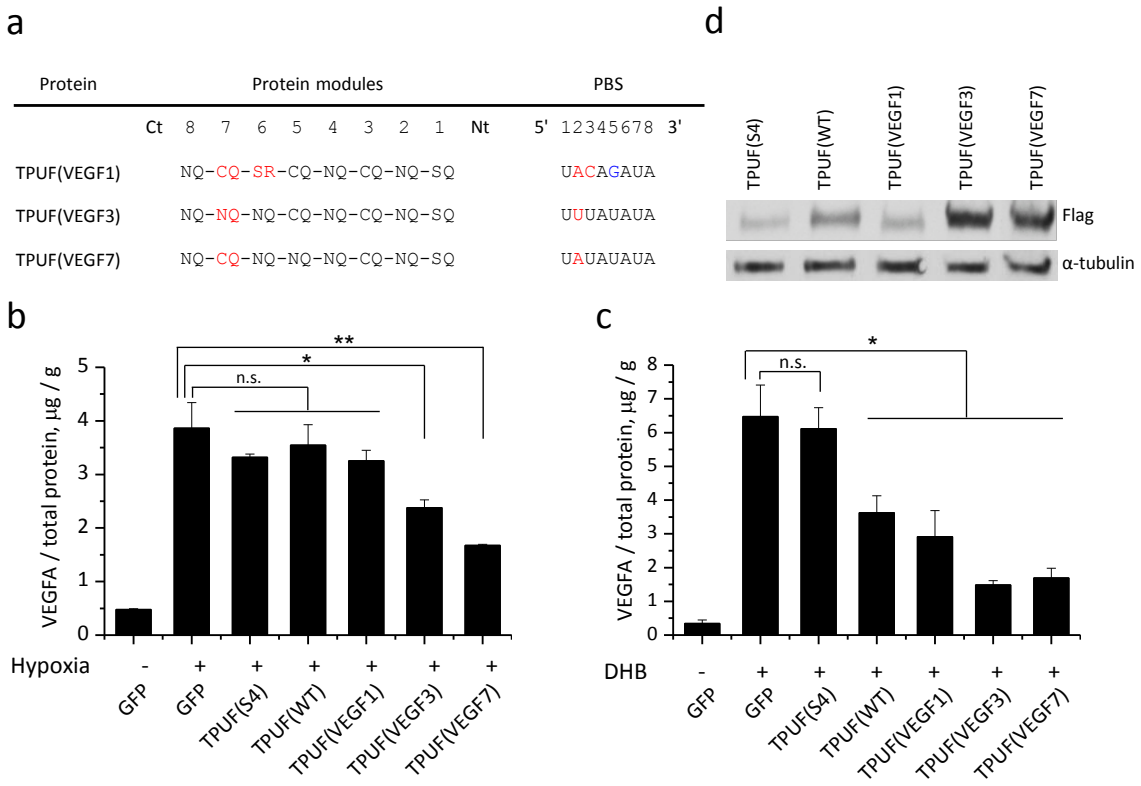


Figure 5.10 TPUF represses endogenously expressed VEGFA gene in HEK293 cell line. **(a)** Mutations and binding sequences of TPUFs designed for VEGFA 3′UTR recognition. Black, WT modules and corresponding RNA bases. Red, mutant modules and corresponding RNA bases. Blue, a mismatch in the recognition sequence. Ct, C-terminus, Nt, N-terminus of the protein. **(b)** Inhibition of hypoxia-induced VEGFA expression by TPUF. Secreted VEGFA levels measure by ELISA were normalized to total protein amounts from lysed cells measured by Bradford Assay. Data represented as mean ± SD: n.s., not significant, *P ≤ 0.05, **P ≤ 0.01, (n=3, t test). **(c)** Inhibition of DHB-induced VEGFA expression by TPUF. Secreted VEGFA levels measure by ELISA were normalized to total protein amounts from lysed cells measured by Bradford Assay. Data represented as mean ± SD: n.s., not significant, *P ≤ 0.05. **(d)** Western blot of effector proteins using anti-Flag antibody. Anti-α–tubulin antibody was used as a loading control.

unchanged for the recognition of 5G (Figure 5.10a, blue), since this repeat is known to be promiscuous[22]. TPUF(S4) was shown to be active for its cognate RNA using the luciferase assay (Figure 5.7a); however, it does not have binding sites in the 3′ UTR of *VEGFA* mRNA, and therefore was not expected to repress the gene. Fortuitously, PBS(WT) is present 3 times in the 3′ UTR of *VEGFA* mRNA, which allowed us to use TPUF(WT) as a reference for assaying activities of mutant TPUFs.

To assay the repression activity, we transiently transfected recombinant HEK293 cells with equimolar amounts of GFP or one of the TPUF constructs, induced the expression of VEGFA 24 hours after transfection, and assayed secreted levels of VEGFA by ELISA 24 hours after induction. VEGFA levels were normalized to total protein concentration, which we assumed to correlate with cell number. Bradford Assay was used to measure the total protein concentrations from cell lysates. We found that VEGFA amounted to less than 1% of total protein concentration by mass (Figure 5.10b and c), hence the variations of VEGFA concentrations in the cell would have no significant effect on total protein concentration. In both induction strategies, the VEGFA levels were not significantly affected by TPUF(S4) compared to samples transfected with GFP, as expected (Figure 5.10 b and c). For the cells treated with $CoCl_2$, only TPUF(VEGF3) and TPUF(VEGF7) significantly repressed VEGFA expression, which was down-regulated by 38% and 57%, respectively (Figure 5.10b). On the other hand, for cells induced with DHB, TPUF(WT) caused considerable (44%) repression of VEGFA, whereas TPUFs VEGF1, VEGF3, and VEGF7 knocked down VEGFA levels by 55%, 77%, and 74%, respectively (Figure 5.10c). We are currently unaware of the reason why cells

exposed to hypoxia were more resistant to VEGFA repression by our TPUF constructs compared with cells that expressed the gene switch, but speculate that it could be due to yet undiscovered transcriptional and/or translational gene regulatory response of cells to hypoxia. The lower TPUF activity at hypoxia might be due to either lower expression of constructs at hypoxic conditions or due to an unknown interference of hypoxia with TTP function. Whatever the reason of this difference might be, we observed a considerable sequence-specific down-regulation of VEGFA in both conditions, and confirmed the efficacy of the TPUF platform in the down-regulation of an endogenous gene in human cells as yet another demonstration of the applicability of PUF-based RBPs at post-transcriptional gene regulation[4-5].

To compare the expression levels of TPUF constructs among each other, we performed a Western blot analysis of the effector proteins under normoxia (Figure 5.10d). We observed a substantially greater soluble expression of TPUF(VEGF3) and TPUF(VEGF7), compared to the other constructs, which might be the reason of these constructs' higher overall activity.

5.3 Discussion

In this work, we demonstrated the implementation of the protein-based post-transcriptional repressor for down-regulation of an endogenous gene using the PUF domain as a scaffold. Due to the development of the GG assembly library, which was described in Chapter 4, the RNA-sequence specificity of the TPUF can be easily engineered for the amino acids of choice at positions 12 and 16 in each of the 8 modules and cloned directly into the TPUF receiving vector, thus enabling easy

modification of the TPUF for down-regulation of mRNA of choice. We showed that the TPUF platform introduced in this work exhibits modular and sequence-specific down-regulation of genes. Analysis of a number of mutant TPUFs showed that modification of the 8^{th} module often results in decreased RNA affinity, poor repression activities, or promiscuity in RNA recognition. We therefore suggest selecting RNA target sequences that would start with a 5′-U (base 1U), which is part of the 5′-UGU consensus that is recognized by most of the known natural PUF domains. In addition, we showed that the H to Y mutation in the 13^{th} position (the "stacking" residue) of the notoriously promiscuous repeat 4 improves the recognition of the U base in the 5^{th} position of the target sequence when the $4N_{12}Q_{16}$ module is used. Similarly, the "wrong" recognition of the G base in the 5^{th} position is diminished by the $4C_{12}Q_{16}$ when the "stacking" residue is tyrosine. Overall, our TPUF platform has been shown to be promising in further engineering of PUF-based translational regulators.

There is still some room for improvement of some TPUF characteristics. For example, TPUF requires 10 PBS in the 3′ UTR of a reporter gene to achieve a robust 60-80 % repression. As a trans-acting repressor, TTP can be more sensitive to interactions with the PUF domain compared with cys-acting repressors, due to blockage of one or more surface areas responsible for the recruitment of cys-acting repressors. Thus, further engineering of TPUF for better architecture can be beneficial for the increase of overall functional activity.

In vivo TPUF activities could also be dependent on various factors, including soluble protein expression; unequal contributions of different repeat-base

interactions to the binding energy of the RNA-protein complexes[1, 23], resulting in different overall binding affinities to cognate RNA; binding to noncognate RNA[24] that would allow sequestration of the protein to non-target RNA *in vivo*; as well as sporadic domain interactions in fusion proteins that would interfere with RNA binding. We cannot yet predict the contribution of each of these factors on PUF activity *in vivo*, and therefore suggest that the effect of these and other factors on engineered PUF activity has to be systematically investigated.

Based on our findings on the repression of VEGFA, we conclude that the implementation of TPUFs for down-regulation of endogenous genes with high efficacy is possible. The VEGFA repression levels that we observed are comparable to those in similar assays conducted with zinc finger transcriptional repressors or small interfering RNA (siRNA). For example, hypoxia-induced VEGFA protein levels in HEK293 cells were knocked down by 74% using an engineered zinc finger transcriptional repressor[25] and VEGFA mRNA was knocked down by 50% by another zinc finger repressor in HEK293 cells grown in normoxia[26]. Similarly, siRNA knocked-down endogenous VEGFA by up to 43% in ID8 cells[27] and up to 71% in hypoxic HEK293 cells[25]. Thus, the TPUF system that we have engineered is as effective as some other existing technologies that are often used to knock down gene expression levels, and can serve as a powerful alternative.

However, factors determining tight binding of Pumilio domains to RNA *in vitro*, as well as factors determining high *in vivo* activities are still largely unknown, and a systematic investigation is needed. *In vitro* RNA binding affinities and *in vivo* activities show only a rough correlation[4, 17], and soluble expression levels do not

always predict activity differences (Figures 5.8d and 5.10d). Therefore, for engineering of TPUFs with novel specificities or PUM-HD-based proteins with novel functionalities, we recommend building a reporter system suitable for the given protein activity, and directly screening the functional efficacy of the assembled PUF variants.

Finally, for certain applications, specific recognition of 8 nt sequences may not be enough for unique recognition in a transcriptome. The specificity can be increased by using a split-protein method, where two PUF domains, each fused to a fragment of a functional domain, can reassemble the functional domain on an RNA scaffold, as was shown for PUF-based fluorescent probe engineering[28-30]. This approach would increase the total number of PUF repeats to 16, thus likely increasing the overall specificity of the system. Alternatively, a GG platform for assembly of single polypeptide PUF domains with more RNA-binding repeats can be developed, as was proposed by Filipovska and coworkers who constructed a PUF domain comprising 16 repeats[31].

5.4 Materials and Methods

5.4.1 Materials

All the chemicals and solutions were purchased from Fisher Scientific (Pittsburgh, PA), unless noted otherwise. Oligonucleotides were purchased from Integrated DNA Technologies (Coralville, IA). All the enzymes were purchased from New England Biolabs (Ipswich, MA), unless noted otherwise.

5.4.2 Construction of Reporter Plasmids

pCMV-Fluc plasmid was created by amplification of the firefly luciferase gene from pGL3 plasmid (Promega) and insertion into SacI and KpnI sites of pCMV5 vector (a gift of Dr. David Russell). All the pCMV-Fluc-10xPBS plasmids, pCMV-Fluc-10xARE, as well as pCMV-Fluc-Random were cloned by primer-extension of 6 primers (Table S3) carrying 10 PBSs and subsequent GA-cloning into PstI and XmaI sites of the pCMV-Fluc plasmid. The 6-18 nt spacers between the 10 PBS in the 3′ UTR of the FL were the same in different FL-PBS sequences, and were designed in such a way to minimize secondary structure formation that would involve these spacers. The pCMV-Fluc-Random was created by replacing all the PBSs in the pCMV-Fluc-10xPBS with 10 different scrambled sequences of 8 nt with approximately 50% GC content. Plasmid pRL-SV40 was a gift of Dr. David J. Shapiro.

5.4.3 Construction of Effector Plasmids

The effector plasmid pCMV-TTP(WT)-PUM-HD was GA-cloned from the following fragments: 2.2 kb and 2.4 kb pCMV5 fragments, GS-PUM-HD amplified from pTYB3-PUM1-HD, and TTP-GS amplified from cDNA (Open Biosystems catalog number MHS4768-99609440 [GenBank: BC009693.1]). pCMV-TTP(WT) and pCMV-PUM-HD have been assembled from the same vector backbone fragments, as well as TTP-stop or Flag-PUM-HD fragments, respectively (for primers, see Table S4).

5.4.4 Golden Gate Assembly of Mutant Effector Plasmids

3xFlag-tagged TPUF constructs for mammalian cell expression were assembled in the pCMV-TTP(C147R)-GG-PUF destination vector. Receiving vector (50 ng) and 8 modules of choice (75 ng each) were combined with 1 µl T4 DNA ligase and 1 µl BsaI in 10 µl 1x T4 DNA ligase buffer. The reactions were cycled 10 times for 5 min at 37°C and 10 min at 16°C, and a final incubation of 15 min at 37°C. TOP10 *E. coli* cells (Invitrogen) were then transformed with the cloning reactions and plated on LB plates (Cell Media Facility, UIUC) with *amp* selection, and supplemented with 10 µl 0.4 M IPTG (GoldBio) and 40 µl 20 mg/ml Bluo-Gal (Invitrogen) for blue-white screening. All the plasmids for mammalian cell expression were purified using Qiagen Plasmid Mini kit.

5.4.5 Cell Line Transfection and Dual Luciferase Assay

Transfection of HeLa cells (ATCC) was performed in triplicates in a 24-well plate format with Fugene-HD transfection reagent (Promega). Transfection mixtures contained 150 ng FL, 2 ng pRL-SV40, and 75 ng TPUF or equimolar amounts of other effector DNA constructs, and empty vector pCMV5 to 500 ng total. Cells were lysed in Passive Lysis Buffer (Promega) 48h after transfection and FL and RL activities were measured in white 96-well plates (Greiner Bio One) using Dual-Glo Luciferase Assay System (Promega) with measurements taken on Analyst HT microplate reader at the High-Throughput Screening Facility at UIUC.

5.4.6 Reverse Transcription Real-Time PCR

Total RNA was isolated from HeLa cells 48 hours after transfection using the RNeasy Mini Kit (Qiagen) following manufacturer's instructions, and DNA was removed from samples with Turbo DNase (Life Technologies). RNA was reverse transcribed into cDNA with ProtoScript First Strand cDNA Synthesis kit (NEB) using the $d(T)_{23}VN$ primer. Reverse transcriptase was omitted in control samples. RT-PCR was performed using Power SYBR Green Master Mix (Life Technologies) with the 7900HT Fast Real-Time PCR System (Applied Biosystems). Reactions were carried out in triplicates in 20 µl reactions with 500 nM of each primer. The primer sequences for FL were 5′-GCGCGGAGGAGTTGTGTTTG and 5′-ATCTTTCCGCCCTTCTTGGC; and for RL 5′-GCAGCATATCTTG AACCATTC and 5′-TTGTACAACGTCAGGTTTACC. ΔΔCT method was used for RNA level analysis, where FL mRNA levels were normalized to RL mRNA, and $FL_{PBS(WT)}$ mRNA levels were normalized to FL_{Random}.

5.4.7 VEGFA Induction and ELISA

For hypoxia-induced VEGFA, HEK293 cells were transfected in a 24-well plate format in triplicates with Fugene-HD. Transfection mixtures contained 500 ng TPUF DNA constructs or 350 ng pmaxGFP (Lonza) and 150 ng pCMV5. The cells were induced with 500 µM $CoCl_2$ 24h after transfection, and the supernatant was collected for assay 24 hours after the induction. For gene-switch-induced VEGFA, HEK293 cell line with retrovirally integrated DHB-inducible V24P-GS60 transcription activator was used[21]. The cells were transfected in a 24-well plate

format in triplicates as above. The cells were induced 24h after transfection with 100 nM DHB, in the presence of *pen/strep* (Gibco). The supernatant was collected 24h after induction and subjected to ELISA. The assay was performed by pre-coating the 96-well clear plate with a goat anti-mouse antibody (Thermo Scientific) at 4°C overnight, and then following the instructions of human VEGF DuoSet kit (R&D Systems). The absorption readings were taken on a SpectraMax 340PC microplate reader. The cell monolayer was saved for Bradford assay.

5.4.8 Bradford Assay

The cell monolayers were lysed using RIPA lysis buffer. The protein concentrations were measured in technical duplicates by mixing 4 µl of cell lysate with 295 µl of Coomassie Plus Protein Assay Reagent (Thermo Scientific) in a 96-well clear plate. Quick Start Bovine Serum Albumine Standard Set (Bio-Rad) was used to build a protein standard curve. A_{595} was measured 5 min later using a SpectraMax 340PC microplate reader. The total protein concentrations measured by Bradford assay were used to normalize the VEGF concentrations.

5.4.9 Western Blotting

V24P-GS60-integrated HEK293 cells as well as HeLa cells were transfected in a 6-well plate format with Fugene-HD and 3 µg of effector plasmid. Cells were lysed using RIPA lysis buffer (Santa Cruz Biotech). The proteins were detected using

mouse anti-Flag and anti-α-tubulin antibodies (GeneScript) and imaged using SuperSignal West Dura chemilluminescent substrate (Thermo Scientific).

5.5 References

(1) Cheong, C. G.; Hall, T. M., Engineering RNA sequence specificity of Pumilio repeats. *Proceedings of the national academy of sciences of the United States of America* **2006,** *103*, 13635-9.
(2) Kim, D. H.; Rossi, J. J., Strategies for silencing human disease using RNA interference. *Nature reviews. Genetics* **2007,** *8*, 173-84.
(3) Sibley, C. R.; Seow, Y.; Wood, M. J., Novel RNA-based strategies for therapeutic gene silencing. *Molecular therapy* **2010,** *18*, 466-76.
(4) Cooke, A.; Prigge, A.; Opperman, L.; Wickens, M., Targeted translational regulation using the PUF protein family scaffold. *Proceedings of the national academy of sciences of the United States of America* **2011,** *108*, 15870-5.
(5) Choudhury, R.; Tsai, Y. S.; Dominguez, D.; Wang, Y.; Wang, Z., Engineering RNA endonucleases with customized sequence specificities. *Nature communications* **2012,** *3*, DOI: 10.1038/ncomms2154.
(6) Marchese, F. P.; Aubareda, A.; Tudor, C.; Saklatvala, J.; Clark, A. R.; Dean, J. L., MAPKAP kinase 2 blocks tristetraprolin-directed mRNA decay by inhibiting Caf1 deadenylase recruitment. *Journal of biological chemistry* **2010,** *285*, 27590-600.
(7) Van Etten, J.; Schagat, T. L.; Hrit, J.; Weidmann, C. A.; Brumbaugh, J.; Coon, J. J.; Goldstrohm, A. C., Human Pumilio proteins recruit multiple deadenylases to efficiently repress messenger RNAs. *Journal of biological chemistry* **2012,** *287*, 36370-83.
(8) Lykke-Andersen, J.; Wagner, E., Recruitment and activation of mRNA decay enzymes by two are-mediated decay activation domains in the proteins TPP and BRF-1. *Genes and development* **2005,** *19*, 351-61.
(9) Fenger-Gron, M.; Fillman, C.; Norrild, B.; Lykke-Andersen, J., Multiple processing body factors and the are binding protein TTP activate mRNA decapping. *Molecular cell* **2005,** *20*, 905-15.
(10) Qi, M. Y.; Wang, Z. Z.; Zhang, Z.; Shao, Q.; Zeng, A.; Li, X. Q.; Li, W. Q.; Wang, C.; Tian, F. J.; Li, Q.; Zou, J.; Qin, Y. W.; Brewer, G.; Huang, S.; Jing, Q., AU-rich-element-dependent translation repression requires the cooperation of tristetraprolin and RCK/p54. *Molecular and cellular biology* **2012,** *32*, 913-28.
(11) Lykke-Andersen, J.; Shu, M. D.; Steitz, J. A., Human UPF proteins target an mRNA for nonsense-mediated decay when bound downstream of a termination codon. *Cell* **2000,** *103*, 1121-31.
(12) Kim, Y. K.; Furic, L.; Desgroseillers, L.; Maquat, L. E., Mammalian Staufen1 recruits UPF1 to specific mRNA 3UTRs so as to elicit mRNA decay. *Cell* **2005,** *120*, 195-208.
(13) Kim, Y. K.; Furic, L.; Parisien, M.; Major, F.; DesGroseillers, L.; Maquat, L. E., Staufen1 regulates diverse classes of mammalian transcripts. *EMBO journal* **2007,** *26*, 2670-81.

(14) Kaygun, H.; Marzluff, W. F., Regulated degradation of replication-dependent histone mRNAs requires both ATR and UPF1. *Nature structural and molecular biology* **2005,** *12*, 794-800.
(15) Wharton, R. P.; Sonoda, J.; Lee, T.; Patterson, M.; Murata, Y., The Pumilio RNA-binding domain is also a translational regulator. *Molecular cell* **1998,** *1*, 863-72.
(16) Lai, W. S.; Carballo, E.; Strum, J. R.; Kennington, E. A.; Phillips, R. S.; Blackshear, P. J., Evidence that tristetraprolin binds to AU-rich elements and promotes the deadenylation and destabilization of tumor necrosis factor alpha mRNA. *Molecular and cellular biology* **1999,** *19*, 4311-23.
(17) Wang, Y.; Cheong, C. G.; Hall, T. M.; Wang, Z., Engineering splicing factors with designed specificities. *Nature methods* **2009,** *6*, 825-30.
(18) Dong, S.; Wang, Y.; Cassidy-Amstutz, C.; Lu, G.; Bigler, R.; Jezyk, M. R.; Li, C.; Hall, T. M.; Wang, Z., Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. *Journal of biological chemistry* **2011,** *286*, 26732-42.
(19) Ferrara, N.; Gerber, H. P.; LeCouter, J., The biology of VEGF and its receptors. *Nature medicine* **2003,** *9*, 669-76.
(20) Ho, V. T.; Bunn, H. F., Effects of transition metals on the expression of the erythropoietin gene: Further evidence that the oxygen sensor is a heme protein. *Biochemical and biophysical research communications* **1996,** *223*, 175-80.
(21) Liang, J.; McLachlan, M. J.; Zhao, H., Orthogonal control of endogenous gene expression in mammalian cells using synthetic ligands. *Biotechnology and bioengineering* **2013,** *110*, 1419-29.
(22) Lu, G.; Hall, T. M., Alternate modes of cognate RNA recognition by human Pumilio proteins. *Structure* **2011,** *19*, 361-7.
(23) Zamore, P. D.; Williamson, J. R.; Lehmann, R., The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins. *RNA* **1997,** *3*, 1421-33.
(24) Gupta, Y. K.; Nair, D. T.; Wharton, R. P.; Aggarwal, A. K., Structures of human Pumilio with noncognate RNAs reveal molecular mechanisms for binding promiscuity. *Structure* **2008,** *16*, 549-57.
(25) Kwon, H. S.; Shin, H. C.; Kim, J. S., Suppression of vascular endothelial growth factor expression at the transcriptional and post-transcriptional levels. *Nucleic acids research* **2005,** *33*, DOI: 10.1093/nar/gni068.
(26) Snowden, A. W.; Zhang, L.; Urnov, F.; Dent, C.; Jouvenot, Y.; Zhong, X.; Rebar, E. J.; Jamieson, A. C.; Zhang, H. S.; Tan, S.; Case, C. C.; Pabo, C. O.; Wolffe, A. P.; Gregory, P. D., Repression of vascular endothelial growth factor a in glioblastoma cells using engineered zinc finger transcription factors. *Cancer research* **2003,** *63*, 8968-76.
(27) Zhang, L.; Yang, N.; Mohamed-Hadley, A.; Rubin, S. C.; Coukos, G., Vector-based RNAi, a novel tool for isoform-specific knock-down of VEGF and anti-angiogenesis gene therapy of cancer. *Biochemical and biophysical research communications* **2003,** *303*, 1169-78.
(28) Ozawa, T.; Natori, Y.; Sato, M.; Umezawa, Y., Imaging dynamics of endogenous mitochondrial RNA in single living cells. *Nature methods* **2007,** *4*, 413-9.

(29) Tilsner, J.; Linnik, O.; Christensen, N. M.; Bell, K.; Roberts, I. M.; Lacomme, C.; Oparka, K. J., Live-cell imaging of viral RNA genomes using a Pumilio-based reporter. *Plant journal* **2009,** *57*, 758-70.
(30) Yamada, T.; Yoshimura, H.; Inaguma, A.; Ozawa, T., Visualization of nonengineered single mRNAs in living cells using genetically encoded fluorescent probes. *Analytical chemistry* **2011,** *83*, 5708-14.
(31) Filipovska, A.; Razif, M. F.; Nygard, K. K.; Rackham, O., A universal code for RNA recognition by PUF proteins. *Nature chemical biology* **2011,** *7*, 425-7.

Chapter 6. Engineering of a PUF-Motor[4]

6.1 Introduction

Subcellular mRNA localization is increasingly believed to play an important role in posttranscriptional regulation. Indeed, a broad role for RNA localization was suggested by the revelation that up to 70% of endogenous RNA was differentially localized in *Drosophila* embryos[1]. Differential mRNA localization and local mRNA translation were found to be implicated in diverse cellular processes including asymmetric cell division, embryogenesis, and cell migration[2-3]. In addition, local mRNA translation is likewise important in highly polarized cells like neurons. In particular, the importance of differential mRNA localization and local protein synthesis was reported in a number neurological processes including growth cone chemotropic response[4-6], synapse formation and plasticity[7-8], transmitter biogenesis[9-10], cell survival and axon maintenance[11-13], and axon regeneration[14-15]. Notably, mislocalization of mRNA in neurites has been linked with neurodegenerative disorders such as fragile X syndrome, spinal muscular atrophy, and amyotrophic lateral sclerosis[16-17].

Although subcellular mRNA localization was in some cases reported to be established through localized protection from degradation[18-19] and localized anchorage[20], active transport on the cytoskeleton using molecular motors is

[4] This work was done in collaboration with Dr. Casper C. Hoogenraad and Dr. Laura F. Gumy at Utrecht University, the Netherlands.

regarded as the major pathway of mRNA transport[2]. In the case of active transport of mRNA, ribonucleoprotein complexes are loaded on motor proteins belonging to all three motor families: myosins, kinesins, and dyneins[21-22]. Kinesins usually transport cellular cargos, such as RNAs, vesicles, and organells, to the (+) end of microtubules[23], whereas dyneins transport cargos to the (-) end of the microtubules[24]. Meanwhile, myosins transport cargos, including RNAs, along the actin filaments of the cytoskeleton[25]. It is generally believed that linking RNA cargo to cellular motors occurs through an interaction between RNA-binding proteins (RBPs), and additional adapter proteins that recruit motors to the RBPs[21]. Although in most cases the identities of these RBPs and adapters remain unknown, some well studied RBPs-adapter complexes include She2p and She3p in budding yeast[26-27], as well as egalitarian and bicaudal-D (BicD) in *Drosophila*[28].

The development of synthetic devices capable of controlling endogenous mRNA transport would contribute to the investigation of significance of subcellular mRNA localization. Furthermore, it would expand our capability in regulating RNA metabolism, with potential future applications in synthetic biology and neurobiology. Our current ability to manipulate mRNA transport and/or local mRNA requires prior tagging of the RNA of interest[27, 29-31]. In order to target unaltered endogenous RNA targets, devices would require RBPs with reprogrammable specificity, such that the RBP could be predictably engineered for specificity for every new RNA target. This criterion is satisfied by the Pumilio and fem3-binding factor (PUF) homology proteins, which are repeat proteins that bind RNA in a modular 1:1 repeat : base fashion[32-33].

The crystal structure of the PUF domain from the human Pumilio1 protein (PUM1) bound to Nanos-response element (NRE) RNA revealed that amino acid residues at positions 12 and 16 in each PUF repeat interact with a single RNA base, and suggested a code for recognition of RNA bases adenine, uracil, and guanine by PUF proteins[32]. Amino acid residue at position 13 of each repeat is sandwiched between two adjacent RNA bases in a stacking interaction[32]. Mutagenesis of these key amino acids according to the code was shown to predictably alter the RNA-binding specificity of PUF[33]. The development of artificial PUF repeats for the recognition of cytosine expanded our ability to engineer PUF specificity to almost any sequence of 8 RNA bases[34-35].

In this work, we created a device that allows directional transport of mRNA in eukaryotes. The device consists of two components: a transporter, which provides a directional movement, and a component containing an RBP. The transporter is tagged with an FRB domain, whereas the RBP engineered from the PUF scaffold is tagged with FKBP. These two domains heterodimerize upon addition of rapamycin analog (AP21967, henceforth called rapalog)[36] (Figure 6.1b). Our transporters come in two flavors: Kif5b, which is a cellular motor that transports the cargo molecules to the (+) end of the microtubules located in the cell periphery, and BicD, which is an adaptor protein that interacts with cellular motor dynein and transports the cargos to the (-) end of microtubules located in the centrosome. We show here that our PUF-motor system is capable of sequence-specific recognition and transport of mRNA in cancerous mammalian cell cultures as well as in hippocampal neurons.

6.2 Results and Discussion

6.2.1 Engineering of the PUF-Motor System

In order to ensure motility function to our device, we implemented the proteins previously shown to exhibit directional transport on microtubules in cultured mammalian cells. Thus, for the minus end-directed transport, we utilized the N-terminal domain of mouse BicD2 protein (amino acid residues 1-594), which was shown to be sufficient for the interaction with cytoplasmic dynein and induction of transport[37-38]. For the plus end-directed transport, we utilized a truncated human kinesin (amino acid residues 1-807) without the tail domain, which was shown to be sufficient for cargo transport[39]. Both of the constructs were fused to a single FRB domain via a GGGS linker.

The RNA-binding component of our system consists of two PUF domains, enhanced green fluorescent protein (eGFP) for tracking the localization of the constructs, and 1 or 2 consecutive FKBP domains, all fused to a single polypeptide chain with GGGS linkers. In order to minimize the unintended interaction of the PUF with the endogenous sites in the human transcriptome, we used a mutant version of the PUF domain from human PUM1 protein, where repeats 6 and 7 were mutated so that the PUF recognizes the RNA sequence 5′-UUGAUAUA-3′ (see Chapter 5). We used two PUF domains and in some cases two FKBP domains with the intention of maximizing the avidity between PUF and RNA, as well as PUF and the transporter proteins (Figure 6.1a). Since a polypeptide chain consisting of 4-5 domains linked by flexible linkers might assume insoluble conformations or conformations prohibitive to interaction with RNA or FRB, we constructed 4 configurations of this component

that had various order of these domains in the polypeptide chain (Figure 6.1a), and tested them all for the ability to be transported by our transporter constructs. The denomination of these four RBP constructs was an abbreviation of the domain names in the order of these domains appearing in the polypeptide chain from the N-terminus to the C-terminus of the fusion protein. Thus, the construct FFGPP consists of two domains of FKBP, eGFP, and two PUF domains in the aforementioned order from the N-terminus to the C-terminus of the fusion protein. All of the constructs belonging to both transporter and RBP parts of the system were cloned under cytomegalovirus (CMV) promoter.

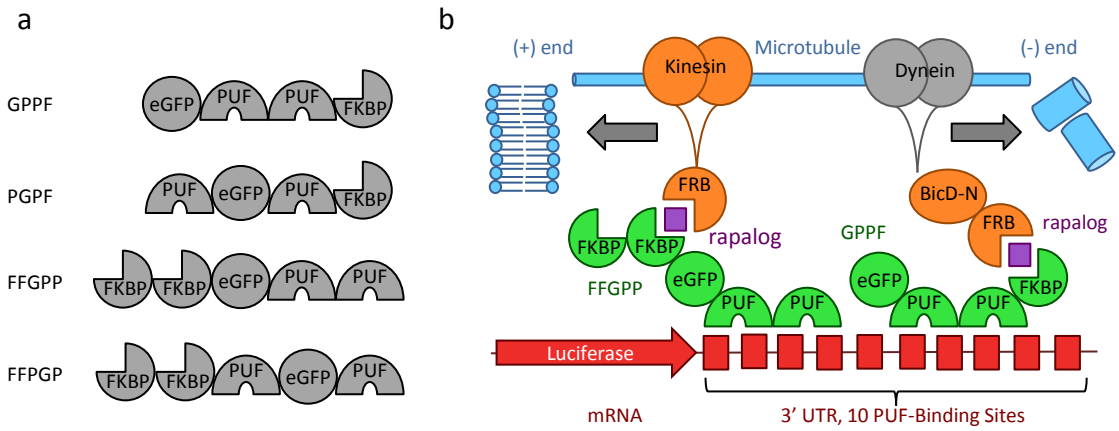


Figure 6.1 Schematic of the PUF-motor system. **(a)** Schematic of four RBP constructs with different order of domains (Left to right: N-terminus to C-terminus of the polypeptides). **(b)** Schematic of PUF/Kinesin and PUF/BicD-N systems for the transport of reporter mRNA to the (+) and (-) end of microtubules, respectfully. The two PUF domains in the FFGPP or GPPF fusions interact with the PUF-binding sites in the 3' UTR of firefly luciferase mRNA. Upon addition of the rapalog, the FKBP domain in FFGPP or GPPF constructs dimerize with the FRB domain, which is in turn fused to either kinesin or BicD-N. BicD-N interacts with the cellular motor dynein. Heterodimerization allows the assembly of the PUF-motor system and the transport of RNA.

In order to test which of the PUF-containing polypeptides were functional, we co-transfected each of these constructs with either Kif5b-FRB or BicD-N-FRB in HeLa cells. Since HeLa cells contain radial microtubule arrays stretching from the centrosome at their minus end to the cell periphery in their plus ends, we expected BicD-N and Kif5b constructs to transport the PUF constructs towards the centrosomes or cell periphery, respectively (Figure 6.1b). Twenty-four hours after transfection, rapalog was added at the final concentration of 1 μM to the cells in the conditioned medium for 1 hour before imaging. Live-cell imaging using widefield fluorescence microscopy revealed that when co-transfected with Kif5b-FRB and treated with rapalog, construct FFGPP localized to the cell periphery (Figure 6.2a), whereas it was spread throughout the cell cytoplasm when co-transfected with an empty vector and treated with rapalog (Figure 6.2b). BicD-N-FRB, on the other hand, interacted best with the construct GPPF, and allowed its localization in the perinuclear space (Figure 6.2c). When co-transfected with an empty vector and treated with rapalog, GPPF did not show the same perinuclear localization, but was spread evenly throughout the cellular cytoplasm (Figure 6.2d), indicating that the transport is indeed dependent on the introduced transporter constructs. Constructs PGPF and FFPGP were both solubly expressed, as judged from similar fluorescence intensities under widefield microscopy (data not shown). However, these constructs showed minimal localization to either cell periphery or perunuclear space upon

treatment with rapalog and in the presence of respective transporter constructs (data not shown).

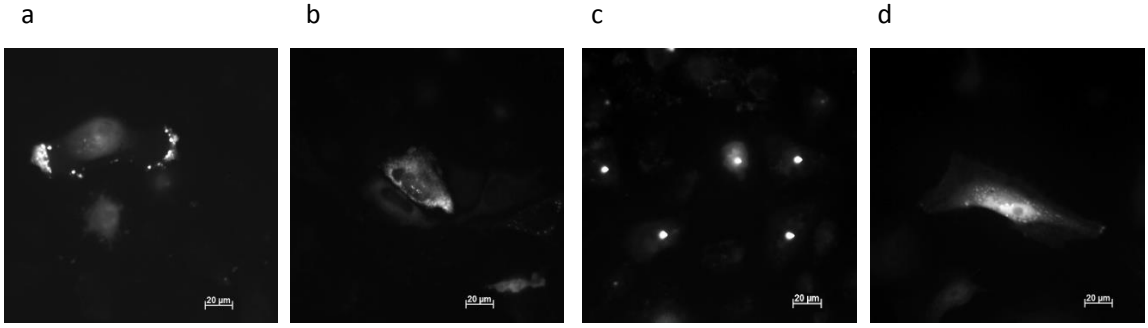


Figure 6.2 Transport of the PUF fusion constructs FFGPP and GPPF to the cell poles corresponding to the (+) and (-) ends of microtubules. **(a)** HeLa cells were transfected with Kif5b-FRB and FFGPP, grown for 24 hours and treated with 1 µM rapalog for 1 hour before imaging by widefield microscopy. GFP-tagged PUF is observed at the cell periphery. **(b)** HeLa cells were transfected with empty vector and FFGPP, grown for 24 hours and treated with 1 µM rapalog for 1 hour before imaging by widefield microscopy. **(c)** HeLa cells were transfected with BicD-N-FRB and GPPF, grown for 24 hours and treated with 1 µM rapalog for 1 hour before imaging by widefield microscopy. GFP-tagged PUF is aggregated at the perinuclear space. **(d)** HeLa cells were transfected with empty vector and GPPF, grown for 24 hours and treated with 1 µM rapalog for 1 hour before imaging by widefield microscopy.

6.2.2 Intracellular Transport of Reporter mRNA by the PUF-Motor

To assay the effectiveness of the device for the transport of mRNA, we utilized a reporter firefly luciferase (FL) mRNA, in the 3′ untranslated region (UTR) of which we placed 2 or 10 cognate PUF-binding sites (PBS) or a random 250 bp-sequence (Figure 6.1b). In the construct with 10 PBS, each PBS was separated from another by a spacer with a random sequence and alternating lengths of 6 or 18 nucleotides (nt), resulting in a 3′ UTR of approximately 250 bases. In the construct with 2 PBS, the

PBS were separated by 6 nt. The three components of the reporter assay (the transporter, corresponding RBP, and reporter mRNA) were co-transfected in HeLa cells. After 48 hours of cell growth, rapalog was added for 1 hour, after which the cells were fixed and the localization of all three constructs was analyzed by RNA fluorescence in situ hybridization (FISH) and immunofluorescence.

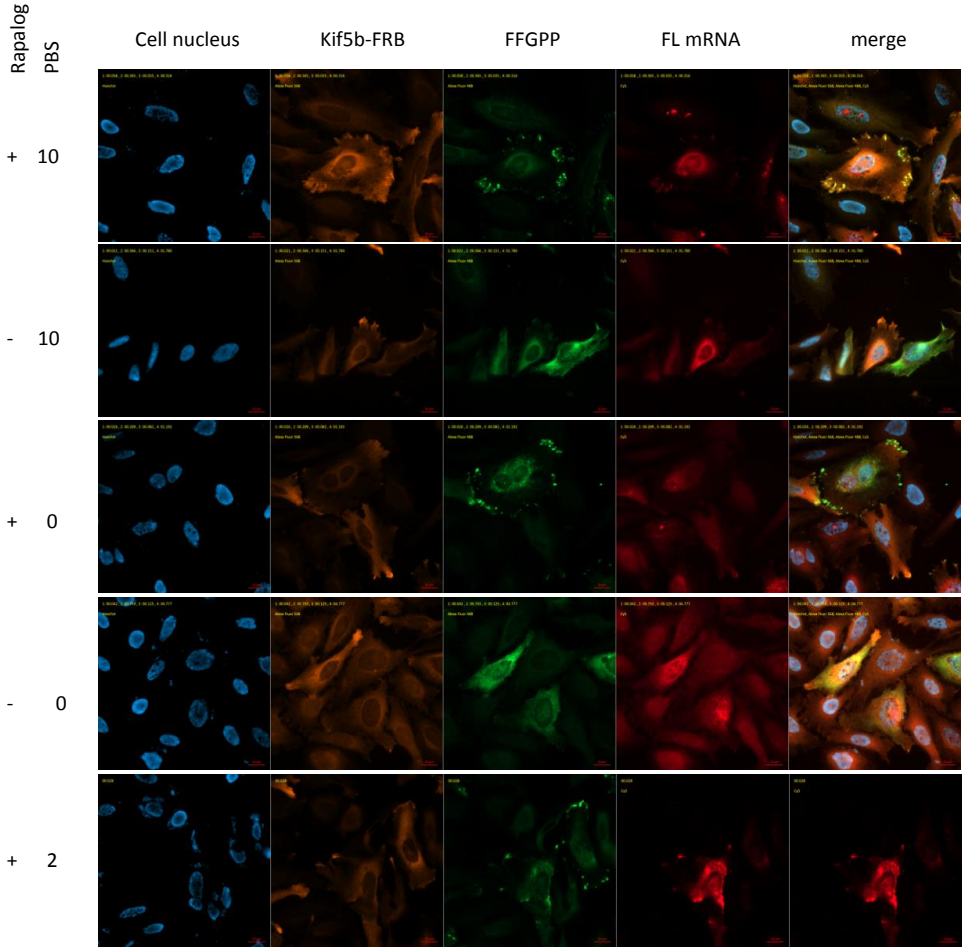


Figure 6.3 Transport of FL mRNA by the PUF-motor system consisting of Kif5b-FRB and the PUF construct FFGPP. HeLa cells were transfected with Kif5b-FRB, FFGPP, and FL with 0, 2, or 10 PBS after the stop codon, grown for 48 hours and treated with 1 μM rapalog for 1 hour before fixing. The samples were subjected to combined immunofluorescence and FISH for the assessment of protein and mRNA localization in the cells. After treatment with rapalog and in the presence of PBS in the 3' UTR of FL mRNA, Kif5b-FRB, FFGPP and mRNA are co-localized in the periphery of the cells.

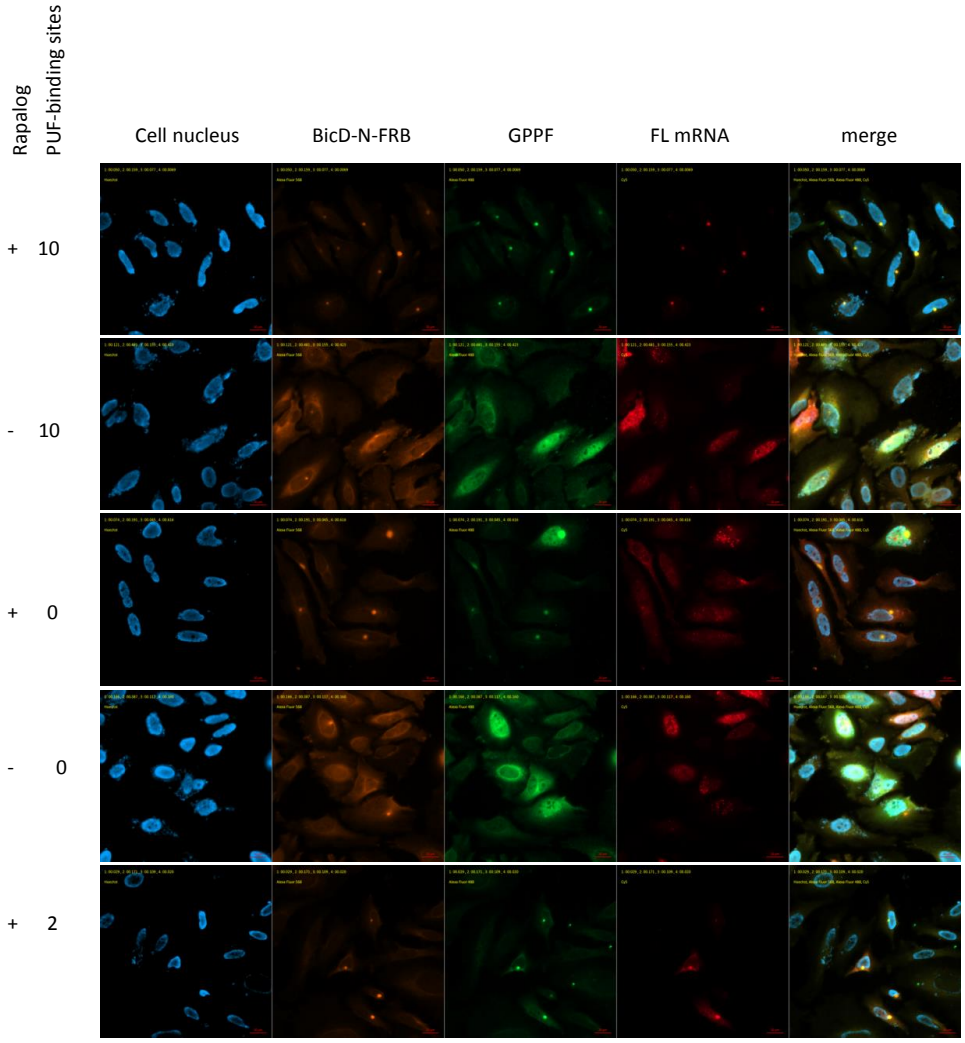


Figure 6.4 Transport of FL mRNA by the PUF-motor system consisting of BicD-N-FRB and the PUF construct GPPF. HeLa cells were transfected with BicD-N-FRB, GPPF, and FL with 0, 2, or 10 PBS after the stop codon, grown for 48 hours and treated with 1 μM rapalog for 1 hour before fixing. The samples were subjected to combined immunofluorescence and FISH for the assessment of protein and mRNA localization in the cells. After treatment with rapalog and in the presence of PBS in the 3' UTR of FL mRNA, BicD-N-FRB, GPPF, and mRNA are co-localized at the centrosome.

Widefield fluorescence microscopy revealed that when FFGPP was co-transfected with Kif5b-FRB, FFGPP co-localized with Kif5b-FRB in the presence of

rapalog, but not in the absence of rapalog (Figure 6.3). Moreover, FL mRNA was co-localized with FFGPP when the 3′ UTR of its mRNA contained 10 PBS, but not when the mRNA contained only 2 or no PBS (Figure 6.3). These results suggest that the transport of mRNA to the minus end of the microtubules by the PUF-motor system is motor-dependent and RNA sequence-dependent. However, the avidity between the FFGPP and the target RNA needs to be strong enough to have an observable transport of mRNA in HeLa cells.

Likewise, when GPPF was co-transfected with BicD-N-FRB, GPPF co-localized with BicD-N-FRB in the presence of rapalog, but not in the absence of rapalog (Figure 6.4). In addition, FL mRNA was co-localized with GPPF when the 3′ UTR of its mRNA contained 10 PBS, but less efficiently when the mRNA contained only 2 PBS. GPPF and FL mRNA did not co-localize when the 3′ UTR of mRNA contained a random sequence (Figure 6.4). These results likewise imply that the transport of mRNA in the plus-end direction of the microtubules is BicD-N as well as RNA sequence-dependent.

In order to quantify the effectiveness of mRNA transport by our PUF-motor system, we calculated the localization index of FL mRNA using the intensity of the fluorescence signal from the probe generated against the FL open reading frame. The localization index was calculated as $\frac{V_l}{V_c}$, where V_l is background-subtracted mean gray value of the RNA FISH signal in a region where most immunofluorescence signal from Kif5b or BicD-N was observed in the cell. V_c is background-subtracted mean gray value of the RNA FISH signal in a region of the same size, but in an adjacent area in the cytoplasm.

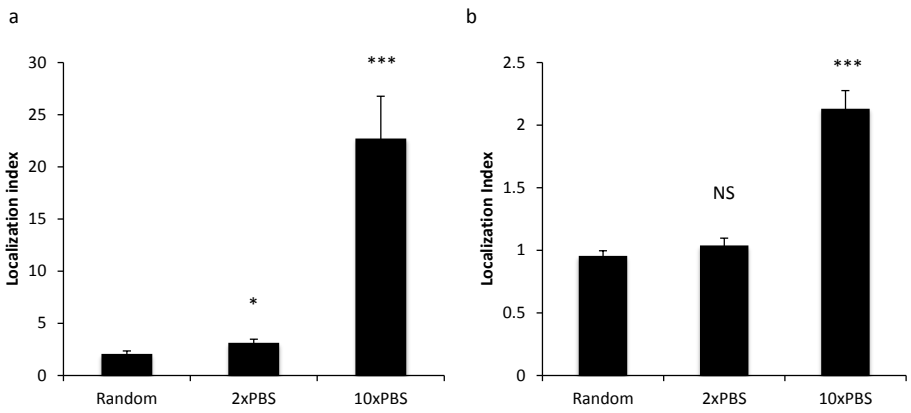


Figure 6.5 Quantification of FL mRNA localization from samples exemplified in Figure 3 and Figure 4. **(a)** Localization index of FL mRNA in samples co-transfected with Kif5b-FRB and FFGPP. *, $P < 0.05$. ***, $P < 0.001$. **(b)** Localization index of FL mRNA in samples co-transfected with BicD-N-FRB and GPPF. NS, not significant. Error bars represent standard error.

We thus calculated and averaged the localization index of RNA in 20-30 randomly picked HeLa cells that exhibited immunofluorescence signals from GFP-tagged PUF, HA-tagged transporter proteins, as well as FISH signal from luciferase mRNA. With the Kif5b/FFGPP combination, the localization index of luciferase mRNA was ~2 times greater in mRNA 3′ UTR of which contained 10 PBS compared to random sequence (Figure 6.5a). The localization index was not significantly greater for mRNA 3′ UTR of which contained only 2 PBS compared to random sequence ($P < 0.05$). On the other hand, with the BicD-N/GPPF combination, the localization index of luciferase mRNA was ~11 times greater in mRNA 3′ UTR of which contained 10 PBS compared to random sequence (Figure 6.5b). The localization index was 50% greater in mRNA 3′ UTR of which contained 2 PBS compared to random sequence ($P < 0.05$). These results corroborate our previous conclusion that the PUF-motor system is RNA sequence-dependent, but the avidity

between the PUF constructs and the target RNA plays an important role in the effectiveness of mRNA transport. These results are the first demonstration of an engineered device that is capable of re-distributing mRNA localization in higher eukaryotes.

6.2.3 Directional Transport of PUF and mRNA in Hippocampal Neurons

To apply our PUF-motor system for the transport of endogenous mRNA, we re-engineered both of the PUF domains for binding to β-actin mRNA and re-assembled them in the GPPF configuration. We chose β-actin mRNA, since its intercellular localization in neurons and local translation are studied better than most other mRNAs. Local translation of β-actin mRNA was implicated in axonal growth cone chemotropic response[4], axonal branching[40], and growth of dendritic filopodia[41].

Target name	Target sequence	Target length
actb1	TGTTTTTTGTTTTTGTTTTTTT	21
actb2	TGTTTTTTTGTTTTTTGTTTTTG	23
actb3	TGTGGCTGAGGACTTTGATTGTACATT	27
actb4	TGTGGCTGAGGACTTTGATTGTA	23
actb5	TGACAGCATTGCTTCTGTGTAAATT	25
actb6	TGTGTAAATTATGTACTTG	19
actb7	TGTATGAAGGCTTTGGTCTCCCTGGGAGTGGTTTGAGGTGTTGAGG	46
actb8	TGGTTTGAGGTGTTGAGG	18

Table 6.1 Target sites in rat β-actin mRNA that were chosen for PUF engineering. Red, 8-nt PUF binding sites.

The natural PUF proteins have a strong consensus in the first three bases of their target sequence (5′-UGU)[32], which are bound by the last three PUF repeats. For this reason, the two adjacent PUF target sites were chosen from the 3′ UTR sequence of rat β-actin mRNA so that minimal mutagenesis would be required in PUF repeats

6, 7, and 8. We chose 8 potential double target sites, with 3-30 nt spacer between the two putative PBS (Table 6.1). We next constructed 8 GPPF (GPPFact1-8) variants with the two PUF domains engineered to bind to the two adjacent target sites, in accordance with the RNA-binding code of PUF[32, 34-35]. The construction of these individual PUF variants was facilitated by the PUF Golden Gate assembly kit previously developed in our laboratory[42]. In order to screen the most effective of these engineered GPPF constructs, we tested them using the FL assay described above. To this end, we individually cloned each of the 8 putative RNA target sites into the 3′ UTR of the FL gene.

The effectiveness of FL mRNA transport by these reprogrammed GPPF was next assayed using the BicDN-FRB/eGPF system. Of all the GPPF constructs, GPPFact1 and GPPFact6 showed significant mRNA transport (Figure 6.6). Since the GPPFact6 yielded higher fluorescence signal compared to other constructs under similar conditions (data not shown), we assumed that GPPFact6 is more soluble and proceeded with this construct for the transport of endogenous β-actin mRNA in hippocampal neurons.

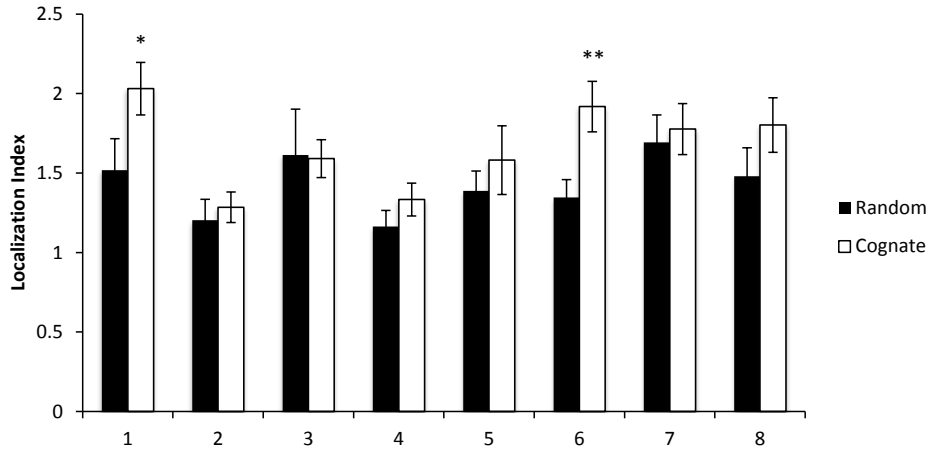


Figure 6.6 Assessment of activity of GPPF designed for binding to β-actin 3' UTR using the FL reporter assay. *, $P \leq 0.05$. **, $P < 0.01$. Error bars represent standard error.

To use this reprogrammed RBP for the transport of endogenous β-actin mRNA, the GPPFact6 construct was sub-cloned under GW1 promoter and used with the BicD-N-FRB construct under β-actin promoter. The two PUF mutants in the GPPFact6 construct were also sub-cloned into the FFGPP configuration under GW1 promoter for the interaction with the Kif5b-FRB. In accordance with previous reports[43], we expected Kif5b and BicD-N constructs to target cargos to axons and dendrites, respectively. FFGPPact6 was co-transfected with Kif5b-FRB and blue fluorescent protein (BFP) into rat hippocampal neurons that were cultured 8 days in vitro (DIV8). On the second day after transfection, rapalog was added to the medium at 0.1 uM concentration, and the neurons were grown for 24 more hours before fixation and analysis with widefield fluorescence microscopy.

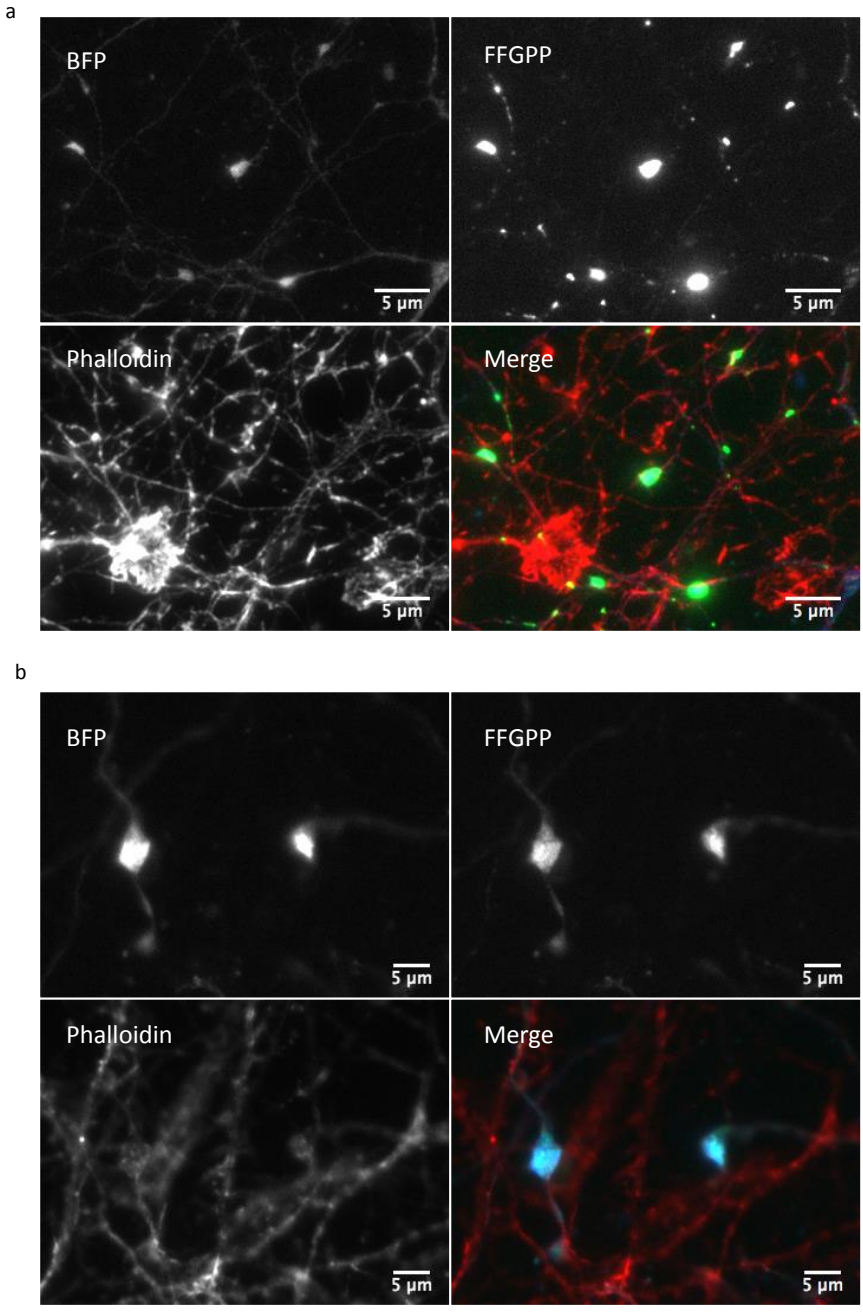


Figure 6.7. Application of the PUF-motor system for targeting endogenous β-actin mRNA. **(a)** FFGPP localization in axonal growth cones in DIV11 rat hippocampal neurons when co-transfected with Kif5b-FRB and treated with rapalog for 24 hours. After fixation, F-actin was stained with Alexa Fluor 568 Phalloidin. **(b)** FFGPP localization in axonal growth cones in DIV11 rat hippocampal neurons when co-transfected with Kif5b-FRB and not treated with rapalog. After fixation, F-actin was stained with Alexa Fluor 568 Phalloidin.

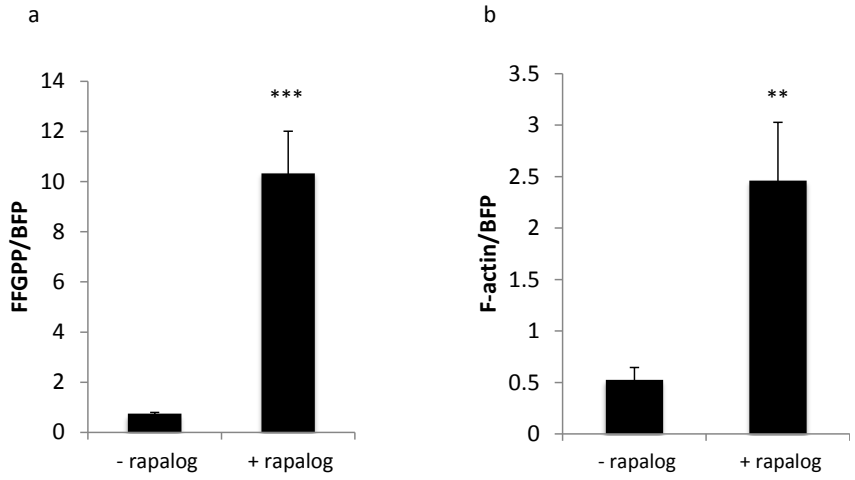


Figure 6.8 Application of the PUF-motor system for targeting endogenous β-actin mRNA. **(a)** FFGPP localization in axonal growth cones in DIV11 rat hippocampal neurons when co-transfected with Kif5b-FRB. The background-subtracted mean gray value of PUF-fused GFP fluorescence intensity was normalized to background-subtracted mean gray value of BFP fluorescence intensity. Fifty axonal growth cones from 10 randomly picked neurons were analyzed and averaged. Data represented as average ± standard error. ***, $P < 0.001$. **(b)** F-actin localization in growth cones in DIV11 rat hippocampal neurons transfected with Kif5b-FRB and β-actin-specific FFGPP. F-actin was phalloidin-stained and its localization in axonal growth cones quantified. The background-subtracted mean gray value of Alexa Fluor 568 fluorescence intensity was normalized to background-subtracted mean gray value of BFP fluorescence intensity. Fifty axonal growth cones from 10 randomly picked neurons were analyzed and averaged. Data represented as average ± standard error. **, $P < 0.01$.

As expected, rapalog addition induced targeting of FFGPPact6 to axonal growth cones by Kif5b (Figure 6.7 a and b). The localization of FFGPP was normalized to BFP intensity, which was assumed to be unaffected by the addition of rapalog. Quantification of 50 growth cones revealed that rapalog addition resulted in ~14-fold enrichment of FFGPPact6 in axonal growth cones (Figure 7.8 a). Of note,

phalloidin-staining of neurons for F-actin revealed 5-fold increase in F-actin localization in axonal growth cones (Figure 6.8b), which could be due to local β-actin mRNA translation in axonal growth cones enriched with FFGPPact6-transported mRNA. To confirm PUF-motor mediated re-distribution of β-actin mRNA in the neurons, we estimated mRNA amounts in axonal growth cones using RNA FISH in the presence and absence of rapalog (Figure 6.9 a and b). Although no obvious difference is observed from representative images, estimation of β-actin mRNA amounts averaged from 20 randomly picked growth cones revealed an almost two-fold increase of β-actin mRNA in axonal growth cones of neurons treated with rapalog (Figure 6.10). Although more investigation is required to definitively confirm this change being due to local protein synthesis, it is tempting to suggest that the increase in F-actin amount in axonal growth cones is due to the increased amount of transported β-actin mRNA.

The effect of BicD-N-FRB and rapalog to the localization of GPPFact6 as well as F-actin is currently being analyzed. We predict that the transfection of these constructs would sequester β-actin mRNA from axonal growth cones to neuronal cell bodies and dendrites, thus reducing the amount of β-actin mRNA in the growth cones.

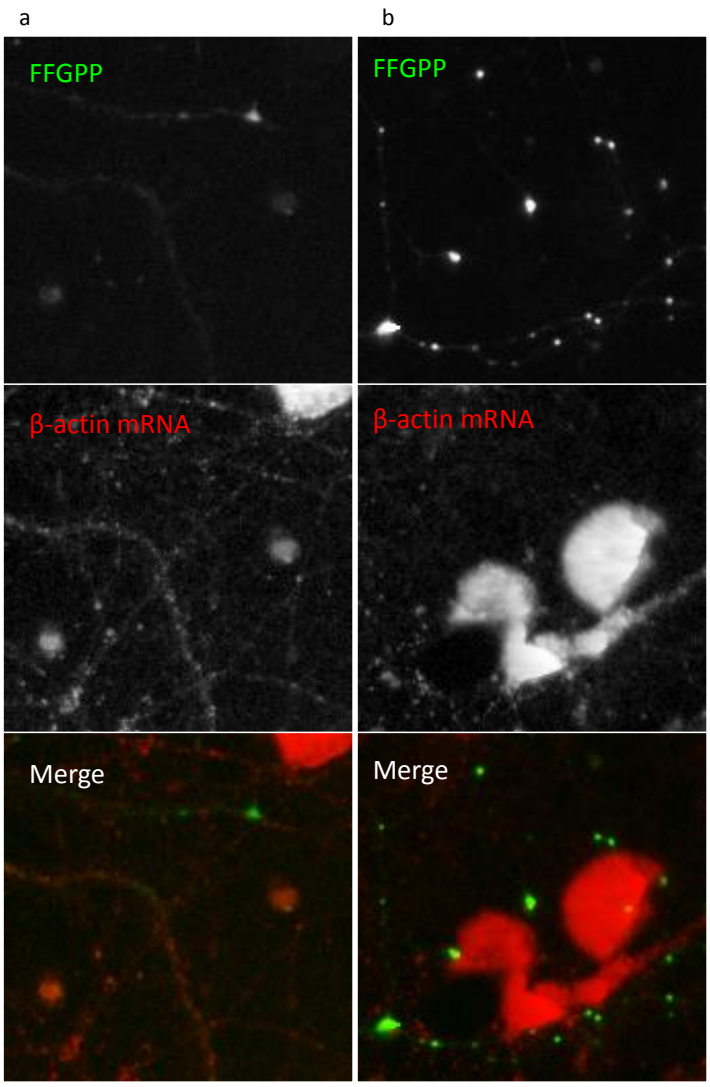


Figure 6.9 Application of the PUF-motor system for targeting endogenous β-actin mRNA. FFGPP localization in axonal growth cones in DIV11 rat hippocampal neurons when co-transfected with Kif5b-FRB and either not treated **(a)** or treated **(b)** with rapalog for 24 hours. After fixation, β-actin mRNA was visualized using RNA-FISH. Representative axonal growth cones are shown.

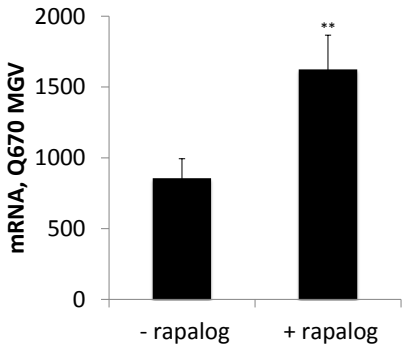


Figure 6.10 Quantification of β-actin mRNA in axonal growth cones estimated by measuring the mean gray value from the RNA-FISH signal. Twenty axonal growth cones from 10 randomly picked neurons were analyzed and averaged. Data represented as average ± standard error. **, $P < 0.01$.

6.3 Conclusions

In this work, we developed a synthetic device for sequence-specific transport of mRNA in eukaryotic cells. To the best of our knowledge, this work is the first demonstration of the efficacy of an engineered device capable of redistributing endogenous mRNA. The transport can be achieved in (+) or (-) directions of the microtubules in cultured cancerous mammalian cells or mammalian neurons. This prototypical synthetic device can potentially control intracellular transport of any endogenous mRNA of interest in eukaryotes, as well as local mRNA translation in polarized eukaryotic cells. Eventually, we believe that this device will broaden our capability in regulating mRNA metabolism and gene expression, which will be helpful in basic science and therapeutics.

6.4 Materials and Methods

6.4.1 Cloning of DNA Constructs

All the constructs used for transfection of HeLa cells were cloned under CMV

promoter. To this end, the backbone, including the start codon and the 3xflag tag were amplified from the pCMV-TTP (C147R)-GG-PUF plasmid (Chapter 5). Kif5b-FRB was PCR-amplified from HA-Kif5b(1-807)-FRB[43], and BicDN-FRB was PCR-amplified from the HA-BICD2-N(1-594)-FRB plasmid[43], which were then assembled into the amplified vector backbone using the Gibson Assembly method. PUF-GFP-FKBP constructs were similarly assembled into the amplified vector using Gibson Assembly method. For this purpose, the modified PUF domain was amplified from TPUF (6SE,7NQ) plasmid (Chapter 5). The eGFP gene was amplified from the pCMV5-eGFP plasmid, and double or single FKBP domains were amplified from the PEX-RFP-FKBP[43] plasmid. For the construction of GPPFact(1-8) plasmids, individual PUF domains were first assembled in pET28 vector using the Golden Gate assembly method and the parts library developed previously[42]. The complete GPPFact(1-8) constructs were then assembled into the pCMV5 backbone from mutant PUF domains, eGFP, and FKBP fragments using the Gibson Assembly method as above. Plasmid maps are available upon request.

The construction of plasmids for transcription of firefly luciferase mRNA containing a 3′ UTR with 10 PBS, or a random sequence were described previously (Chapter 5). The plasmid for the transcription of firefly luciferase mRNA containing a 3′ UTR with 2 PBS was constructed by Gibson Assembly. For this purpose, double stranded (ds) oligomer was prepared by annealing single stranded (ss) oligomers 5′-TATCGATAAGCTTGCATGCCTGCAGTTGATATAAAGTGTTTGATATAGGTGG

and 5′-GGGGTCACAGGGATGCCACCTATATCAAACACTTTATATCAACTGCAG and assembled with the *Pst*I and *Xma*I-digested pCMV5-Fluc (construction of this plasmid was described in Chapter 5). The plasmids for transcription of firefly luciferase mRNA with 3′ UTRs containing short inserts from the 3′ UTR of rat β-actin mRNA were prepared by ligation of ds oligomers into the *Pst*I and *Xma*I-digested pCMV5-Fluc plasmid. The ds oligomers were prepared by annealing two ss oligomers and subsequent 5′-phosphorylation using the T4 polynucleotide kinase (NEB), as per manufacturer's protocol.

For expression in hippocampal cells, the BicD-N-FRB construct was subcloned under the β-actin promoter. To this end, the BicD-N-FRB fragment was amplified from pCMV5-BicD-N-FRB and ligated into *Hind*III and *Sal*I sites in the β-actin vector. Similarly, the Kif5b-FRB fragment was amplified from pCMV5-Kif5b-FRB and ligated into *Asc*I and *Sal*I sites of the β-actin2b-MCS-ATG vector. FFGPPact6 and GPPFact6 constructs were cloned under GW1 promoter. For this purpose, GPPFact6 was amplified from pCMV5-GPPFact6 and ligated into the HindIII and SalI sites of the GW1-HA-2b vector. FFGPPact6 was cloned by amplification of the GW1 vector, 2xFKBP fragment from the PEX-RFP-FKBP plasmid, and the eGFP-PUF-PUF fragment (with a primer-added stop codon) from the GPPFact6 construct, and subsequent splicing using the Gibson Assembly method.

6.4.2 Cell Culture and Transfection

HeLa cells were maintained in modified Eagle's Medium (MEM) supplemented with 10 % Fetal Bovine Serum (FBS). HeLa cells were plated a day before transfection at a density of $1x10^4$ cells/well in 8-well poly-L-lysine-coated μ-slide (Ibidi) in 300 μl

of MEM with 10% FBS. The next day, transfection mixtures were prepared by mixing 15 µl OptiMEM (Life Technologies), 300 ng DNA, and 0.9 µl Fugene HD reagent (Promega). The DNA mixture contained 100 ng transporter plasmid, 100 ng PGF plasmid, and 100 ng firefly luciferase plasmid. The transfection mixtures were incubated for 15 minutes at room temperature and added to the wells with cells. The cells were allowed to recover and grow for 48 hours before heterodimerization of motor and PGF constructs in 200 µl media supplemented with 1 µM A/C heterodimerizer (Clontech Laboratories) for 1 hour at 37C.

Primary hippocampal cultures were prepared from embryonic day 18 (E18) rat brains as described previously[43]. Briefly, cells were plated on coverslips coated with poly-L-lysine (30 µg/ml) and laminin (2 µg/ml) at a density of 75,000/well. Hippocampal cultures were grown in Neurobasal medium (NB) supplemented with B27, 0.5 mM glutamine, 12.5 µM glutamate, and penicillin-streptomycin. Hippocampal neurons were transfected at DIV8 with Lipofectamine 2000 (Invitrogen). DNA (1.8 g µg/well) was mixed with 3.3 ml Lipofectamine 2000 in 200 µl NB, incubated for 30 min, and then added to the neurons in NB at 37°C in 5 % CO_2 for 45 min. Next, neurons were washed with NB and transferred to the conditioned medium at 37 °C in 5 % CO_2 . Rapalog was added to the media at DIV10 to a final concentration of 0.1 µM and the neurons were grown for additional 24 hours at 37 °C in 5 % CO_2.

6.4.3 RNA-FISH, Immunofluorescence, and Phalloidin Staining

For combined RNA-FISH and immunofluorescence, HeLa cells were washed twice with phosphate buffer saline (PBS) and fixed with 4% formaldehyde (Polysciences)

in 2x saline-sodium-citrate (SSC) buffer for 10 minutes at room temperature (RT). The cells were washed twice in PBS and permeabilized with 70% ethanol at 4C overnight, after which they were washed in a washing buffer (2xSSC, 10 % formamide -Sigma) for 5 minutes at RT. RNA-FISH co-immunofluorescence staining was performed by incubation with 170 nM Stellaris Quasar 670-tagged firefly luciferase 48 probe mix (Biosearch Technologies), 1 μg/ml mouse monoclonal anti-HA tag IgG antibody (Abcam), and 5 μg/ml rabbit monoclonal anti-GFP IgG Abfinity antibody (Invitrogen) in the FISH-IF buffer (10 % dextran sulfate, 10 % formamide, 2x SSC, 2mM vanadyl rybonucleoside complex, 0.02 % RNase-free bovine serum albumin (BSA)) overnight at 37 °C in the dark. The cells were washed three times in the washing buffer and incubated with 4 μg/ml Alexa-Fluor568-tagged anti-mouse IgG antibodies and 10 μg/ml Alexa-Fluor488-tagged anti-rabbit antibody (Invitrogen) in the FISH-IF buffer for 1 hour at RT in the dark, after which they were subsequently washed three times in the washing buffer. The cells were subsequently conterstained in 130 ng/ml Hoechst 33342 (Pierce) for 5 minutes at RT, and washed three times in 2xSSC buffer. After washing in the oxygen scavenger (GLOX) buffer (2xSSC, 0.4 % glucose, 10 mM Tris-HCl pH 8) for 5 minutes at RT, the cells were imaged in GLOX solution (37 μg/ml glucose oxidase (Sigma) and 100 μg/ml catalase (Sigma) in GLOX buffer).

For RNA-FISH of embryonic rat hippocampal cells, transfected DIV11 neurons were fixed in 4% paraformaldehyde supplemented with 4 % sucrose for 10 min at RT, washed 3x with PBS, and permeabilized in 70% ethanol overnight at 4 °C. The coverslips were next washed in the washing buffer for 5 minutes at RT. RNA-FISH

was performed by incubation with 170 nM Stellaris Quasar 670-tagged 48 probe mix specific for rat actb (Biosearch Technologies) in the FISH-IF buffer overnight at 37 °C in the dark. The cells were washed three times in the washing buffer and imaged in the GLOX buffer as above.

For visualization of F-actin in neurons, fixed cells were permeabilized in 0.1 % Triton X-100 in PBS for 15 minutes at RT, and washed in PBS. The coverslips were next treated with 0.4 units/ml Alexa Fluor 568 Phalloidin (Life Technologies) in 1 % goat serum and 1x PBS for 20 minutes at RT, after which they were washed in PBS and mounted using Vectashield mounting medium (Vector laboratories).

6.4.4 Image Acquisition and Quantification

Images were acquired using the Zeiss Axiovert 200M widefield fluorescence microscope equipped with the 40x NA 1.4 oil objective, Zeiss Axiocam 506 high resolution camera, and DAPI, FITC/GFP, Rhodamine, and Cy5 excitation/emission filter set (Sutter). Localization index was determined by calculating the ratio of Cy5 (RNA) intensity co-localized with either Kif5b-FRB or BicDN-FRB versus cytoplasmic intensity by measuring the mean gray value of the localized spot and in an area of equal size in the adjacent cytoplasm using ImageJ (http://rsb.info.nih.gov/ij/index.html). Localization indexes were averaged over 20-30 cells and a statistical analysis was performed with student's t test assuming a two-tailed and unequal variation. Origin 8.5 software was used to plot the localization indexes on box-plot charts.

Neurons were imaged using a Nikon Eclipse 80i microscope equipped with a

Plan Fluor 60x N.A. 1.40 oil objective, Chroma ET-DAPI (49000), Chroma ET-GFP (49002), and Chroma ET-mCherry (49008) filters, and a Photometrics CoolSNAP HQ2 CCD camera. Mean gray values of GFP, BFP, and Alexa Fluor 568 fluorescence signals ere measured using ImageJ. The background-subtracted mean gray value of PUF-fused GFP fluorescence intensity or Alexa Fluor 568 fluorescence intensity was normalized to background-subtracted mean gray value of BFP fluorescence intensity. Fifty axonal growth cones from 10 randomly picked neurons were analyzed and averaged.

6.5 References

(1) Lecuyer, E.; Yoshida, H.; Parthasarathy, N.; Alm, C.; Babak, T.; Cerovina, T.; Hughes, T. R.; Tomancak, P.; Krause, H. M., Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* **2007,** *131*, 174-87.

(2) Holt, C. E.; Bullock, S. L., Subcellular mRNA localization in animal cells and why it matters. *Science* **2009,** *326*, 1212-6.

(3) Jung, H.; Gkogkas, C. G.; Sonenberg, N.; Holt, C. E., Remote control of gene function by local translation. *Cell* **2014,** *157*, 26-40.

(4) Leung, K. M.; van Horck, F. P.; Lin, A. C.; Allison, R.; Standart, N.; Holt, C. E., Asymmetrical beta-actin mRNA translation in growth cones mediates attractive turning to netrin-1. *Nature neuroscience* **2006,** *9*, 1247-56.

(5) Yao, J.; Sasaki, Y.; Wen, Z.; Bassell, G. J.; Zheng, J. Q., An essential role for beta-actin mRNA localization and translation in Ca2+-dependent growth cone guidance. *Nature neuroscience* **2006,** *9*, 1265-73.

(6) Campbell, D. S.; Holt, C. E., Chemotropic responses of retinal growth cones mediated by rapid local protein synthesis and degradation. *Neuron* **2001,** *32*, 1013-26.

(7) Miller, S.; Yasuda, M.; Coats, J. K.; Jones, Y.; Martone, M. E.; Mayford, M., Disruption of dendritic translation of CaMKIIalpha impairs stabilization of synaptic plasticity and memory consolidation. *Neuron* **2002,** *36*, 507-19.

(8) Wang, D. O.; Kim, S. M.; Zhao, Y.; Hwang, H.; Miura, S. K.; Sossin, W. S.; Martin, K. C., Synapse- and stimulus-specific local translation during long-term neuronal plasticity. *Science* **2009,** *324*, 1536-40.

(9) Melia, K. R.; Trembleau, A.; Oddi, R.; Sanna, P. P.; Bloom, F. E., Detection and regulation of tyrosine hydroxylase mRNA in catecholaminergic terminal fields: possible axonal compartmentalization. *Experimental neurology* **1994,** *130*, 394-406.

(10) Trembleau, A.; Melia, K. R.; Bloom, F. E., BC1 RNA and vasopressin mRNA in rat neurohypophysis: axonal compartmentalization and differential regulation during dehydration and rehydration. *The European journal of neuroscience* **1995,** *7*, 2249-60.
(11) Andreassi, C.; Zimmermann, C.; Mitter, R.; Fusco, S.; De Vita, S.; Saiardi, A.; Riccio, A., An NGF-responsive element targets myo-inositol monophosphatase-1 mRNA to sympathetic neuron axons. *Nature neuroscience* **2010,** *13*, 291-301.
(12) Cox, L. J.; Hengst, U.; Gurskaya, N. G.; Lukyanov, K. A.; Jaffrey, S. R., Intra-axonal translation and retrograde trafficking of CREB promotes neuronal survival. *Nature cell biology* **2008,** *10*, 149-59.
(13) Yoon, B. C.; Jung, H.; Dwivedy, A.; O'Hare, C. M.; Zivraj, K. H.; Holt, C. E., Local translation of extranuclear lamin B promotes axon maintenance. *Cell* **2012,** *148*, 752-64.
(14) Verma, P.; Chierzi, S.; Codd, A. M.; Campbell, D. S.; Meyer, R. L.; Holt, C. E.; Fawcett, J. W., Axonal protein synthesis and degradation are necessary for efficient growth cone regeneration. *The Journal of neuroscience* **2005,** *25*, 331-42.
(15) Yan, D.; Wu, Z.; Chisholm, A. D.; Jin, Y., The DLK-1 kinase promotes mRNA stability and local translation in *C. elegans* synapses and axon regeneration. *Cell* **2009,** *138*, 1005-18.
(16) Dahm, R.; Macchi, P., Human pathologies associated with defective RNA transport and localization in the nervous system. *Biology of the cell / under the auspices of the European Cell Biology Organization* **2007,** *99*, 649-61.
(17) Liu-Yesucevitz, L.; Bassell, G. J.; Gitler, A. D.; Hart, A. C.; Klann, E.; Richter, J. D.; Warren, S. T.; Wolozin, B., Local RNA translation at the synapse and in disease. *The Journal of neuroscience* **2011,** *31*, 16086-93.
(18) Bashirullah, A.; Cooperstock, R. L.; Lipshitz, H. D., Spatial and temporal control of RNA stability. *Proceedings of the national academy of sciences of the United States of America* **2001,** *98*, 7025-8.
(19) Wolke, U.; Weidinger, G.; Koprunner, M.; Raz, E., Multiple levels of posttranscriptional control lead to germ line-specific gene expression in the zebrafish. *Current biology* **2002,** *12*, 289-94.
(20) Forrest, K. M.; Gavis, E. R., Live imaging of endogenous RNA reveals a diffusion and entrapment mechanism for nanos mRNA localization in Drosophila. *Current biology* **2003,** *13*, 1159-68.
(21) Gagnon, J. A.; Mowry, K. L., Molecular motors: directing traffic during RNA localization. *Critical reviews in biochemistry and molecular biology* **2011,** *46*, 229-39.
(22) Bullock, S. L., Messengers, motors and mysteries: sorting of eukaryotic mRNAs by cytoskeletal transport. *Biochemical Society transactions* **2011,** *39*, 1161-5.
(23) Hirokawa, N.; Noda, Y.; Tanaka, Y.; Niwa, S., Kinesin superfamily motor proteins and intracellular transport. *Nature reviews.* **2009,** *10*, 682-96.
(24) Vallee, R. B.; Williams, J. C.; Varma, D.; Barnhart, L. E., Dynein: An ancient motor protein involved in multiple modes of transport. *Journal of neurobiology* **2004,** *58*, 189-200.
(25) McCaffrey, M. W.; Lindsay, A. J., Roles for myosin Va in RNA transport and turnover. *Biochemical Society transactions* **2012,** *40*, 1416-20.

(26) Bohl, F.; Kruse, C.; Frank, A.; Ferring, D.; Jansen, R. P., She2p, a novel RNA-binding protein tethers ASH1 mRNA to the Myo4p myosin motor via She3p. *The EMBO journal* **2000,** *19*, 5514-24.
(27) Long, R. M.; Gu, W.; Lorimer, E.; Singer, R. H.; Chartrand, P., She2p is a novel RNA-binding protein that recruits the Myo4p-She3p complex to ASH1 mRNA. *The EMBO journal* **2000,** *19*, 6592-601.
(28) Dienstbier, M.; Boehl, F.; Li, X.; Bullock, S. L., Egalitarian is a selective RNA-binding protein linking mRNA localization signals to the dynein motor. *Genes & development* **2009,** *23*, 1546-58.
(29) Belmont, B. J.; Niles, J. C., Inducible control of subcellular RNA localization using a synthetic protein-RNA aptamer interaction. *PloS one* **2012,** *7*, DOI: 10.1371/journal.pone.0046868.
(30) Chartrand, P.; Meng, X. H.; Singer, R. H.; Long, R. M., Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle *in vivo*. *Current biology* **1999,** *9*, 333-8.
(31) Kislauskis, E. H.; Li, Z.; Singer, R. H.; Taneja, K. L., Isoform-specific 3'-untranslated sequences sort alpha-cardiac and beta-cytoplasmic actin messenger RNAs to different cytoplasmic compartments. *The journal of cell biology* **1993,** *123*, 165-72.
(32) Wang, X.; McLachlan, J.; Zamore, P. D.; Hall, T. M., Modular recognition of RNA by a human pumilio-homology domain. *Cell* **2002,** *110*, 501-12.
(33) Cheong, C. G.; Hall, T. M., Engineering RNA sequence specificity of Pumilio repeats. *Proceedings of the national academy of sciences of the United States of America* **2006,** *103*, 13635-9.
(34) Dong, S.; Wang, Y.; Cassidy-Amstutz, C.; Lu, G.; Bigler, R.; Jezyk, M. R.; Li, C.; Hall, T. M.; Wang, Z., Specific and modular binding code for cytosine recognition in Pumilio/FBF (PUF) RNA-binding domains. *The Journal of biological chemistry* **2011,** *286*, 26732-42.
(35) Filipovska, A.; Razif, M. F.; Nygard, K. K.; Rackham, O., A universal code for RNA recognition by PUF proteins. *Nature chemical biology* **2011,** *7*, 425-7.
(36) Banaszynski, L. A.; Liu, C. W.; Wandless, T. J., Characterization of the FKBP.rapamycin.FRB ternary complex. *Journal of the american chemical society* **2005,** *127*, 4715-21.
(37) Hoogenraad, C. C.; Akhmanova, A.; Howell, S. A.; Dortland, B. R.; De Zeeuw, C. I.; Willemsen, R.; Visser, P.; Grosveld, F.; Galjart, N., Mammalian Golgi-associated Bicaudal-D2 functions in the dynein-dynactin pathway by interacting with these complexes. *The EMBO journal* **2001,** *20*, 4041-54.
(38) Hoogenraad, C. C.; Wulf, P.; Schiefermeier, N.; Stepanova, T.; Galjart, N.; Small, J. V.; Grosveld, F.; de Zeeuw, C. I.; Akhmanova, A., Bicaudal D induces selective dynein-mediated microtubule minus end-directed transport. *The EMBO journal* **2003,** *22*, 6004-15.
(39) Kapitein, L. C.; Schlager, M. A.; van der Zwan, W. A.; Wulf, P. S.; Keijzer, N.; Hoogenraad, C. C., Probing intracellular motor protein activity using an inducible cargo trafficking assay. *Biophysical journal* **2010,** *99*, 2143-52.
(40) Donnelly, C. J.; Park, M.; Spillane, M.; Yoo, S.; Pacheco, A.; Gomes, C.; Vuppalanchi, D.; McDonald, M.; Kim, H. H.; Merianda, T. T.; Gallo, G.; Twiss, J. L.,

Axonally synthesized beta-actin and GAP-43 proteins support distinct modes of axonal growth. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **2013,** *33*, 3311-22.
(41) Eom, T.; Antar, L. N.; Singer, R. H.; Bassell, G. J., Localization of a beta-actin messenger ribonucleoprotein complex with zipcode-binding protein modulates the density of dendritic filopodia and filopodial synapses. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **2003,** *23*, 10433-44.
(42) Abil, Z.; Denard, C. A.; Zhao, H., Modular assembly of designer PUF proteins for specific post-transcriptional regulation of endogenous RNA. *Journal of biological engineering* **2014,** *8*, DOI: 10.1186/1754-1611-8-7.
(43) Kapitein, L. C.; Schlager, M. A.; Kuijpers, M.; Wulf, P. S.; van Spronsen, M.; MacKintosh, F. C.; Hoogenraad, C. C., Mixed microtubules steer dynein-driven cargo transport into dendrites. *Current biology* **2010,** *20*, 290-9.

Appendix A

DNA sequences of target and random double stranded oligonucleotides used in fluorescence polarization experiments. Note that the abbreviation /56-FAM/ corresponds to 5' 6-FAM, which is a single isomer derivative of fluorescein.

Cognate DNA sequence

56FAM-5'-ATCTAGCAACCTCAAACAGACACCATACG-3'

3'-TAGATCGTTGGAGTTTGTCTGTGGTATGC-5'

Random DNA sequence

56FAM-5'-ATCAGACCGACATCTAATCCGCAACACAA-3'

3'-TAGTCTGGCTGTAGATTAGGCGTTGTGTT-5'

TALE21.5 and its target DNA sequence

target nucleotide position:		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	21.5
target nucleotide:	T	A	G	C	A	A	C	C	T	C	A	A	A	C	A	G	A	C	A	C	C	A	T
RVD:		NI	NN	HD	NI	NI	HD	HD	HG	HD	NI	NI	NI	HD	NI	NN	NI	HD	NI	HD	HD	NI	NG

Amino acid sequence of Naldt-TALE21.5

MGPLCTPSRSSHHHHHHSSGLVPRGSHMLDTSLLDSMPAVGTPHTAAAPAECDEVQSGL
RAADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKP
KVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVTYQDIIRALPEATHEDIVGVG
KQWSGARALEALLTEAGELRGPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLN
LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETVQRL
LPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGK
QALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVV
AIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQ
RLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIG
GKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQ
VVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLC
QDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALET
VQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASN
IGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTP
DQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASNGGGKQALESIVAQLSRPDPALAALTNDHLVALACLGGRPAL
DAVKKGLPHAPELIRRINRRIPERTSHRVA

Amino acid sequence of Naldt-NTR

MGPLCTPSRSSHHHHHHSSGLVPRGSHMLDTSLLDSMPAVGTPHTAAAPAECDEVQSGL
RAADDPPPTVRVAVTAARPPRAKPAPRRRAAQPSDASPAAQVDLRTLGYSQQQQEKIKP
KVRSTVAQHHEALVGHGFTHAHIVALSQHPAALGTVAVTYQDIIRALPEATHEDIVGVG
KQWSGARALEALLTEAGELRGPPLQLDTGQLLKIAKRGGVTAVEAVHAWRNALTGAPLN

Amino acid sequence of Naldt-CRD-CTR

MGLCTPSRSSHHHHHHSSGLVPRGSHMLDTSLNLTPDQVVAIASNIGGKQALETVQRLL
PVLCQDHGLTPDQVVAIASNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQ
ALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVA
IASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDH
GLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQR
LLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGG
KQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQV
VAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQ
DHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNNGGKQALETV
QRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHD
GGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPD
QVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALE
SIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPELIRRINRRIPERTS
HRVA