Single molecule characterization of proteins involved in replication of *E. coli*, the LINE1 retrotransposon, and retroviruses.

by Mohamed Nabuan Naufer
BSc. in Physics, University of Colombo

A dissertation submitted to

The Faculty of
the College of Science of
Northeastern University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

January 6, 2019

Dissertation directed by

Mark C. Williams
Professor and Chair
Department of Physics, Northeastern University
Acknowledgments

Being able to spend a better part of my adult life working on the science that excites me as a PhD candidate at a prestigious university, living in the hub of the world-class higher education has been an outstanding privilege enabled by many fantastic people.

I am utterly grateful for my advisor Mark Williams for taking me under his wing, sharing his knowledge and nurturing my outlandish ideas by giving me the freedom and resources to execute my research projects. Having him as my PhD advisor is most certainly the best thing happened in my science journey and one could have not asked for a better guidance, support and mentorship, which made me the scientist I am today.

I have been very fortunate to work with a group of excellent interdisciplinary scientists for every project I did during my PhD and none of those would have been possible without their resilient collaborations. Penny Beuning made the indispensable biochemistry possible facilitating the tremendous support from David Murison and Bilyana Koleva in Beuning lab. I am extremely appreciative of the time she spent to foster my research work as a collaborator as well as one of my dissertation committee members. Anthony Furano introduced me to the world of retrotransposons and supported me throughout the journey from learning the definition of the word to publishing a review article, with his utmost knowledge, patience and sense of humor, being the best in the field. I thank Ioulia Rouzina for solidifying theoretical framework for my research projects while going through enormous data sets and indulging my scientific thoughts during conference- and workshop-breaks.

I am extremely thankful for Alain Karma and Paul Whitford for agreeing to be on my dissertation committee and being the most patient and flexible on scheduling my dissertation defense. It is my honor to have them both in my committee.

Williams lab has always been filled with pleasant and fun vibes. The senior research scientist Micah McCauley, alumni Kathy Chaurasiya and Hao Wu patiently trained me on optical tweezers through catching my first DNA, which still doesn’t cease to amaze me! It has been a wonderful science ride since then with all the past and present Williams lab members; Kiran Pant, Thayaparan Paramanathan, Sam Lone, Jailin Li, Ali Almaqwashi, Ran Huo, Divakaran Murugesapillai, Andrew Clark, Ben Cashen and Gudfridur Moller. My time in Williams lab was significantly
enriched with the giant contribution made by Mike Morse, with his invaluable ideas and support on the exciting research work we accomplished together.

I would like to thank all the physics graduate students at Northeastern for making the department an exceptional place to do research, especially, Rohan Gala who has been continuously inspiring and encouraging me to explore new avenues in research, while finding the potential of my pet-projects which were rough around the edges and contributing to carry them through to completion.

I had a memorable time at Northeastern with Hasnain Hafiz, Alex Hyde, Younggil Song, Ross Altman from philosophical discussions to entertaining hangouts.

Many thanks to Nancy Wong, Lindsay Day, Alina Mak, and Sheila Magee Beare, et al. at the Physics department office, and Ilana Zeltser, Melissa Rischall at the College of Science, for making life easy on all administrative matters and for ensuring I met all the real-life deadlines and requirements to become a PhD. I would like to thank Harold Marshman, and Daniel Sierpina at the Physics department machine shop, for gracefully transforming my schematics into functional components that were vital in my research.

Initiating this endeavor was possible due the unconditional love, hard work and sacrifices made by my parents; Segmadar Mohamed Naufer and Haleemath Sahithiya Naufer. I am humbled by knowing I could make them both proud. I thank my wife Mia Edirisinghe for being extremely supportive and putting up with my long and odd work routine, my siblings Nafhan and Nusran Naufer for the inspiration and encouragement throughout this journey.
Abstract of Dissertation

Single molecule force spectroscopy is a powerful tool used to quantitatively probe protein-DNA interactions. The analyses discussed in this dissertation involves quantifying and modeling protein-DNA interactions to elucidate the mechanisms involved in E. coli, retroviral, and L1 retrotransposon replication.

E. coli is a model organism that has been abundantly used to study the complex processes involved in DNA replication. Pol III core is the three-subunit subassembly of the E. coli replicative DNA polymerase III holoenzyme. It contains the catalytic polymerase subunit α, the 3′ → 5′ proofreading exonuclease ε, and a subunit of unknown function, θ. Here, we characterize the pol III core activity on a single DNA substrate. We observe polymerization at applied template forces F<25 pN and exonucleolysis at F>30 pN. Both polymerization and exonucleolysis occur as a series of short bursts separated by pauses. For polymerization, the initiation rate after pausing is independent of the force. In contrast, the exonucleolysis initiation rate strongly depends on force. The measured force and concentration dependence of exonucleolysis initiation fit well to a two-step reaction scheme in which pol III core bimolecularly binds to the primer-template junction, and then converts at rate $k_2$ into an exo-competent conformation. Fits to the force dependence of initiation rate $k_{init}$ show that exo initiation requires fluctuational opening of two base pairs, in agreement with temperature- and mismatch-dependent bulk biochemical assays. The results support a model in which the pol and exo activities of pol III core are effectively independent, and in which recognition of the 3′ end of the primer by either α or ε is governed by the primer stability. Thus, binding to an unstable primer is the primary mechanism for mismatch recognition during proofreading, rather than duplex defect recognition described in an alternative model.

Long stretches of single-stranded DNA (ssDNA) are formed during genomic maintenance processes such as replication, recombination and repair. ssDNA binding proteins (SSBs) comprise a class of proteins that rapidly binds and sequesters these transiently formed ssDNA intermediates, primarily to protect the DNA from nucleophilic degradation among other functions. The SSB from E. coli (EcSSB) is a model protein used to study SSB function. It is a stable homotetramer and is known to wrap ssDNA with multiple modes depending on the solution conditions. Here, we directly measure the competitive binding dynamics of SSB to a long ssDNA template. At saturating concentrations (>1 nM), SSB binds to ssDNA in a biphasic manner in which a rapid
contraction of the ssDNA is followed by a slower partial elongation. Subsequently, removal of free SSB from the solution results in further contraction of the ssDNA, and addition of free SSB again elongates the ssDNA. Oscillations between the two states controlled by free protein concentration can be repeated over many cycles. Thus, the total degree of ssDNA contraction is inversely related to the free SSB concentration, indicating that competitive SSB binding inhibits further wrapping of ssDNA. This mechanism also provides compelling evidence for a minimally wrapped SSB mode, for the first time to our knowledge. We show that while SSB binding induces the unwrapping of wrapped proteins, SSB wrapping facilitates the dissociation of unwrapped proteins. We propose a general two-step kinetic model for SSB binding to ssDNA, in which the binding state of SSB is regulated through ssDNA occupancy. Finally, using the measured rates and equilibrium extensions of SSB-ssDNA complex, we derive the fundamental parameters to conduct numerical simulations, which show the proposed two-step model well describes the observed results.

Long interspersed nuclear element 1 (LINE-1 or L1) is an obligatory intragenomic parasitic element that makes up 20% of the human genome. It amplifies in the host genome by copying its RNA transcript into genomic DNA, a process called retrotransposition. L1 encodes two proteins, ORF1p and ORF2p, both associate with their parent transcript to form a ribonucleoprotein complex (L1 RNP), an essential intermediate in L1 retrotransposition. Detailed mechanistic understanding of L1 retrotransposition is sparse, particularly with respect to ORF1p, a coiled coil-mediated homotrimeric nucleic acid chaperone that can form tightly packed oligomers on nucleic acids. Although the coiled coil motif is highly conserved, it is uniquely susceptible to evolutionary change. We studied three ORF1 proteins: a modern human one (111p), its resuscitated primate ancestor (555p), and a mosaic modern protein (151p) wherein 9 of the 30 coiled coil substitutions retain their ancestral state. While 111p and 555p equally supported retrotransposition, 151p was inactive. All three proteins were indistinguishable in ensemble biochemical assays of nucleic acid interactions including nucleic acid chaperone activity, despite their in vivo functional differences. However, using single molecule DNA stretching assays, we show that 151p trimers form stably bound oligomers on ssDNA at <1/10th the rate of the active proteins, revealing that oligomerization rate is a novel critical parameter of ORF1p activity in retrotransposition conserved for at least the last 25 million years of primate evolution.
In an extended study (unpublished to date), we characterize ORF1p-nucleic acid interactions using a long ssDNA molecule. The binding kinetics of ORF1p are measured while the ssDNA is held at a constant force. We show that ORF1p binds to ssDNA in a biphasic manner, in which a rapid decrease in the extension is followed by a relatively slow increase in extension. The biphasic behavior observed here qualitatively mimics the EcSSB binding dynamics to ssDNA. Therefore, for the first time, this study provides compelling evidence that ORF1p wraps the ssDNA in multiple conformations. The ssDNA-bound ORF1p exhibits negligible dissociation within the timescale of our experiments. In addition to wrapping the ssDNA, we show that the ORF1p-ssDNA complex undergoes secondary compaction due to well-established protein-protein interactions, providing novel insights on ORF1p function.

The human immunodeficiency virus type 1 (HIV-1) nucleocapsid (NC) protein contains 15 basic residues located throughout its 55-amino acid sequence, and one aromatic residue in each of its two CCHC-type zinc finger motifs. NC facilitates nucleic acid rearrangements via its chaperone activity, but the structural basis for this activity and its consequences in vivo are not completely understood. Here, the role played by basic residues in the N-terminal domain, the N-terminal zinc finger, and the linker region between the two zinc fingers is investigated. In vitro ensemble and single-molecule DNA stretching experiments are used to measure the characteristics of wild-type and mutant HIV-1 NC proteins, and correlate these results with cell-based HIV-1 replication assay. All the cationic residue mutations lead to nucleic acid interaction defects, as well as reduced HIV-1 infectivity, and these effects are most pronounced upon neutralizing all five N-terminal cationic residues. HIV-1 infectivity in cells is correlated most strongly with NC’s nucleic acid annealing capabilities as well as its ability to intercalate the DNA duplex. Although NC’s aromatic residues participate directly in DNA intercalation, our findings suggest that specific basic residues enhance these interactions, resulting in optimal nucleic acid chaperone activity.

HIV-1 infection of humans resulted from cross-species transmission of a chimpanzee simian immunodeficiency virus (SIV). SIV NC shares 53% amino acid sequence identity with the HIV-1 protein, and its chaperone activity is characterized for the first time. Only modest differences are observed in the ability of SIV NC to facilitate reactions that mimic the minus-strand annealing and transfer steps of reverse transcription relative to HIV-1 NC, with the latter displaying slightly higher strand transfer and annealing rates. Quantitative single molecule DNA stretching studies and dynamic light scattering experiments reveal that these differences are due to significantly
increased DNA compaction energy and higher aggregation capability of HIV-1 NC relative to the SIV protein. Using salt-titration binding assays, we find that both proteins are strikingly similar in their ability to specifically interact with HIV-1 Psi RNA. In contrast, they do not demonstrate specific binding to an RNA derived from the putative SIV packaging signal.

In summary, this dissertation extensively discusses the experimental procedures, analysis methods, models, and simulations to dissect the mechanisms involved in *E. coli*, retroviral, and L1 retrotransposon replication.
# Table of Contents

Acknowledgments 2

Abstract of Dissertation 4

Chapter 1: Introduction 14

- Replication proteins from *E. coli*, the LINE 1 retrotransposon, and retroviruses. 14
- Force Spectroscopy with Optical tweezers 17
- Stretching DNA, one molecule at a time 20
- Single molecule DNA stretching to probe protein-DNA interactions 23

Chapter 2. Single molecule characterization of replication proteins in *E. coli* 30

2.1 Background .............................................................................................................................................. 30

- *E. coli* DNA Polymerase III 31
- Single-stranded DNA binding protein from *E. coli* 32

2.2 Mechanochemical characterization of *E. coli* pol III core catalytic activity ........................................ 35

- Introduction 35
- Results 37
- Pol III core activity at constant force 37
- Force-dependent polymerization and exonucleolysis velocities 40
- Concentration-dependent pause times during exonucleolysis 44
- Force-dependent pause times during exonucleolysis 46
- Pauses during polymerization 47
- Dwell times for polymerization and exonucleolysis 47
- Number of processively catalyzed base pairs during polymerization or exonucleolysis 49
- Discussion 49
- Force-dependent instantaneous velocity of pol III core exonucleolysis 53
- Force-dependent instantaneous velocity and pausing of pol III core polymerization 54
Two-step exonucleolysis initiation from concentration-dependent measurements 54
Exonucleolysis initiation is determined by primer-template junction stability 55
Materials and Methods 57

2.3 *EcSSB*-ssDNA binding dynamics regulated through competitive binding ................................ 61

Introduction 61
Results 64
EcSSB binds ssDNA in a biphasic manner at saturating protein concentrations 64
Interconversion of EcSSB-ssDNA modes is regulated through protein concentration 65
Interconversion rates between wrapped and unwrapped EcSSB-ssDNA modes 68
Force-dependence of EcSSB-ssDNA binding kinetics 70
Simulations of EcSSB-ssDNA binding dynamics 72
Discussion 73

Chapter 3. Single molecule characterization of LINE-1 ORF1p-nucleic acid interactions 77

3.1 Background........................................................................................................................................... 77

ORF1p structure 80
Nucleic acid binding properties of ORF1p 83
Nucleic acid chaperone activity of ORF1p 86

3.2 *L1* retrotransposition requires rapid ORF1p oligomerization, a novel coiled coil-dependent

property conserved despite extensive remodeling .................................................................................. 91

Introduction 91
Results 94
Comparisons of modern, ancestral and mosaic ORF1p 94
Interactions of modern, ancestral and mosaic ORF1p with oligonucleotides 96
Single-molecule measurements of ORF1p-ssDNA binding kinetics 99
ssDNA-bound ORF1p exhibits three distinct kinetic states 104
Quantifying oligomerization rates of ORF1p variants on ssDNA 106
4.3 Mechanistic differences between HIV-1 and SIV nucleocapsid proteins and cross-species HIV-1 genomic RNA recognition

Introduction 149

Results 151

Comparison of SIV and HIV-1 NC proteins and predicted secondary structures of TAR and Psi RNAs 151

Minus-strand annealing and strand transfer activities of SIV and HIV-1 NCs 153

SIV RNA interactions with HIV-1 Gag 161

SAXS reveals overall shape of SIV Psi RNA 161

Discussion 162

Materials and Methods 166

Chapter 5: Conclusions and future work 175

Conclusions 175

Future work 180

Publications 182

Appendices 183

Appendix A: UTC, a complete solution for optical tweezers data acquisition 183

Appendix B: iFextA, a complete automated solution incorporated with image processing for high-precision optical tweezers data calibration 185

Appendix C: pBacgus-11, 3' recessed, digoxigenin and biotin labeled 8 kbp DNA construct 189

Appendix D: Permission details for reprinting 192

References 193
List of Figures

Figure 1.1 : Ray diagram of an optical trap
Figure 1.2 : The schematic diagram of the dual beam optical tweezers
Figure 1.3 : Stretching DNA, one molecule at a time
Figure 1.4 : Examples of protein-DNA interactions probed by force-extension cycles
Figure 1.5 : Examples of protein/small molecule-DNA kinetic measurements
Figure 2.1.1 : DNA polymerase III holoenzyme
Figure 2.1.2 : Topology of an EcSSB bound in the (SSB)$_{65}$ mode
Figure 2.1.3 : The domain structure of E. coli SSB
Figure 2.2.1 : Experimental procedure to probe polymerase activity using optical tweezers
Figure 2.2.2 : Force-dependent catalytic activity of pol III core
Figure 2.2.3 : Bimodal catalytic velocities of pol III core.
Figure 2.2.4 : Force dependence of the pol III core instantaneous velocities
Figure 2.2.5 : Concentration-dependent exonucleolysis of pol III core
Figure 2.2.6 : Force-dependent exonucleolysis of pol III core
Figure 2.2.7 : Force-dependent pauses and catalytic bursts
Figure 2.2.8 : Temperature- and force-dependent exonucleolysis initiation
Figure 2.3.1 : Experimental Procedure to probe EcSSB-ssDNA binding dynamics.
Figure 2.3.2 : ssDNA-EcSSB binding dynamics at 12 pN.
Figure 2.3.3 : Concentration-dependent EcSSB interconversion.
Figure 2.3.4 : Force-dependant EcSSB-ssDNA binding dynamics
Figure 2.3.5 : MonteCarlo simulations of the two-step kinetic model:
Figure 2.3.6 : Proposed self regulation of protein density (SRPD) mechanism
Figure 3.1.1 : L1 Retrotransposition cycle and target-primed reverse transcription.
Figure 3.1.3 : Alignment of mouse and human ORF1p sequences
Figure 3.1.4 : ORF1p polymerization
Figure 3.1.5 : Single molecule characterization of nucleic acid chaperone activity.
Figure 3.2.1 : ORF1p variants.
Figure 3.2.2 : Binding of ORF1p variants to mismatched duplex oligonucleotide.
Figure 3.2.3 : NA chaperone activity of ORF1p variants
Figure 3.2.4 : Trimer oligomerization
Figure 3.2.5 : Single molecule analysis reveals and quantifies three ssDNA-bound ORF1p populations
Figure 3.2.6 : Representative return curves upon different incubation times with the specified ORF1p variants
Figure 3.2.7 : Direct single molecule measurements of fast and intermediate dissociation time-constants.
Figure 3.2.8 : Single molecule measurements to quantify slow, intermediate and fast ORF1p-bound ssDNA
Figure 3.2.9 : Single molecule method to quantify fractions of ssDNA-bound ORF1p populations
Figure. 3.3.1 : Experimental Procedure to probe ORF1p-ssDNA binding dynamics
Figure 3.3.2 : Concentration-switch experiments at 30 pN.
Figure. 3.3.3 : Concentration-dependence binding dynamics of ORF1p.
Figure. 3.3.4 : Force-dependence binding dynamics of ORF1p.
Figure 3.3.5 : Protein-mediated secondary compaction of ORF1p-ssDNA complex
Figure 4.1.1 : Replication cycle of a retrovirus
Figure 4.1.2 : Schematic of HIV-1 virion organization
Figure 4.2.1 : Sequence of WT HIV-1 NC
Figure 4.2.2 : FEC of DNA in the presence-HIV-1 NC:
Figure 4.2.3 : DNA stretching (solid line) and relaxation (dotted line) in the presence (red) and absence (black) of HIV-1 NC cationic mutants:
Figure 4.2.4 : Concentration-dependence of transition slope and hysteresis
Figure 4.3.1 : Sequence and structural features of HIV-1 and SIV NC proteins.
Figure 4.3.3 : Kinetics of minus-strand annealing with SIV and HIV-1 substrates in the presence of SIV and HIV-1 NC proteins.
Figure 4.3.4 : Kinetics of SIV minus-strand transfer in the presence of SIV and HIV-1 NC proteins
Figure 4.3.5 : Force-extension curves for dsDNA in the presence of 30 nM SIV NC or HIV-1 NC
Figure 4.3.6: : Force-induced compaction by protein-DNA interactions.
Figure A1 : Main controller panel of the UTC
Figure A2 : Extension-time panel of UTC
Figure A3 : UTC camera module
Fig. B1 : A representative screen shot from a data analyses example from iFexta
Fig. B2 : Image analysis module of iFexta
Chapter 1: Introduction

Replication proteins from *E. coli*, the LINE 1 retrotransposon, and retroviruses.

The genetic code is the blueprint of life and is stored and protected in the form of a double stranded DNA (dsDNA) helix. Replication of the genetic code is one of the most fundamental processes in biology. In this dissertation, DNA replication processes are discussed and approached in the context of nucleic acid interactions to elucidate replication mechanisms involved in bacteria, retroviruses, and the long interspersed nuclear element-1 (LINE1 or L1) retrotransposon from mammalian cells.

*E. coli* DNA Pol III: The second chapter discusses the characterization of the replicative DNA polymerase III (pol III) in *E. coli*\(^1\)\(^-\)\(^6\). Replicative DNA polymerases are the molecular machines that are primarily responsible for duplicating the genetic code before cell division. While the genetic code of a bacterium is on the order of millions of base pairs long, the human genome contains several billions of base pairs. Therefore, the duplication of the genetic code requires not only proceeding faithfully but also operating with high efficiency. To achieve this, the typical replicative DNA polymerase mechanisms are evolved to be more complex than the other polymerase mechanisms that are involved primarily in damage repair\(^7\). These polymerases also possess the ability to proofread during nucleotide incorporation, a function that significantly enhances the fidelity of replication. Proofreading is achieved by an enzyme called exonuclease that begins to excise newly synthesized nucleotides after incorporation of an incorrect nucleotide by the polymerase subunit. How are these two complementary functions regulated? How does
the polymerase ‘sense’ that it has made a mistake and switch to its proofreading mode? Chapter two contains a comprehensive analysis of our work to address these questions.

**E. coli Single-stranded DNA binding protein:** Although the genetic code is stored and protected in the form of dsDNA, during DNA replication as discussed above, or during DNA recombination and repair, these duplexes are perturbed to transiently generate long stretches of ssDNA that are vulnerable to nucleophilic degradation. Single-stranded DNA binding proteins (SSBs) rapidly bind and sequester these single stranded intermediates, primarily to protect the DNA from degradation, among other functions. The SSB from *E. coli* (EcSSB) is a model protein that has been extensively used to study SSB function. EcSSB is a homotetramer in which each subunit contains a single-stranded DNA binding domain (also known as oligonucleotide or OB domain). Due to its structure, EcSSB can wrap the substrate ssDNA with multiple topologies. In the third part of chapter two, I discuss our most recent work (unpublished to date) that investigates the competitive binding of EcSSB to ssDNA. Our results not only propose a general two-step kinetic model for the SSB binding, but also discover surprising new properties of EcSSB that will significantly enhance the understanding of its function.

**ORF1p of the L1 retrotransposon:** The chapter three discusses our studies on the protein encoded by the open reading frame 1 (ORF1p) of the L1 retrotransposon. The L1 retrotransposon is a 6 kbp parasitic transposable element that has been reshaping and amplifying in the mammalian genome for millions of years. While L1 itself makes up 20% of the human genome, owing to L1’s activity, it is responsible for generating more than 40% of the human genome. Although a majority is epigenetically silenced, a few L1 copies remain active, causing detrimental effects. L1 autonomously replicates (L1 retrotransposition) with the help
of two self-encoded proteins, ORF1p and ORF2p\textsuperscript{34}. These two proteins associate with their parent transcript to form a ribonucleoprotein complex (L1 RNP). The L1 RNP mediates the insertion of a new L1 copy at a target site in the genomic DNA via a unique mechanism called target primed reverse transcription (TPRT), during which ORF2p functions as the reverse transcriptase and the endonuclease\textsuperscript{35-38}. However, detailed mechanistic understanding of this process is sparse, especially with respect to the function of ORF1p. Chapter three presents a comprehensive discussion on our published and unpublished work to date, which aims to elucidate ORF1p function in the L1 retrotransposition.

**Nucleocapsid proteins from HIV-1 and SIV:** In chapter four, I discuss two projects on characterizing the nucleocapsid proteins from human immunodeficiency virus-1 (HIV-1) and simian immunodeficiency virus (SIV). The HIV-1 and HIV-2 lentiviruses are the culprits of the deadly Acquired Immune Deficiency Syndrome (AIDS) outbreak that emerged in the late 20th century and infected more than 75 million people by the year 2014\textsuperscript{39-41}. It was later discovered that HIV-1 infection of humans resulted from cross-species transmission of a chimpanzee simian immunodeficiency virus (SIVcpz), a recombinant virus generated from two distinct monkey SIV lineages\textsuperscript{42,43}. Like other retroviruses, SIV and HIV-1 have a nucleocapsid protein (NC), a small basic structural protein containing two zinc-binding domains. NC possesses nucleic acid chaperone activity, by which it facilitates the rearrangement of nucleic acids into conformations that are thermodynamically more stable than the original structure\textsuperscript{44-48}. Chapter four contains a comprehensive analysis on the functional role of the basic residues found in HIV-1 NC. Furthermore, in-depth analysis is discussed on the chaperone activities of SIV and HIV-1 NCs in the context of biologically relevant reactions.
**Force Spectroscopy with Optical tweezers**

Optical tweezers enable applying forces in piconewton range to micron-scale dielectric objects while simultaneously measuring displacement with nanometer-level precision, and hence characteristically applied to study structures spanning from molecular motors to colloidal systems to polymers and biopolymers at the single-molecule level\textsuperscript{43,49-51}. The birth of optical tweezers dates back to 1970s with the detection of optical scattering and gradient forces on micron sized particles by Arthur Ashkin who pioneered this technology. His work on levitating micron-sized dielectric particles in both water and air, and the stable, three-dimensional trap with counterpropagating laser beams\textsuperscript{52,53} led to the development of the optical tweezers, originally known as “the single-beam gradient force optical trap”\textsuperscript{54,55}. Not only did he develop this remarkable tool but also employed it in experiments ranging from trapping neutral atoms to live bacteria\textsuperscript{56,57}, winning the 2018 Nobel prize in physics for his seminal work on “optical tweezers and their applications in biological systems”.

The phenomenon of optical trapping exploits interaction of light with dielectric materials of which the sizes are similar to the wavelength of light. The incident photons impart a momentum flux and hence a force on the dielectric particle. The force on a dielectric object induced by refraction of the light by the object, is given by\textsuperscript{53}

\[
\vec{F} = \left( \frac{n}{c} \right) \int \left( \vec{S}_{\text{in}} - \vec{S}_{\text{out}} \right) dA
\]

(1.1)

where \(n\) is the index of refraction, \(c\) is the speed of light, \(\vec{S}\) is the Poynting vector, and \(dA\) is an element of area normal to \(S\). When a laser beam is tightly focused using a microscope objective lens with a high numerical aperture, the resulting optical force can be represented by two components; first, the scattering force in the direction of the propagating beam, originated from both the absorption and specular reflection by the trapped object and second, the gradient force,
in the direction of spatial intensity gradient of the beam, originated by the light intensity
gradient. For a three-dimensional stable trapping, the gradient force component, which pulls the
dielectric particle towards the focus of the beam should surpass the scattering force component
pushing it away from the focus. In order to satisfy this condition, a very steep gradient in the
light produced by sharply focusing the trapping laser beam is required.

The optical tweezers were originally proposed as an “atom trap”, which operates in the Rayleigh
scattering regime; where the dielectric particle is smaller than the wavelength of light. It was
shown that the optical tweezers can be used to cover the full spectrum of Mie and Rayleigh
regimes\(^\text{58}\), that is the spectrum of particles which are larger and smaller than the wavelength of
light.

The particle acts as a point dipole when its diameter is smaller than the wavelength of the light.
The scattering force component arises due to the absorption and retardation of light by the dipole
and is pointing in the direction of the incident light. The gradient force component arises
pointing in the direction of the intensity gradient of the light from the interaction of the induced
dipole with the inhomogeneous field. When the dielectric particle is larger than the wavelength
of light (Mie scattering regime), simple ray optics can be used to describe the forces involved in
a stable trap. Fig. 1.1 shows how the resulting force facilitates stable trapping of the dielectric
particle for arbitrary displacements. As qualitatively depicted in Fig. 1.1, the edges of the
objective lens have a higher contribution to the trapping force, which emphasizes the
requirement of a microscope objective with a high numerical aperture in order to generate higher
trapping forces.

When two counter-propagating laser beams are used with two microscope objectives facing each
other and focus the two separate laser beams to the same spot, a stable trap can be created with
the ability to generate higher trapping forces for a given laser power with lower numerical aperture microscope objectives.

Figure 1.1: Ray diagram of an optical trap. The schematic of a trapped spherical dielectric particle in the Mie scattering regime represented with a pair of rays a and b. The net restoring force $F$ directs the sphere back to the focus when it undergoes arbitrary displacements as shown in A and B. The net gradient force due to refraction is enhanced in the dual beam trap due to counteracting scattering forces exerted by reflection.

This dual-beam optical tweezer design is able to cancel out the force due to reflection ($F_{reflection}$) at the surface of the dielectric particle, since that is approximately the same for each laser. This design demands high precision alignment of the two laser beams, within less than the diameter of the particle. Care should be taken to correct for errors due to drifts in the relative alignment of the laser beams. Fig.1.2 shows a schematic of the dual-beam optical tweezer system that I rebuilt with improved performance and used for the majority of the experiments discussed in this dissertation.

The optical trap can be considered as a Hookean spring for small enough displacements of the dielectric particle, in which the characteristic stiffness is proportional to the light intensity. In this case the gradient restoring force is proportional to the offset from the equilibrium position and hence measuring the position of the bead in the trap determines the force on the bead in the trap. The stiffness of the trap depends on the design of the optical tweezers and the size of the sphere. The stiffness of the optical tweezers system used in this study is $\sim 100$ pN/µm for a 3-µm
polystyrene bead. In the Williams lab, we have harnessed the capabilities of optical tweezers to implement force spectroscopy studies to dissect the nucleic acid interactions involved in the machinery of life to address important biological questions.

Figure 1.2: The schematic diagram of the dual beam optical tweezers system at the Williams lab. Two counterpropagating laser beams are focused with a pair of water immersion high NA objectives, forming an optical trap inside a custom-made flow cell. The flow cell is composed of five inlet tubes on one end and an outlet tube on the other end and is fixed to a piezoelectric stage with sub-nanometer translation resolution. The piezoelectric stage is digitally interfaced with pulsed frequency modulation and differential quadrature encoding to achieve 0.1 nm spatial resolution. The cMOS cameras capture the beads inside the flow-cell and integrate to data acquisition software for simultaneous image acquisition. Position sensing diodes (PSDs) detect and send out an analog signal of the laser deflections on the trap, which measures the force exerted on the trapped bead.

**Stretching DNA, one molecule at a time**

A typical single molecule force spectroscopic experiment, which uses optical tweezers to probe how a protein or other ligand interacts with DNA, begins by tethering a termini-labeled DNA molecule between two functionalized beads. Depending on the experiment, one of three DNA constructs was used in the work discussed in this dissertation; a 48.5 kbp λ DNA labeled with biotin at both the 3′ termini, a 38.5 kbp λ DNA that is biotinylated at the 3′ and 5′ termini on the same strand with a 3′ recessed end, and an 8.1 kbp linearized pBACgus-11 DNA that is attached to a Digoxigenin (DIG)-labeled 100 bp dsDNA handle at one terminus and a
biotinylated 5′ terminus with a 3′ recessed end. The biotinylated and DIG-labeled termini are tethered to streptavidin and anti-DIG functionalized beads, respectively.

Force as a function of the end-to-end extension or the force-extension cycle (FEC) is the most fundamental measurement in single molecule DNA stretching experiments. At very low forces (< 6 pN) the dsDNA is coiled, and the measured force gradually increases as it uncoils. At 30 pN, the extension approaches the dsDNA contour length; the backbone resists further extension and the force increases dramatically as an elastic response. The overstretching transition of a bare double-stranded (ds) DNA occurs at ~62 pN and is observed as a plateau, a rapid increase of the extension in the narrow 60–65 pN force regime in the FEC. Although this plateau may represent the transition from B-form dsDNA into S-DNA in high salt conditions (>0.15 M), at salt conditions used in the studies discussed in this dissertation, the plateau primarily represents the helix-coil transition due to force-induced melting of the dsDNA. The return curve of the FEC of the dsDNA is reversible, although it may exhibit a slight discrepancy, or hysteresis, as it approaches the forces below the overstretching transition, depending on the solution conditions and the quality of the DNA molecule. The small hysteresis observed is primarily because the delayed reannealing of the end-peeled DNA segments. However, if the DNA molecule is damaged and contains nicks in its backbone, multiple fraying events will result in a large hysteresis.

The FEC of ssDNA is also almost completely reversible, nevertheless, at high monovalent salt conditions or in the presence of Mg$^{2+}$ a hysteresis that is indicative of secondary structures is observed at forces lower than 10 pN. At forces higher than 6 pN, the end-to-end extension of an ssDNA molecule is significantly larger than a dsDNA molecule with the same number of bases,
but due to the heavily coiled nature, the extension of a ssDNA below 6 pN is shorter than that of an equivalent dsDNA molecule.

![A](image1.png) ![B](image2.png)

**Figure 1.3: Stretching DNA, one molecule at a time.** (A) Schematic depiction of a single DNA molecule tethered in the optical tweezers system. (B) Force-extension curves of a dsDNA (blue) and a ssDNA (yellow). The respective lines are the extensible WLC and FJC polymer models.

Polymer models that characterize the force dependence of the end-to-end extension of DNA molecules are often used in single molecule analyses. Although these models may not completely describe DNA properties, they fit the FECs within uncertainty in the force-ranges that are used in the analyses in the studies described here, providing useful physical parameters that approximate the long length scale of the average polymer behavior. The extensible Worm-like chain (WLC) model describes the extension $b_{ds}(F)$ of a dsDNA molecule as a function of force, where $^{62}$

$$b_{ds}(F) = B_{ds} \left\{ 1 - \frac{1}{2} \left( \frac{k_b T}{FP_{ds}} \right)^{1/2} + \frac{F}{S_{ds}} \right\}$$

(1.2)
Similarly, the extensible freely-jointed chain (FJC) model describes the extension of an ssDNA (b_{ss}(F)) as a function of force, where:

\[
b_{ss}(F) = B_{ss} \left[ \coth \left( \frac{2 P_{ss} F}{K_B T} \right) - \frac{1}{2} \frac{k_B T}{P_{ss} F} \right] \left[ 1 + \frac{F}{S_{ss}} \right].
\]

Here S is the stretch modulus (S_{ds} = 1361 pN, S_{ss} = 720 pN), P is the persistence length (P_{ds} = 45 nm, P_{ss} = 0.75 nm) and B (B_{ds} = 0.34 nm/bp, B_{ss} = 0.55 nm/bp) is the contour length and the ss and ds subscripts denotes ssDNA and dsDNA, respectively.

### Single molecule DNA stretching to probe protein-DNA interactions

There are several methods by which single molecule DNA stretching can be utilized to characterize protein-DNA interactions. The choice of method depends primarily on the protein or ligand of interest and the information one seeks to learn about the molecule. In force spectroscopy experiments, irrespective of the system being in or out of equilibrium, the interactions of the protein are captured by the alterations induced on the substrate DNA because of the protein’s mechanism of action. Although the term “protein” is frequently used, it generally applies for any ligand or enzyme that interacts with DNA. Typically, alterations of DNA properties by proteins are accompanied by length changes at the binding site. Therefore, these interactions are affected by the force exerted on the DNA molecule, providing an additional variable to characterize such protein-DNA interactions.

One of the commonly used methods to characterize protein-DNA interactions is to examine the FEC of a DNA molecule in the presence of the protein of interest. Because the FEC of a DNA molecule is extremely sensitive to the presence of DNA binding proteins, significant variation in the FEC is observed when the protein binding results in DNA deformation. The FEC may or may
not be in equilibrium, depending on the solution conditions and the data acquisition parameters. Therefore, the alterations in FEC may not only provide information on the binding conformation, but may also reflect binding kinetics, allowing us to quantitatively characterize the protein-DNA interactions. Because a protein-free dsDNA molecule exhibits many transitions during the FEC (Fig. 1.3B), the alterations in each of these transitions in the presence of protein will yield distinct protein-DNA binding properties.

**Figure 1.4: Examples for protein-DNA interactions probed by force-extension cycles.** The FECs of a dsDNA in the absence of proteins are shown in black. Solid and empty lines represent stretch and return curves of the FECs, respectively. (A) FECs of dsDNA in the presence of HMGB1 protein with force-induced melting transition occurring at higher forces than a free dsDNA. The additional shift in overstretching transition characterize the properties of protein-DNA interactions (B) FECs of which the overstretched (~70 pN) DNA molecule is incubated with ORF1p for different time periods as denoted in the legend to the figure. During the incubation the protein binds to the ssDNA regions generated by force-induced melting and thereby inhibits duplex-formation during the return curve. This is registered as a hysteresis, which in turn is a precise quantitative measure of the protein-bound ssDNA fraction. FECs in the presence of 30 nM SIV NC and HIV-1 NC. The increase in the DNA transition slope is a characteristic measure the of nucleic acid chaperone properties of the protein.
For instance, as shown in Fig. 1.4A, in the presence of dsDNA stabilizing proteins such as the high mobility group box 1 (HMGB1) protein, the force at which the overstretching transition occurs will rise as the force required to melt the DNA is increased. The force shift in the overstretching transition reflects the additional force required to disrupt the protein binding, thereby providing a precise quantitative measure to probe the DNA binding characteristics of the protein.

Binding of proteins to ssDNA can be characterized by allowing the protein to bind at forces above the melting transition. Protein binding to ssDNA will prevent reannealing and result in a significantly large hysteresis in the FEC. The relative extent of the FEC’s hysteresis in the presence of the protein is a quantitative measure of the protein-bound ssDNA fraction. Fig. 1.4B shows an example of such a study to characterize the oligomerization properties of L1 ORF1p when interacting with ssDNA (discussed in detail in chapter 3).

The most distinct feature of the changes induced by the presence of nucleic acid chaperone proteins such as L1 ORF1p or retroviral nucleocapsid (NC) proteins (discussed in detail in chapter 4) in the FEC is the significant increase in the slope of the overstretching transition due to protein binding. Nucleic acid chaperone proteins, such as retroviral nucleocapsid (NC) proteins and L1 ORF1p, rearrange nucleic acids to achieve the optimal number of base pairing. The overstretching transition is an empirical quantity defined to be the force interval over which the DNA molecule transitions from an entirely ds- to primarily ss- state at low salt solution conditions (<0.15 M). In the absence of DNA-binding ligands this transition occurs in a highly cooperative manner. In the presence of nucleic acid chaperone proteins, the overstretching transition width is significantly increased. Although to date, we do not completely understand the exact molecular mechanism by which the overstretching transition width is
increased by the protein, this provides a useful quantitative measure to evaluate nucleic acid chaperone properties such as rapid binding kinetics and nucleic acid rearrangements. For instance, altering the overstretching transition must require the ability to bind both ss- and ds-forms of the DNA with rapid kinetics. Increasing the overstretching transition width must presumably lower the energy barrier of rearranging nucleic acids and thereby reduce the cooperativity of the transition from the ds- to ss- forms of DNA.

The ability to directly measure the kinetics of protein-DNA interactions is one of the most attractive and powerful features of single molecule DNA stretching experiments. Because the protein deforms the DNA upon binding, a binding event or an unbinding event will alter the length at the binding site, which in turn will result in a sudden change in the force. Because the end-to-end extension of DNA remains a constant at a given force, the extension change due to a binding event will also remain unchanged at a given force. In other words, the substrate-DNA construct can be visualized as a long lattice in which the binding-site size remains a constant at a given force. Therefore, to accurately characterize the kinetic measurements, the sudden force changes caused by these binding events are typically transformed to an extension change via a force feedback mechanism, thus maintaining a constant force throughout the substrate DNA (force-clamped experiments).

While the sign of the extension changes (whether increasing or decreasing) upon binding and unbinding provides information on the nature of the deformation caused by the protein, the temporal evolution of the extension of such an experiment is a direct measurement of the binding dynamics of the protein to the DNA molecule.
**Fig 1.5: Examples for protein/small molecule-DNA kinetic measurements.** Free dsDNA FEC is shown in black (left panels) (A) DNA elongation at a constant force in the presence of metalorganic ruthenium threading intercalator, Δ,Δ-B, is shown in red. (B) DNA elongation as a function of time in the presence of 0.6 (violet), 2.3 (blue), 6.0 (green), 12 (yellow), 25 (orange), and 90 (red) nM Δ,Δ-B at a constant force of 29.8 ± 0.3 pN. (C) A partially single-stranded dsDNA molecule (red) in the presence of 0.2 μM pol III core. Changes in extension due to dsDNA-ssDNA conversion at different constant forces are shown in the multicolor profile. The arrows represent the direction of the pol III core velocity, indicating exonuclease (exo) or polymerization (pol). (D) Bottom: Extension time profiles corresponding to colors in (A), at constant forces shown on the top panel. (E) Extension-change at a constant-force (green, also zoomed in inset) due to E. coli SSB (EcSSB) binding to preformed ssDNA (orange). (F) Extension-time profiles that probe the EcSSB-ssDNA binding dynamics at the constant forces shown in the legend.
For instance, an intercalating agent will bind and insert itself in between the base pairs of a dsDNA molecule. This process will elongate the dsDNA backbone and thus the binding and unbinding events will register as positive and negative length-changes of the DNA molecule, respectively, at a constant force. Fig. 1.5A is an example of such characterization for the small metalorganic ruthenium threading intercalator (ΔΔ-B).

Single molecule DNA stretching can also be used to study the catalytic activity of DNA or RNA polymerases (Fig. 1.5C-D) that alter the DNA substrate itself by nucleotide synthesis and/or excision (discussed in detail in Chapter 2). Unlike single molecule protein-binding assays where the substrate DNA is deformed due to protein binding, in these experiments the DNA substrate is transformed between its ds and ss forms, by the addition or excision of nucleotides, through the catalytic activity of the enzymes. Because mechanical work is involved during the catalysis (synthesis or excision), the enzymatic activities are typically studied at a constant template tension at which the length change associated with transforming a single nucleotide position between its ss and ds forms is also well defined. In addition, it has been shown that the template tension also can be used as a “mechanical switch” to alternate between the complementary nucleotide- excision or synthesis functions, for several DNA polymerases.

Fig. 1.5C shows how the template tension alters the catalytic activity of the E. coli DNA polymerase III core subassembly.

In the Williams lab we have extensively used FECs and constant force measurements to study protein binding to dsDNA over many years. We have recently developed methods and protocols to efficiently conduct these dynamic and equilibrium studies on preformed ssDNA (see chapter 2.3.3, and appendix C). This not only allows us to accurately study the binding dynamics of proteins such as APOBEC3 proteins and SSBs (Fig. 1.5E) of which ssDNA is the primary
substrate, but also provides the means to deconvolve and isolate the ssDNA and dsDNA binding properties of proteins such as L1 ORF1p and retroviral NC proteins, which exhibit complex nucleic acid interactions while binding to both ds and ss forms of DNA.
Chapter 2. Single molecule characterization of replication proteins in *E. coli*

*The work described in this chapter is conducted in collaboration with Beuning Lab, Northeastern University and Ioulia Rouzina at The Ohio State University, Columbus, OH*

2.1 Background

Replicative DNA polymerases duplicate chromosomal DNA during cell division. Chromosomal DNA molecules are extremely large, even in bacteria, and the vast amount of information they carry must be replicated with high accuracy to sustain life. Each time a cell divides, the entire length of its chromosomal DNA is replicated. The enzymes that are largely responsible for this are known as DNA polymerases. Elucidating the structure, function, and catalytic activity of these molecular motors is essential to an understanding of the complex mechanisms of DNA replication.

![Figure 2.1.1: DNA polymerase III holoenzyme (dimer form) at a replication fork.](image)

Figure 2.1.1: DNA polymerase III holoenzyme (dimer form) at a replication fork. Figure reprinted from with permission. The two polymerase cores, which are tethered to the β clamps, contain the three subunits α, ε, and θ. The γ complex, assembled with two τ subunits, couples the polymerization of both the leading strand and the lagging strand. The single-stranded lagging strand is threaded through DnaB helicase and is coated with SSB. Primase (not shown) synthesizes the RNA primers (red) on the lagging strand. The ψ and χ subunits are not shown but would connect the γ complex to SSB. Upon completing replication of the lagging strand Okazaki fragment, the polymerase core and β clamp are loaded onto the next SSB-coated lagging strand template for synthesis of the next Okazaki fragment.
DNA polymerases catalyze the addition of deoxynucleoside triphosphate (dNTP) units to the DNA backbone in DNA replication. The addition of the dNTPs occurs directly on the DNA template strand, and the base of the new dNTP is complementary to the base on the template strand. Since bases are added to the 3' end of the nascent strand, the polymerization reaction must proceed in the 5' to 3' direction. The tertiary structure of DNA polymerase is such that the enzyme fits over the previously formed base pairs \(^{60}\). These bases must be paired correctly for the polymerase to adopt its functional conformation \(^{86-88}\). This is a vital part of the enzyme’s fidelity because it enables proofreading \(^{89}\).

**E. coli DNA Polymerase III**

DNA polymerase III (pol III) is the replicative DNA polymerase in *E. coli* \(^ {1,2,86}\). The polymerase subunit α is a C family DNA polymerase, a family that is found only in prokaryotes \(^ {3}\). Pol III is an asymmetric dimer or trimer that synthesizes the leading and lagging strands simultaneously at the replication fork (Fig. 2.2.1) \(^ {1,2}\). A helicase unwinds the double-stranded DNA (dsDNA) into two anti-parallel template strands. After primase synthesizes the RNA primer, DNA pol III replicates the leading strand continuously and the lagging strand in Okazaki fragments in the 5' to 3' direction \(^ {1-3,90}\). DNA pol III has an extremely high catalytic rate, at 10\(^3\) bases/sec \(^ {86,91}\), and high fidelity, with error frequencies of approximately 10\(^{-5}\)/bp before proofreading \(^ {91,92}\) and 10\(^{-8}\)/bp with proofreading \(^ {93}\). In the presence of the β clamp, which allows α to remain bound to its DNA substrate for many catalytic cycles, α exhibits very high processivity \(^ {91}\).

Although the dimer is asymmetric to allow simultaneous polymerization of both template strands, the core of each branch consists of the same α ε θ complex (pol III core) \(^ {94}\). The α subunit is the DNA polymerase and the ε subunit is the 3' to 5' proofreading exonuclease \(^ {91}\). The exact role of the θ subunit is still to be determined, but its presence increases the accuracy of pol III.
and it has been suggested to stabilize the interaction between α and ε\textsuperscript{91,95,96}. The β subunit is a sliding DNA clamp, which is responsible for the high processivity of the complex\textsuperscript{97}. The clamp loader complex, consisting of combinations of the τ, γ, δ, and δ’ subunits, loads the β clamp onto DNA and with χ and ψ coordinates synthesis on the leading and lagging strands\textsuperscript{2,91,92,94,98}.

**Single-stranded DNA binding protein from *E. coli***

Single-stranded DNA binding proteins (SSBs) form a family of proteins, found in all domains of life, that binds transiently-exposed ssDNA regions with high affinity and protects (among other functions) the ssDNA during genomic maintenance processes such as DNA replication, recombination and repair\textsuperscript{8-19}.

![Figure 2.1.2: Topology of the ssDNA-EcSSB in (SSB)\textsubscript{65} mode](image)

**Figure 2.1.2: Topology of the ssDNA-EcSSB in (SSB)\textsubscript{65} mode:** Crystal structure (PDB ID : 1EYG)\textsuperscript{23} and schematic representation of an EcSSB bound to ssDNA in the (SSB)\textsubscript{65} mode, resembling the seams of a tennis ball. Individual yellow, purple, green and red subunits (open access figure reprinted from\textsuperscript{99}).

The SSB from *E. coli* (EcSSB) is a homotetramer (Fig. 4.1.2) in which each subunit (19 kDa) contains an oligonucleotide binding domain (OB-fold) that can interact with ssDNA\textsuperscript{23}. For this reason, it is able to bind and wrap ssDNA with multiple topologies such as (SSB)\textsubscript{65}, (SSB)\textsubscript{35}, and (SSB)\textsubscript{17}, where the subscripts denote the numbers of nucleotides occluded by the tetramer\textsuperscript{99,100}.

Although the crystal structure of a C-terminal truncated EcSSB in the (SSB)\textsubscript{65} mode was solved to show that the ssDNA is wrapped around the tetramer in a manner resembling the seams of a tennis ball\textsuperscript{23} (Fig. 2.1.2), precise topological information on other binding modes is not known to
date. EcSSB can interconvert between these modes depending on the solution conditions such as salt concentration and type, pH, and temperature, as well as the protein density on the substrate DNA\textsuperscript{14,100}. In addition, recent single-molecule analyses revealed that these distinct binding modes exist in a dynamic equilibrium, with the ability to diffuse along the ssDNA substrate without dissociation\textsuperscript{24,25}.

![Figure 2.1.3: The domain structure of EcSSB](image)

**Figure 2.1.3: The domain structure of EcSSB**: Schematic depictions of the OB-fold containing N-terminal domain, the C-terminal domain (CTD) that contains the intrinsically disordered linker (IDL) and the conserved TIP. The numbers represent the amino acid residue positions of domain boundaries. The amino acid composition of the TIP is denoted in green text.

The structure of EcSSB is organized with the OB-fold-containing N-terminal domain that interacts with ssDNA, which also mediates tetramer-formation, a C-terminal conserved tip (TIP) that facilities interactions with other proteins, and a non-conserved intrinsically disordered linker (IDL). The 9 amino acid C-terminal TIP is known to mediate the recruitment of SSB-interacting proteins (SIPs) to the ssDNA during genome maintenance\textsuperscript{26}. Although the function of the IDL is not completely understood, it has been shown that this region is crucial for EcSSB inter-tetramer interactions\textsuperscript{22,101}.

Although extensive studies have significantly improved the understanding of the EcSSB-ssDNA binding dynamics, several formidable biological questions remain to be addressed. These include the following: What is the molecular mechanism by which the EcSSB interconverts between its modes? What are the functional roles of the distinct binding modes of EcSSB? What is the
mechanism by which a tightly wrapped EcSSB eventually dissociates to ensure its transient role?
The studies described here provide significant new insights into the answers to these questions.
2.2 Mechanochemical characterization of *E. coli* pol III core catalytic activity

*The work described in this section is adapted from the following publication.*


*Pol III core purification and ensemble results are the work of Dave Murison in the laboratory of Penny Beuning, at Northeastern University, Boston, MA*

**Introduction**

During replicative polymerization, tight coordination between the polymerization and exonucleolysis cycles is expected to exist, to permit efficient and faithful replication. It has been shown that mutations that lead to a loss in fidelity during *E. coli* replication are found in the *dnaQ* gene, which encodes the ε subunit.\(^{90,102,103}\) However, the molecular mechanism of the switching between the polymerase and exonuclease subunits is poorly understood. Elucidating the structure, function, and catalytic activity of these molecular motors is essential to understand the complex mechanisms of DNA replication. Here we report a single-molecule approach to manipulate these molecules and characterize the dynamics of the pol III core polymerization and exonucleolysis.

We observe that the mechanical tension applied to the substrate DNA promotes the switching between exonucleolysis and polymerization functions, which agrees with previous single molecule studies on DNA polymerases Klenow Fragment, T7 gp5, and φ29 DNA polymerase.\(^{84,104,105}\) The force dependence of T7 polymerization velocity is modeled as a function of the free energy change involved in ssDNA-dsDNA conversion. The kinetic scheme proposed for φ29 DNA polymerase describes the intra-molecular primer transfer as a consequence of a conformational change in the φ29 pol - DNA assembly induced by the applied
tension on the DNA template\textsuperscript{106}. The key difference between pol III core and these polymerases is that the editing and polymerization activities of pol III core are carried out by distinct subunits, $\varepsilon$ and $\alpha$, respectively. Hence the primer transfer between the catalytic exo and pol domains occurs intermolecularly. In addition, the exo activity of isolated $\varepsilon$ is similar to that of pol III core and $\varepsilon$ is considered to be a highly efficient 3′-5′ exonuclease, capable of functioning independently of $\alpha$\textsuperscript{105,107}.

![Diagram of experimental procedure](image)

**Figure 2.2.1: Experimental procedure to probe polymerase activity using optical tweezers.** (A) Linearized pBACgus11 is ligated with a DIG-dsDNA handle and a biotinylated oligonucleotide at its free ends, providing a single primer-template junction for pol III core to bind. At forces above 30 pN, exonucleolysis is observed. Conversion from dsDNA to ssDNA upon the excision of nucleotides is registered as an increase in extension. Similarly, at forces below 30 pN incorporation of nucleotides due to polymerization is registered as a decrease in extension. (B) Force-extension curves of dsDNA (blue) and ssDNA (red). Stretch and release are represented as solid and empty circles, respectively. The ssDNA is obtained by completely removing one strand of the dsDNA via exonucleolysis. Solid lines represent the theoretical polymer models: extensible wormlike chain\textsuperscript{108} for dsDNA and extensible freely jointed chain\textsuperscript{109} for ssDNA. The arrows show the direction of extension change of exonucleolysis and polymerization at constant force experiments.

The source of this exonucleolytic editing specificity is found to be the greater melting capacity of a mispaired 3′ terminus for both isolated $\varepsilon$ and pol III core\textsuperscript{110,111}. According to previous bulk biochemical assays, the exonuclease activity of both $\varepsilon$ and pol III core is more efficient with
Furthermore, a two-step kinetic scheme for the exonuclease reaction of isolated ε subunit suggests that the physiologically relevant substrate for the ε subunit within the holoenzyme complex is ssDNA at least three nucleotides in length111.

Here, we investigate the force dependence of pol III core polymerization and exonucleolysis. We are able for the first time to characterize these individual catalytic events on a single primer-template DNA substrate. We propose a two-state reaction scheme to describe the rate of force-induced exo initiation. According to our model, pol III core bimolecularly binds at the primer-template junction and subsequently transforms to an exo-active conformation that is strongly affected by the applied template force. We show that this result is in quantitative agreement with the previously measured temperature-dependence of exo-activity. This analysis shows that the intermolecular switching of the primer between the polymerase and exonuclease subunits is a thermally driven process governed by destabilization of the primer-template junction, rendering it more susceptible to exonuclease binding.

Results

Pol III core activity at constant force

We used optical tweezers to characterize the dynamics of pol III core activity at the single molecule level. Both polymerization and exonuclease activity were measured at constant applied tensions on a single DNA substrate. To do this, a single dsDNA molecule with a 3’ recessed end (~30 nt) was attached by its covalently-labeled free ends to polystyrene spheres, one held in an optical trap and the other immobilized on the end of a glass micropipette (Fig. 2.2.1A). By gradually moving the micropipette, the applied mechanical tension and the extension of a single DNA molecule was measured. In the absence of protein, an approximately constant force phase transition, referred to as DNA overstretching, is observed112. At the low-salt conditions used in
these experiments, this transition occurs at about 62 pN and represents a conversion of DNA from dsDNA to ssDNA as the DNA is destabilized by force and primarily peels from its free end\textsuperscript{113-117}.

![Figure 2.2](image)

**Figure 2.2: Force-dependent catalytic activity of pol III core** (A) Force-extension curves of a bare dsDNA molecule (black) and a partially single-stranded dsDNA molecule (red) in the presence of 0.2 μM pol III core. Changes in extension due to dsDNA-ssDNA conversion at different constant forces are shown in the multicolor profile. The arrows represent the direction of the pol III core velocity, indicating exonucleolysis or polymerization. (B) Temporal trajectories of the change in extension corresponding to the shown colors in (A) at the given constant forces (red).

The force-extension profiles of ssDNA and dsDNA cross at ~6 pN (Fig. 2.2.1B). Pol III core activity was measured at forces greater than this crossover force, at which ssDNA is longer than dsDNA. Therefore, at constant forces below the melting transition, conversion between ssDNA and dsDNA is registered as an increase in extension due to pol III exonucleolysis (exo) and a decrease in extension due to pol III core polymerization (pol) activity (Fig. 2.3.1B). To measure pol III core activity, we introduced purified pol III core at fixed concentration to the flow cell containing a single DNA molecule captured between the beads as shown (Fig. 2.2.1A).
DNA was maintained at a constant force through a feedback loop and the change in DNA length at constant force was determined as a function of time. The number of nucleotides incorporated or excised as a function of time was then obtained by dividing the observed change in extension by the expected change in extension at a given force accompanying the conversion of one single-stranded nucleotide into its double-stranded counterpart, as described in Materials and Methods.

As shown in Fig. 2.2.2, polymerization is observed at forces below 30 pN and exonuclease activity is observed at forces higher than 30 pN. From the measured extension vs. time trajectories, we obtained a distribution of instantaneous velocities for both exo and pol activity at several forces by applying a moving average filter as described in Materials and Methods. An example of a velocity distribution for exo activity at 50 pN is shown in Fig. 2.2.3. Each distribution was fit as a sum of Gaussian functions, one at zero velocity and the other at finite velocity. The fit at finite velocity represents the distribution of instantaneous velocities characteristic of enzyme catalytic activity, while the zero-velocity distribution represents instrument noise as well as protein fluctuations, and these data are not included in the velocity analysis. The average instantaneous velocities given by the mean of the Gaussian were obtained for each trajectory and averaged over at least three trajectories for all forces, and these averages for both exo and pol activity are shown in Fig. 2.2.4. The instantaneous velocities are sensitive to the template tension, suggesting that the rate-limiting step is force-dependent. In addition, the sharp transition from pol to exo activity as a function of force also suggests that switching between these functions in the absence of force is expected to be thermally driven and the initiation of an exo event is facilitated by force.
Force-dependent polymerization and exonucleolysis velocities

The instantaneous velocities of T7 and Klenow Fragment DNA polymerases were previously modeled, primarily attributing the force dependence to the activation enthalpy of converting n bases from the ss to the ds geometry. In that model, the projections of DNA segments along the direction of applied external force were determined “globally” by the change in the extension of ds and ss DNA extension as a function of force, averaged over thousands of bases. Therefore, that model does not distinguish the local orientations in the DNA segments inside the active site from the DNA segments further away from the active site. Consequently, Goel et al. suggested a “local” model attributing the force dependence of polymerization velocity to the orientation change of two DNA segments neighboring the active site.

Figure 2.2.3: Bimodal catalytic velocities of pol III core. Representative bimodal Gaussian distribution of pol III core (0.2 μM) instantaneous velocities at 50 pN template tension determined from the moving average filter (yellow) shown in the inset. Zero (dashed blue) and nonzero (dashed red)-peaked Gaussians represent the paused states and the moving states of pol III core, respectively. The peak position of the non-zero distribution represents the pol III core mean instantaneous velocity at 50 pN. Inset: Representative temporal trajectory (magnified) of the moving and paused states of pol III core. The pause detection (black, see Materials and Methods) is done using the moving average (yellow) of the extension data (blue).
It was also shown that large conformational changes in the pol-template complex and a conserved active-site geometry that induce a sharp kink at the 5’ end of the template during a catalytic pol event is a universal property shared by three families of polymerases\textsuperscript{120}.

**Figure 2.2.4:** Force dependence of the pol III core instantaneous velocities. Positive velocities represent polymerization and negative velocities represent exonucleolysis in the presence (red) and absence (blue) of dNTPs. The force dependence of polymerization is fitted to the RCLM\textsuperscript{109} (black line) as described in the text, yielding 84.8 ± 5.7 nt/s as the zero-force velocity ($v_{0,\text{pol}}$). The exo force dependence is modeled as a simple exponential function of force, which yields $v_{0,\text{exo}} = -20 \pm 5$ nt/s, $d = 0.11 \pm 0.02$ with dNTPs (negative red line) and $v_{0,\text{exo}} = -15 \pm 10$ nt/s, $d = 0.14 \pm 0.05$ nm without dNTPs (blue line), where $d$ is the force-independent length change required for each exonucleolysis event. Error bars are standard errors of at least three independent measurements and uncertainties in the fitting parameters are from the standard deviation of the $\chi^2$-minimized fit.

This model was further modified in Andricioaei et al.\textsuperscript{119} by restricting the orientations of these local DNA segments (Restricted-Cone Local Model, RCLM) due to steric effects that were determined using molecular dynamic simulations on the *Thermus aquaticus* (Taq) DNA polymerase I complex. In this model, the instantaneous polymerization velocity ($v(F)$) is given by\textsuperscript{121}

$$v(F) = v_0 e^{-\Delta G(F)/k_BT}$$

\textsuperscript{2.2.1}
where $\Delta G$ is the force-dependent free energy contribution determined by the additional enthalpy and change in entropy associated with converting the two DNA segments from their “open” to “closed” forms, in the presence of force. For a given DNA segment $d$, the free energy contribution $\Delta G_d$ is given by

$$
\Delta G_d(F) = -\int_0^F \langle \cos \theta \rangle dF'
$$

Here $\langle \cos \theta \rangle$ is the average angular orientation of a given DNA segment along the direction of force and is given by

$$
\langle \cos \theta \rangle = \frac{\cos \theta_m e^{\xi \cos \theta_m} - \cos \theta_M e^{\xi \cos \theta_M}}{e^{\xi \cos \theta_m} - e^{\xi \cos \theta_M}} - \frac{1}{\xi}
$$

The subscripts $m$ and $M$ refer to the minimum and maximum values used in RCLM, $\delta$ is the length of a given DNA segment, and $\xi = F \delta / k_B T$. Here $v_0$ is the zero-force instantaneous velocity, $k_B$ is the Boltzmann constant, and $T$ is the absolute temperature. The reported $\theta$ and $\delta$ values for these two segments in the closed and open conformation in Andricioaei et al. were used in the force ranges less than and greater than 9 pN. Although this model used Taq pol I, which is an A family polymerase, and $E. coli$ pol III $\alpha$ is a C family polymerase, another C family polymerase Taq pol III $\alpha$ was also shown to exhibit a bend of the template at the active site, suggesting that the basic features of this model are applicable in the present case.

The best fit of our data to the model of Andricioaei et al. yields $v_{0,\text{pol}} = 84.8 \pm 5.7$ nt/s (Fig. 2.2.4). Pol III core has a high velocity in the context of the complete replisome; however, the rate of replication of pol III core alone, which is the relevant comparison here, was measured in bulk biochemical experiments to be 20 nt/s. This significant difference in the velocities with and without the $\beta$ clamp is likely due to the weaker association of pol III core with the substrate.
DNA in the absence of the β clamp. In contrast to single molecule assays, in bulk biochemical assays the catalytic rates are averaged over the paused states as well. Hence it is not surprising that the zero-force velocity \( v_{0,\text{pol}} = 84.8 \pm 5.7 \text{ nt/s} \) in our measurements is higher than the 20 nt/s rate observed in bulk biochemical experiments.

In the case of exo activity, there is not a similar previously applied model, so we will initially assume a simple exponential dependence on force given by

\[
v(F) = v_0 e^{-Fd/k_B T}
\]  

(2.2.4)

Here \( v_0 \) is the zero force exo velocity and \( d \) is the force-independent length change required for each exonucleolysis event. The best fit yields, \( v_0 = -20 \pm 5 \text{ nt/s} \), \( d = 0.11 \pm 0.02 \text{ nm} \) with dNTPs and \( v_0 = -15 \pm 10 \text{ nt/s} \), \( d = 0.14 \pm 0.05 \text{ nm} \) in the absence of dNTPs. The value of \( d \) reflects an elongation of DNA that occurs during each exo event, which is likely the slightly extended state of the terminal base pair when it is positioned for cleavage during the exo rate-limiting step. The fact that \( d \) (in both cases) is slightly less than the total change in DNA length (~0.22 nm/bp) during an exo event supports this hypothesis.

Other alternative models, such as the model that describes \( \varphi 29 \) pol and exo activity\textsuperscript{123}, are not consistent with our data. For \( \varphi 29 \), the total velocity was presented as a sum of exo and pol activity, such that fluctuations between the states determined the total velocity. However, we do not observe a significant change in exo velocity in the presence and absence of dNTP, showing that pol activity is not present at forces that primarily induce exo activity. Also, as is shown below, in contrast to \( \varphi 29 \), the time that pol III core spends on the DNA is much shorter than the switching time between pol and exo activities, such that in each processive event only one process is observed. Because pol and exo activity are accomplished by two separate proteins for
pol III core, and ε by itself is an independent ssDNA nuclease, it is not surprising that these activities are so well separated.

**Concentration-dependent pause times during exonucleolysis**

The time spent during an exo event is measured as the dwell time ($\tau_d$). The time spent between two consecutive exo events is measured as the pause time, determined from the pause-detecting trajectory shown in the inset of Fig. 2.2.3. The reciprocal of the average $\tau_p$ is the exo initiation rate ($k_{\text{init}}$) at a given force and concentration. The resulting concentration dependence of the pause and dwell times is given in Fig. 2.2.5A. Although $k_{\text{init}}$ (Fig. 2.2.5B) initially increases with concentration, this rate saturates at high concentrations. To describe this concentration dependence, we propose the following kinetic scheme.

---

**Figure 2.2.5: Concentration-dependent exonucleolysis of pol III core** (A) Concentration dependence of pause (blue) and dwell (green) times during exo. Average pause time is significantly decreased with concentration. Dwell times do not exhibit a significant dependence on concentration. (B) Exo initiation rates, $k_{\text{init}}$ (C) as a function of pol III core concentration. The rate $k_{\text{init}}(C)$ is the reciprocal of the pause times shown in (A) and fitted to the two-state model as described in the text (Eq. 2.2.6). The best fit yields the dissociation constant $K_d = 0.13 \pm 0.07 \mu M$ and transition rate to exo-active state $k_2 = 11.3 \pm 3.1 \text{ s}^{-1}$. Error bars are standard errors of at least three independent measurements and uncertainties in the fitting parameters are from the standard deviation of the $\chi^2$-minimized fit.
Here E is pol III core in the solution, C is the pol III core concentration, and DNA$_n$ is the substrate that is the primer terminus at the primer-template junction. The first step is the bimolecular pol III core binding, which is in pre-equilibrium to the subsequent exo-active E•DNA$^*$ state. The quantity $k_2$ is the rate of pol III core transition to its exo-active state, and $k_{-2}$ is the rate of exo activity termination, due to dissociation that is determined by the dwell time measured below. As suggested by our measured force-dependence of that rate (also discussed below) as well as by its strong temperature dependence measured in the previous studies,$^{111,112}$ exo initiation is rate-limited by the requirement for destabilization of 2-3 bp at the primer terminus. Based on the pol III core structure,$^{112}$ such destabilization is most likely followed by strand transfer of the 3’ end of the primer from the pol to exo catalytic site. This step is not accompanied by a net change in DNA construct length and is therefore not directly observed in our experiment. However, conventional biochemical measurements of the exo initiation rates$^{111,112,124}$ also suggest that DNA duplex destabilization is the rate-limiting step of the process, supporting the notion that the subsequent strand transfer between the pol and exo sites occurs rapidly.

The next step is the catalytic exo activity that transforms E•DNA$^*_n$ to E•DNA$^*_{n-1}$. This process ($k_{\text{exo}}$) is fast, and is not observed in the kinetics of exo-initiation. This is because our analysis decouples the paused states from the moving states that are measured as a velocity ($v_{\text{exo}}$), which ranges from 20 to 100 nt/s, depending on the force, as described in the previous section. For the proposed reaction scheme (Eq. 2.2.5), the predicted exo initiation rate ($k_{\text{init}}$) is given by,

$$k_{\text{init}}(C) = \frac{k_1 C}{k_1 C + k_{-1}} k_2 = \frac{1}{1 + K_d/C} k_2 \quad \text{(2.2.6)}$$
where,

$$K_d = \frac{k_{-1}}{k_1}$$

(2.2.7)

is the equilibrium dissociation constant for pol III core binding to the substrate. Note that because $k_{\text{init}}$ is the reciprocal of the average pause time ($1/\tau_p$) for a single pol III core molecule to rebind to the substrate, this explicitly represents the on rate of an exo-competent state and therefore, $k_2$ is disregarded in the equation. The best fit to the observed dependence of $k_{\text{init}}$ on concentration (Fig. 2.2.5B), yields $K_d$ to be $0.13 \pm 0.07 \mu$M and $k_2$ to be $11.3 \pm 3.1 \text{ s}^{-1}$.

Eq. 2.2.6 assumes that the first step of pol III core binding to DNA occurs in pre-equilibrium to the slower catalysis initiation step $k_2$, $(k_{-1} \gg k_2)$. According to our fitted values of $k_2$ ($11.3 \pm 3.1 \text{ s}^{-1}$) and $K_d$ ($0.13 \pm 0.07 \mu$M), the bimolecular association rate $k_1 (k_{-1}/K_d)$ is much higher than $k_2$ ($k_1 > 108 \text{ M}^{-1} \text{ s}^{-1}$), which is on the order of the diffusion rate ($109 \text{ M}^{-1} \text{ s}^{-1}$). Thus, initial pol III core binding to the primer-template junction is a nonspecific diffusion-limited process, leading to the slower step of catalysis initiation, which subsequently results in either pol or exo activity, depending on the stability of the primer-template junction.

**Force-dependent pause times during exonucleolysis**

We observe a significant increase in the exo initiation rate $k_{\text{init}}$ with increasing template tension (Fig. 2.2.6). The data in Fig. 2.2.6 is measured at 0.2 μM pol III core concentration and therefore reflect the protein-saturated value of $k_{\text{init}}$ that is $\sim k_2$. Thus, the observed force-dependence of $k_{\text{init}}$ primarily corresponds to the rate at which the bound protein-DNA complex transforms into the exo-active state. It has been shown that destabilization of the primer-template junction increases the susceptibility of DNA to the exonuclease activity of pol III core and its exo subunit.
Because template tension uniformly destabilizes all base pairs of stretched DNA\textsuperscript{111}, the force dependence of $k_{\text{init}}$ is likely due to the increased probability of the duplex fraying, which can be described as\textsuperscript{125}

$$k_{\text{init}} = k_F e^{n_{\text{exo,init}} (\Delta G^r(F) - \Delta G_0^r)/k_BT}$$  \hspace{1cm} (2.2.8)

Where,

$$\Delta G^r(F) = \int_{0}^{F} x_{ss}(F')dF' - x_{ds}(F')dF'$$  \hspace{1cm} (2.2.9)

Here $x_{ss}$ and $x_{ds}$ are the ss- and dsDNA extensions, respectively (Fig. 2.2.1), and $k_F$ is the maximum exo rate on completely destabilized dsDNA or on ssDNA. This maximum exo rate is expected to be reached at the melting force $F_m$ (62 pN), at which the work performed by force to destabilize dsDNA, $\Delta G^r(F)$, is equal to the free energy of bp melting in the absence of force $\Delta G_0^r$, $2.23 k_BT$\textsuperscript{125}. Best fit yields $n_{\text{exo,init}}$ to be $1.84 \pm 0.20$ bp and $k_F$ to be $17.3 \pm 1.3$ s$^{-1}$ µM$^{-1}$.

This value for $n_{\text{exo,init}}$ is very similar to the number of mismatches required for optimal exo activity by the subunit $\varepsilon$ observed in our bulk primer extension assay for pol III core, presented in Fig. 2.2.6, and previous biochemical assays using isolated $\varepsilon$\textsuperscript{126}.

### Pauses during polymerization

The observed pauses between consecutive catalytic pol bursts appear to be unaffected by the applied force within the accuracy of our measurement (Fig. 2.2.7). The weighted average of the pause time over all forces during pol activity is $0.90 \pm 0.22$ s. This suggests that in contrast to that observed for exo, initiation of a pol event is force-independent.

### Dwell times for polymerization and exonucleolysis
The observed catalytic bursts during both pol (Fig. 2.2.7) and exo (Figs. 5-6) occur at short time scales (\( \tau_{d,\text{pol}} = 0.21 \pm 0.05 \) s, \( \tau_{d,\text{exo}} = 0.15 \pm 0.05 \) s) that are independent of applied force or protein concentration (data not shown for pol).

![Figure 2.2.6: Force-dependent exonucleolysis of pol III core](image)

(A) Force-dependence of average pause times (blue) and dwell times (green) during exo at 0.2 \( \mu \text{M} \) pol III core. Average pause times decrease with force. Average dwell times (green) exhibit insignificant force dependence. (B) Exo initiation rates, \( k_{\text{init}}(F) \) as a function of force. The rate \( k_{\text{init}}(F) \) is the reciprocal of the average pause times shown in (A) and is fitted to an exponential function of force as described in the text (Eq. 2.2.8), which yields \( n_{\text{exo,init}} \) to be 1.84 \( \pm \) 0.20 bp and \( k_r \) to be 17.3 \( \pm \) 1.3 s\(^{-1}\). Here \( n_{\text{exo,init}} \) is the minimum number of base pairs required to melt at the primer-template junction in order for the bound-pol III core to transform to the exo-competent conformation. The quantity \( k_r \) is the force-independent attempt rate at the given concentration. Error bars are standard errors of at least three independent measurements and uncertainties in the fitting parameters are from the standard deviation of the \( \chi^2 \)-minimized fit. (C) Exonuclease activity of DNA pol III core complex is stimulated by base-pair mismatches at the primer-template junction. Fully-extended polymerization products also decrease when 1, 2, or 3 non-complementary bases are present at the junction. Reactions were quenched after 0, 1, 2.5, 5, and 10 min and analyzed by 12\% denaturing polyacrylamide gel electrophoresis.
The reciprocal of the measured dwell time represents the termination rate of pol or exo activity ($k_{\text{off}}=1/\tau_d = 5-7 \text{ s}^{-1}$) that is orders of magnitude higher than that observed for the pol III core complex in the presence of the β clamp$^{85,112,127}$. This significant difference in termination rates may account for the much weaker association of pol III core with the primer-template junction in the absence of the β clamp. Because the observed $k_{\text{init}}$ (C) (Fig. 2.2.5) for exo increases and saturates at high concentrations, this termination of catalysis primarily represents protein dissociation from the active conformation rather than intrinsic pausing during catalysis. Overall, this suggests that each exo event is associated with bimolecular protein binding to the primer-template junction, which is followed by a complex transition to its exo-active conformation.

**Number of processively catalyzed base pairs during polymerization or exonucleolysis**

Fig. 2.2.7B shows the number of nucleotides polymerized or excised per pol or exo activity burst, obtained as a product of the velocity (Fig. 2.2.4) and dwell time (Figs. 2.2.7A). The number of catalyzed nucleotides vanishes at ~25 pN for both exo and pol, indicating a transition from pol to exo. Moreover, from the predicted zero-force pol velocity (Fig. 2.2.4) and the measured dwell time (Fig. 2.2.7), we can estimate the zero-force value of the processivity to be ~18 nt. This is consistent with previously measured bulk studies, in which pol III core was shown to incorporate 10-15 nucleotides before dissociating from the primer-template junction$^{91}$.

**Discussion**

The force dependence of the catalytic functions of polymerases from bacteriophage φ29 and T7 as well as the *E. coli* polymerase Klenow Fragment (KF) has been previously studied using single-molecule stretching experiments$^{84,105,128}$. Our single molecule results for the force dependence of pol III core catalytic activity qualitatively agree with the previously studied
polymerases to the extent that force inhibits or facilitates polymerization and exonuclease activity, depending on the force relative to the 6 pN crossover point. However, pol III core is a weakly processive polymerase that only incorporates ~20 nucleotides before dissociating from the primer-template junction\textsuperscript{106}.

This weak processivity of pol III core imposes an additional challenge in studying its function and characterizing its activity, especially at the single-molecule level, demanding high resolution data acquisition. In this study we have successfully probed the force dependence of both the polymerase and exonucleolysis functions of pol III core, and by modeling the force dependence

**Figure 2.2.7:** Force-dependent pauses and catalytic bursts (A) Force-dependence of average pause times (red) and dwell times (blue) during pol at 0.2 μM pol III core. Average pause and dwell times are independent of force during polymerization. The weighted average over all the forces of dwell time (dashed blue) is 0.20 ± 0.05 s and pause time (dashed red) is 0.90 ± 0.22 s shown by the dashed line. Error bars are standard errors of at least three independent measurements. (B) Average number of nucleotides (\(N\)) polymerized (red) or excised (blue) per single burst of catalytic activity of pol III core as a function of force. \(N\) was calculated as a product of measured velocities and dwell times of pol and exo activities, respectively. The red triangle is the \(N\) determined from the zero-force velocity predicted from the RCLM (Fig. 2.2.4) and the red square is the zero-force measurement reported in a previous biochemical study\textsuperscript{121}.
of the observed pol and exo velocity and pausing, we obtain significant new insights into how these processes are regulated.

One major difference between pol III core and the previously studied T7 and φ29 polymerases is that pol III core is a multienzyme assembly in which the pol and exo domains are different subunits, α and ε, which can function independently even when not part of core 128. In fact, a recent study estimated that the distance between the polymerase and exonuclease active sites in pol III core is greater than 7 nm 112. Therefore, for the exo activity to initiate after pol III binding to duplex DNA, the 3’ end of the primer strand has to move from the pol catalytic site in α into the exo catalytic site in ε. The length change associated with the force dependence of the exo initiation rate \( k_{\text{init}}(F) \) is only ~0.44 nm (2 bp × ~0.22 nm/bp, where ~0.22 nm/bp is the extension

---

**Figure 2.2.8:** Temperature- and force-dependent exonucleolysis initiation (A) Temperature-dependence of the exo initiation rate as reported in Brenowitz et al 111,128. Red circles are the reported exo initiation rates on a G-C paired primer-template junction, scaled by a factor of 0.02 to match the experimental conditions in this study (see discussion). The reported data is fit to an Arrhenius function as described in Eq. 2.2.2.11. The best fit (red line) yields \( n\Delta H_0 = 47 k_BT \). The dashed blue line at 8.1 s\(^{-1}\) is the maximum rate reported, which is observed on ssDNA at 310 K. (B) The rate \( k_{\text{init}} \) determined from force- (blue) and temperature- (red) 111 dependent measurements as a function of total free energy \( \Delta G^\dagger(F,T) \) required to destabilize terminal dsDNA. The global fit is from the expression \( k_{\text{init}} \exp\left( -n_{\text{exo,init}} \Delta G^\dagger /k_BT \right) \left( 1 + \exp\left( -n_{\text{exo,init}} \Delta G^\dagger /k_BT \right) \right) \), that describes the probability of destabilizing \( n_{\text{exo,init}} \) terminal base pairs as a function of \( \Delta G^\dagger(F,T) \). Here \( \Delta G^\dagger(T) = \Delta H_0 - T\Delta S_0 \), where \( \Delta H_0 = 17 k_BT \) and \( T\Delta S_0 = 0.0471 T k_BT \), as described in the text. \( \Delta G^\dagger(F) \) is given by Eq. 2.2.8. The best fit yields \( n_{\text{exo,init}} = 2.16 \pm 0.13 \) bp and \( k_0 = 27.9 \pm 1.9 \) s\(^{-1}\).
change associated with melting one DNA bp at F >30 pN, Fig. 2.3.1B). Because this length change is much smaller than the distance between the active sites of α and ε, α binding to the template must be disrupted to allow ε to bind to the frayed strand.

The structural autonomy of the two catalytic domains of pol III core may result in more independent functions between the two proteins. In the case of T7 and φ29 DNA polymerases, duplex DNA binds to the polymerase active site and exonucleolysis is facilitated via an intramolecular transfer through several intermediate steps\textsuperscript{124,129,130}. Although the applied force favors exo and suppresses the pol activity of φ29, the underlying mechanochemistry is significantly different from pol III core. Specifically, in contrast to pol III core, the dwell time of φ29 is much longer than its time of switching between pol and exo activities. The average velocity of φ29 catalysis appears to be a continuous function of the force with several fine features suggesting intermediate steps in the pol to exo switching process\textsuperscript{105}. In contrast, ε by itself has been shown to be an ssDNA exonuclease, and the catalytic activity of ε is similar on ssDNA and mispaired primer termini. The ε subunit preferentially binds ssDNA, whereas α binds both ssDNA and dsDNA and prefers a primer-template junction\textsuperscript{105,107}. In addition, a recent NMR study showed that a primer destabilized due to mismatches increases the propensity of the mismatch to reach the ε subunit, enabling ε to correct for the mismatches in a passive manner\textsuperscript{131}. Thus, a simple model for the regulation of exo and pol activity is based primarily on the preferential binding of each protein for specific DNA substrates. Because α binds strongly to a stable primer-template junction, while ε binds strongly to ssDNA free ends, the switch between pol and exo is determined by the stability of the primer-template junction, which is reduced upon the application of large forces or at high temperatures. These conditions, as well as mismatches
at the primer-template terminus, induce a shift from pol to exo activity, and these activities will be considered independently below.

**Force-dependent instantaneous velocity of pol III core exonucleolysis**

We have modeled the force-dependence of the pol and exo instantaneous velocities of pol III core as independent processes. Here, for the first time we were able to directly measure the instantaneous exo velocity on a properly paired dsDNA substrate uniformly destabilized by a stretching force parallel to the DNA axis. Interestingly, this velocity ranges between 40 and ~110 nt/s as the stretching force increases from 30 to 55 pN (Fig. 2.2.4). Because at 55 pN the dsDNA is very close to its melting force of 62 pN, we approximate the latter measurement as the maximum catalytic exo rate of pol III core. The observed velocities at higher forces (~100 nt/s) are independent of the pol III core concentration (data not shown) and are also not strongly facilitated by the force. The exponential dependence of the observed exo velocities yields $v_{0,exo} = 20 \pm 5$ nt/s and $d = 0.11 \pm 0.02$ nm with dNTPs and $v_{0,exo} = 15 \pm 10$ nt/s and $d = 0.14 \pm 0.04$ nm without dNTPs. The independence of both parameters within uncertainty to the presence of dNTPs suggests that $\varepsilon$ acts independently from the polymerase $\alpha$, even in the context of pol III core. The length change required during a rate-limiting step of processive catalytic excision is significantly smaller than that of the length change required for exo initiation (2 bp ~ 0.44 nm, Fig. 2.2.6B), which is the rate-limiting step for the exo process. Thus, while exo initiation is relatively slow and strongly dependent on the primer-template terminus stability, the instantaneous exo velocity is about ~10-100 fold faster and depends weakly on the base pair stability. This result indicates that the catalytic excision by itself is likely not strongly affected by the presence of mismatches or the DNA sequence at the junction.
Force-dependent instantaneous velocity and pausing of pol III core polymerization

The instantaneous pol velocity is strongly affected by the applied force (Fig. 2.2.4). However, the dwell times during pol are approximately force-independent (Fig. 2.2.7A). The zero-force pol velocity ($v_{0,\text{pol}}$) was found to be $84.8 \pm 5.7$ nt/s based on the model of Eqs. 1 and 2. Therefore, the predicted number of nucleotides synthesized by pol III core at zero force during an average dwell time of 0.21 s is ~18 nt (Fig. 2.2.7B). Pauses during consecutive pol bursts are relatively longer than in exo and are poorly or not at all affected by the applied force. We find the average pause over all forces to be $0.90 \pm 0.22$ s at 0.2 μM pol III core. Thus, in strong contrast to the exo activity, the instantaneous catalytic velocity of the pol activity is strongly affected by force, while the initiation of a pol event is insensitive to the applied force. This result is consistent with the fact that pol binds strongly to a stable primer-template junction, the presence of which does not depend strongly on force at pol-competent forces. In contrast, exo requires a highly force-dependent destabilized primer-template junction for initiation, as discussed below. However, once initiation of exo occurs, exo activity can proceed without further requirements for base pair destabilization.

Two-step exonucleolysis initiation from concentration-dependent measurements

We model the observed concentration dependence of exo-initiation rates using the proposed reaction scheme shown in Eq. 2.2.6. The pol III core-saturated value for $k_{\text{init}}=k_2$ ($11.3 \pm 3.1$ s$^{-1}$ at 55 pN), the rate at which an exo-active state is achieved, can be compared to the rate of a single exo-cut by pol III core measured previously using bulk biochemical assays$^{132}$. After correcting the results in Brenowitz et al$^{111}$ (as discussed below) to match our experimental conditions, we find that $k_{\text{init}}$ for ssDNA at 310 K is $\sim8.1$ s$^{-1}$ (dashed line, Fig. 2.2.8A). This represents the
maximum value that characterizes $k_{\text{init}}$ for a completely destabilized DNA substrate, obtained by increasing force or temperature. The similarity of this result to our single molecule measurements at high forces (55 pN) approaching the DNA melting force (62 pN) brings confidence that we are measuring the same process (Fig. 2.2.6). Furthermore, we measured $K_d$ to be $0.13 \pm 0.07 \mu\text{M}$, which agrees with the previously measured values of $0.14$-$0.46 \mu\text{M}^{111}$ for comparable solution conditions.

**Exonucleolysis initiation is determined by primer-template junction stability**

We model the force dependence of the exo-initiation rate as a function of the complete work done by force to transform DNA from its double- to single-stranded form as described in Eqs. 2.2.8-9$^{107}$. We find that at least two base pairs ($n_{\text{exo,init}}=1.84 \pm 0.20$ bp) are required to be destabilized through thermodynamic fluctuations to enter the exo-active state of pol III core. As shown in our primer-extension assay with pol III core (Fig. 2.2.6C) and as demonstrated by Miller et al$^{125}$ for isolated ε, there is a significant increase in exo activity when the number of mismatches at the primer-template terminus is altered from one to two, in remarkable agreement with our single-molecule results.

At concentrations above the pol III core-DNA binding $K_d$, the force dependence measured is determined primarily by $k_2$. Thus, simple diffusion-limited bimolecular binding of the enzyme to its DNA substrate occurs in pre-equilibrium to exo initiation. The exo initiation rate is strongly affected by dsDNA stability, leading to a strong force and temperature dependence of $k_{\text{init}}$. If pol III core exo activity is primarily rate-limited by the destabilization of the primer-template terminus, our force-dependent measurements should be comparable to the previously measured temperature dependence of its activity$^{112}$. The force-dependence of exo initiation $k_{\text{init}}(F)$ in our
study is analogous to the rate of single nucleotide excision on matched GC-terminated and single-stranded DNA reported in Brenowitz et al\textsuperscript{111}. To compare these values to those measured here, we scale the reported parameters of the Michaelis-Menten excision reaction ($V_{\text{max}}/K_m$) to determine $k_{\text{init}}(T)$, where,

$$k_{\text{init}}(T) = \left(\frac{V_{\text{max}}}{K_m}\right) \times \left(\frac{K_M}{C}\right)$$  \hspace{1cm} (2.2.10)

Here we took into account that the $K_M$ is $\sim 400$ nM as measured in the same work at 2 nM pol III core concentration (C), and shown in Fig. 2.2.8 A\textsuperscript{111}. Furthermore, we modeled $k_{\text{init}}(T)$ as an Arrhenius function analogous to the Eq. 2.2.8 where,

$$k_{\text{init}}(T) = k_T e^{-\frac{n_{\text{exo,init}} \Delta G^*(T)}{k_B T}} = k_T e^{-\frac{n_{\text{exo,init}} (\Delta H_0 - T \Delta S_0)}{k_B T}}$$  \hspace{1cm} (2.2.11)

Here $\Delta G^*(T)$, $\Delta H_0$ and $\Delta S_0$ are the free energy, enthalpy and entropy, respectively, of a single bp melting at a reference temperature $T$, and $n_{\text{exo,init}}$ is the number of base pairs melted during the rate-limiting step of exo-initiation. The best fit yields $n_{\text{exo,init}} \Delta H_0$ to be $47$ $k_B T$. Because the enthalpy of a single base pair melting, $\Delta H_0$, is $\sim 17$ $k_B T^{111}$, $n_{\text{exo,init}}$, the minimum number of destabilized base pairs required in the rate-limiting step for exo, can be estimated to be $\sim 2.5$ bp.

Furthermore, a global fit (Fig. 2.2.8B) to both the force and temperature dependence as a function of the free energy required to destabilize dsDNA, determined using the same values, yields $n_{\text{init,exo}} = 2.16 \pm 0.13$ bp. This agrees remarkably well with the value determined for $n_{\text{exo,init}}$ (1.8 $\pm 0.2$ nt) obtained only from our force-dependent measurements. The excellent compatibility between the force and temperature dependence of the pol III core exo activity confirms that the rate limiting step $k_2$ is primarily dependent on the stability of the primer-template terminus.

Furthermore, the exo catalytic functions of pol III core and $\epsilon$ are shown to be similar once scaled
with their appropriate $K_d$ values. This suggests that the requirement for destabilization at the primer-template terminus for the onset of exo activity by either $\epsilon$ or pol III core are very similar. However, pol III core alters the geometry of bound $\epsilon$, in which the 3’ end of the primer strand is required to be displaced a significant distance to the catalytic site of $\epsilon$ to trigger the onset of exo activity.

**Materials and Methods**

**Pol III core expression, purification, and biochemical analysis**

Wild-type DNA pol III core was expressed from the plasmid pET16b-dnaE-holE$^H$-dnaQ, which features a His-tag on the $\theta$ subunit (a generous gift from Mark Sutton, Univ. at Buffalo), as described. Core was purified from a cell pellet harvested from 1 L of culture and stored at -80 °C. The cells were thawed on ice and lysed by sonication. Clarification was carried out by centrifugation at 12000 x g for 1 h at 4 °C. The supernatant containing soluble proteins was passed through a 0.45 μm filter before loading on a 5-mL His-Trap HP column (GE Healthcare) equilibrated with buffer HisA [20 mM HEPES; 500 mM NaCl; 50 mM imidazole; 10% glycerol, pH 7.5]. Bound protein complex was eluted using buffer HisB [20 mM HEPES; 500 mM NaCl; 300 mM imidazole; 10% glycerol, pH 7.5]. Fractions containing the desired protein were pooled and diluted 10-fold using buffer HeparinA [50 mM HEPES (pH 7.5); 0.1 mM EDTA; 10% glycerol; 1 mM DTT] before loading onto a 5-mL Hi-Trap Heparin HP column (GE Healthcare) equilibrated with HeparinA. Bound proteins were eluted by addition of buffer HeparinA + 1 M NaCl in a linear gradient. Fractions containing intact core complex were pooled and diluted 6-fold with buffer HydroxyA [50 mM HEPES (pH 7.5); 150 mM NaCl; 1 mM DTT; 10% glycerol] and loaded onto a 5-mL hydroxyapatite column (BioRad Bioscale Mini CHT Type 1, 5 ml, 40 mm cartridge) equilibrated with buffer HydroxyA. Bound protein was eluted with buffer.
HydroxyB [200 mM sodium phosphate (pH 6.5); 150 mM NaCl; 1 mM DTT; 10% glycerol] in a step gradient. Fractions containing protein complex were pooled and dialyzed overnight at 4 °C against 2 L of storage buffer [30 mM HEPES (pH 7.5); 100 mM NaCl; 0.5 mM EDTA; 2 mM DTT; 20% glycerol]. Protein purity was determined by SDS-PAGE, proteins were quantified by Bradford assay, and purified complex was stored at -80 °C.

Primer extension assays were carried out as described previously\textsuperscript{134} using \textsuperscript{32}P-labeled primers annealed to 61-mer template. Reactions contained a final concentration of 25 nM DNA polymerase, 100 nM primer/template DNA, 100 µM dNTPs, 7.5 mM MgSO\textsubscript{4}, 30 mM HEPES (pH 7.5), 20 mM NaCl, 2 mM DTT, 1% (w/v) bovine serum albumin, and 4% glycerol. Reaction products were separated by denaturing 16% polyacrylamide gel electrophoresis and analyzed by phosphorimaging. The template sequence is 5’-ggttactcagtcgaagacctggcgtgcagttactatcatgc; the primer sequences are

- **Match:** 5’-gcatatgatagtacagctgcagccggacgcc;
- **MmT:** 5’-gcatatgatagtacagctgcagccggacgcct;
- **MmTC:** 5’-gcatatgatagtacagctgcagccggacgcctc;
- **MmTT:** 5’-gcatatgatagtacagctgcagccggacgcctt;
- **MmTTT:** 5’-gcatatgatagtacagctgcagccggacgccttt.

**Single molecule DNA constructs**

Either a 38.5-kbp λ DNA or an 8.1 kbp pBacgus11 DNA were used in the single molecule stretching experiments. The 48.5 kbp linear λ DNA (Roche) with 12-nt 5’ overhangs at both the termini was digested with ApaI (New England Biolabs, NEB). The 5’ overhang of the resultant 38.5 kbp substrate was filled-in with KF in the presence of dGTP, dATP, biotin-14-dATP, and biotin-14-dCTP (NEB). At the opposite end a biotinylated oligonucleotide (5’-bbCTCbTCTCbTCTTCTCTTCTCTTCTTGGCC-3’, Integrated DNA Technologies, IDT) consisting of a 3’ end complementary sequence to the ApaI-digested site was ligated with T4
DNA ligase. The 8.1 kbp construct was created by first linearizing the pBacgus11 (a gift from Borja Ibarra) dsDNA vector (8041 bp) with BamHI and SacI (NEB). A digoxigenin (DIG) labeled dsDNA handle with a complementary sticky end to the BamHI sequence was generated as described\(^{135}\). A biotinylated oligonucleotide (5’-bbCTCbbTCCTCTCTCTCTCTCTCTCTCTCTCTGTCCAGCT-3’, IDT) with a 3’ end complementary to SacI sequence, and the dsDNA DIG-handles were then ligated to their complementary positions at the linearized pBacgus11 DNA using T4 DNA ligase. In both of the biotinylated oligos, the position of biotin is indicated by b.

**Single molecule optical tweezers experiments**

We used optical tweezers to induce tension in single DNA molecules and thereby facilitate the pol III core activity. Here, a single DNA molecule was attached by its labeled ends to derivatized polystyrene spheres. The 38.5 kbp \(\lambda\) DNA biotinylated at both the termini or 8.1 kbp pBACgus11 DNA, with biotinylated and DIG handles ligated at its respective termini, were tethered at the ends with streptavidin or streptavidin and anti-digoxigenin coated beads. One bead was immobilized by a glass micropipette attached to a flow cell while the other was held in a dual beam optical trap. By moving the glass micropipette attached to the flow cell, the DNA molecule was stretched, and the force required to extend the DNA molecule was measured. The solution surrounding a single DNA molecule was replaced with pol III core (20, 50, 100, 200 and 500 nM) diluted in the reaction buffer, 50 mM Hepes at pH 7.5, 25 mM Na\(^+\), 10 mM MgCl\(_2\), 5 mM DTT and 1% BSA. In addition, 0.3 mM (each) dNTP were added to the experiments with dNTP at 0.2 \(\mu\)M pol III core. For some experiments, the exo activity of T7 DNA polymerase (NEB) was used initially to create a partial ssDNA substrate, then exchanged for pol III core. Data were collected at constant forces at 25 Hz, in which a detected change in the tension of the
DNA substrate is compensated with a change in extension via a force feedback loop. The conversion between dsDNA-ssDNA upon exonucleolysis or polymerization, at constant DNA tensions, is registered as a change in extension as a function of time.

**Single molecule data analysis**

The extension-time trajectory was filtered with a moving average window of 8 Hz. The change in extension was converted to number of replicated or excised nucleotides by dividing the observed distance change by the expected change in extension at a given force accompanying the event of a single nucleotide incorporation. Theoretical polymer models, extensible worm like chain \(^{105}\) for dsDNA, and extensible freely jointed chain\(^{109}\) for ssDNA, were used to calculate the expected change in extension at a given force. Polymerization or exonucleolysis rate distributions were obtained from the moving average trace and fit to a bimodal Gaussian to find the instantaneous catalytic rate. Pauses were captured in the filtered trace by setting the consecutive events less than a cutoff of \(5/\sqrt{w}\) nt, to their mean values using a custom MATLAB code. Here 5 nt is the experimental noise and \(w\) is the number of data points included within the chosen window. We tested our algorithm with simulated data and the accuracy of recovered results were >90%, when the random noise level was set as 5 nt.
2.3 EcSSB-ssDNA binding dynamics regulated through competitive binding

The analysis described in this section is a work in progress and is currently unpublished.

Introduction

Single-stranded DNA binding proteins (SSBs) are proteins that rapidly sequester and protect the single-stranded DNA (ssDNA) stretches formed during genome maintenance\(^8\text{-}^{18,109}\). They exhibit high affinity ssDNA binding and may also play regulatory roles by interacting with other proteins involved in genome maintenance\(^{19,136}\). The SSB from \textit{E. coli} (EcSSB) is one of the most extensively studied SSB proteins, and it plays critical roles during DNA replication. EcSSB protects the transiently formed ssDNA intermediates from nucleolytic degradation and facilitates the lagging strand synthesis by eliminating secondary structures in the template strand. It interacts with the replisome by binding to the χ subunit, which in turn is bound to the clamp loader complex (Fig. 2.1.1).

Monomeric EcSSB is 19 kDa, contains an oligonucleotide/oligosaccharide (OB) binding motif\(^{26}\), forms a stable homotetramer in solution, and exhibits multiple ssDNA binding conformations\(^{18,23,26,137}\). The distinct binding modes of EcSSB, \((SSB)\)_\text{n}, are identified based on the number of nucleotides (n) occluded by the tetramer upon binding to ssDNA. These modes can interconvert depending on the solution conditions such as the salt type, salt concentration and protein to DNA ratio. Studies on EcSSB binding to poly-dT substrates uncovered the \((SSB)_{35}\), \((SSB)_{56}\), \((SSB)_{65}\) modes\(^{137}\). At saturating protein concentrations and low monovalent salt conditions EcSSB binds ssDNA predominantly in the \((SSB)_{35}\) mode with high cooperativity. In contrast, at high salt conditions (NaCl \(>0.2\) M, 2 mM Mg), the relatively low cooperative \((SSB)_{65}\) mode is favored\(^{14}\). X-ray crystallographic structural studies revealed a model for the
(SSB)$_{65}$ binding topology in which the ssDNA is ‘fully’ wrapped to associate all four SSB subunits, resembling the seams of a tennis ball. However, the precise binding topologies of the other binding modes are not completely understood.

Recent single molecule analyses have greatly enhanced the understanding of the binding dynamics of these distinct modes. Stopped flow and single molecule FRET (smFRET) experiments were used to study the spontaneous transitions between the (SSB)$_{65}$ and (SSB)$_{35}$ binding modes to characterize the dynamic equilibrium between these well-defined functional and structural states. smFRET and optical tweezers studies have demonstrated the ability of the tetramer to diffuse along the DNA substrate without dissociation, presumably by a reptation mechanism. Using force spectroscopy with smFRET analysis, Suksombat et al. identified an intermediate (SSB)$_{17}$ mode that existed at higher template (> 6 pN) tension.

Here, using optical tweezers we directly measure the binding of EcSSB to a long ssDNA substrate that allows us to precisely probe the interconversion dynamics of distinct competitive binding modes. For the first time we show that at saturating concentrations (>1 nM), SSB binds to ssDNA in a biphasic manner, and equilibrates in a conformation that negligibly wraps the substrate DNA. Using numerical simulations, we show that our results are consistent with a model in which the conformation of the SSB-ssDNA complex is regulated through competitive binding, and we directly measure the transition rates between these different modes. By directly measuring the changes in ssDNA extension associated with SSB binding over a range of forces as well as concentrations, we quantify the energy associated with binding and wrapping, which allows us to characterize the energy landscapes for SSB-ssDNA interactions.
**Figure. 2.3.1:** Experimental procedure to probe EcSSB-ssDNA binding dynamics. (A) Schematic depiction of the experimental method to probe EcSSB-ssDNA binding dynamics. 1. An ~8 kbp dsDNA that is recessed (30 nt) at its 3′ end is tethered with its DIG- and biotin-labeled termini to antiDIG- and streptavidin-functionalized beads, respectively. 2. The DNA molecule is held at ~50 pN in the presence of T7 DNA polymerase (T7DNAp) to activate exonucleolysis. 3. A long ssDNA molecule is generated after complete excision by exonucleolysis of the bottom strand. 4. The ssDNA is held at a constant force (F ~12 pN) via a force feedback module and the surrounding solution is exchanged with a solution containing EcSSB. (B) Force extension curves corresponding to the steps described in (A). (C) Change in extension of an ssDNA molecule upon binding to EcSSB as a function of time. EcSSB is introduced to the ssDNA at $t=0$, as denoted from the dashed line. A biphasic change in extension is observed, where a rapid decrease is followed by a relatively slow increase in extension. These two phases are indicative of two distinct ssDNA-EcSSB wrapping configurations.
Results

EcSSB binds ssDNA in a biphasic manner at saturating protein concentrations

To characterize the ssDNA binding kinetics of EcSSB we generate an 8100 nt long ssDNA in situ (Fig. 2.3.1) in an optical tweezers system as described in the legend to the Fig. 2.3.1. The ssDNA is then stretched, held, and maintained at a given tension via a force feedback module (Fig. 2.3.1B). The buffer surrounding the ssDNA molecule is then rapidly exchanged (<1 s) with a solution containing a fixed EcSSB concentration. The binding kinetics of EcSSB to ssDNA is registered as a change in the ssDNA extension. Interestingly, we observe a biphasic binding profile at saturating EcSSB concentrations (<1 nM) wherein a rapid decrease in the ssDNA extension is followed by a relatively slow increase that equilibrates at an extension less than that of a protein-free ssDNA molecule (Fig. 2.3.1C). Because EcSSB is known to wrap the ssDNA in different modes, the biphasic profile is indicative of rapid wrapping events upon protein binding that are followed by unwrapping events as more EcSSB populates the ssDNA molecule. To further investigate this mechanism, we obtained the EcSSB-ssDNA binding kinetics as a function of protein concentration. Interestingly, the timescale at which the second transition occurs increases as the protein concentration is decreased from 50 nM down to 1 nM. This transition completely diminishes at 0.1 nM (Fig. 2.3.2A) resulting in a single-phase binding kinetics profile. Taken together, the results are indicative of a competitive binding mechanism between the wrapped and unwrapped modes of EcSSB. These modes are self-regulated based on the ssDNA occupancy, as the stability of the unwrapped state increases with increasing protein density and vice versa (Fig. 2.3.2A). Accordingly, we propose the following two-step kinetic model to describe the competitive EcSSB binding (Fig. 2.3.2B):
Here, $\Theta_0$, is the free ssDNA fraction, $\Theta_b$ is the ssDNA fraction that is occupied by the bound but unwrapped (or minimally wrapped) EcSSB, and $\Theta_w$ is the ssDNA fraction that is wrapped by EcSSB. The rates $k_b$ and $k_{b^{-}}$ are the bimolecular association and dissociation rates, respectively. The rates $k_w$ and $k_{w^{-}}$ are the wrapping and unwrapping rate of a single EcSSB, respectively and these rates are stimulated by the ssDNA occupancy. We phenomenologically impose this additional dependency by introducing the multiplication factors $\Theta_0$, and $\Theta_b$, on $k_w$ and $k_{w^{-}}$, respectively, which account for influence of ssDNA-occupancy on the effective unwrapping and wrapping rates.

$$\Theta_0 \frac{k_b[SSB]}{k_{b^{-}}} \Theta_b \frac{\Theta_0 k_w}{\Theta_b k_{w^{-}}} \Theta_w.$$  \hspace{1cm} (2.3.1)

Figure 2.3.2: ssDNA-EcSSB binding dynamics at 12 pN. (A) Change in the ssDNA extension ($\Delta x$) upon EcSSB binding at different protein concentrations. At saturating protein concentrations (>1 nM) a biphasic binding profile is observed where a decrease in the extension is followed by an increase of extension before equilibration. At 0.1 nM EcSSB, the second transition vanishes resulting in a single-phase binding profile. (B) Top: Schematic depiction of the proposed kinetic model in Eq. 2.3.1. Bottom: Description of the biphasic binding mechanism. EccSSB binding regulates its own wrapped state, where initial wrapping at low lattice occupancy changes to unwrapped binding at high occupancy.

**Interconversion of EcSSB-ssDNA modes is regulated through protein concentration**

To further investigate the dynamics of the competitive binding of EcSSB to ssDNA and test the proposed model, we conducted concentration-switch experiments at 12 pN (Fig. 2.3.3A). Here, after the EcSSB-ssDNA complex is equilibrated upon initial binding, we rapidly switched the
protein solution with buffer alone, effectively reducing the free protein concentration surrounding the protein-ssDNA complex down to zero. Interestingly, this resulted in a significant extension decrease in the EcSSB-ssDNA complex, indicating a transition to a further wrapped EcSSB-ssDNA state. We denote this the ‘dissociation-wrap’ transition because these wrapping events must be concomitant with protein dissociation (Fig. 2.3.3A). When the protein-free buffer is switched back to the protein solution, the protein-ssDNA complex equilibrates back to the original protein-saturated extension, indicating unwrapping events upon rebinding of EcSSB (rebind-unwrap transition, where $\Delta X_{bu} = \Delta X_{ru}$). The oscillations between the two states that are controlled by free protein concentration can be repeated over many cycles. Thus, the total degree of ssDNA wrapping is inversely related to the free EcSSB concentration, indicating that the high density of bound proteins inhibits further wrapping of ssDNA. To quantify these interconversions, we measure the extension change ($\Delta X$) associated with each transition, as a function of free protein concentration in the solution (Fig. 2.3.3A). At saturating protein concentrations the extension-change associated with the transient bind-wrap transition ($\Delta X_{bw}$), and the final equilibrium extension-change of the bind-unwrap transition ($\Delta X_{bu}$), strongly depend on the protein concentration. However, regardless the starting protein-concentration, the extension change upon removing the free proteins (dissociate wrap transition, $\Delta X_{dw}$) converges at ~0.8 nm/nt, which is also the total extension-change observed with single-phase binding at 0.1 nM EcSSB. This suggests that when the free protein in solution is scarce, the EcSSB-ssDNA complex equilibrates to the maximal wrapping state. The extension change associated with the maximal wrapping state (0.08 nm/nt) observed is consistent with the previously observed extension change associated with a single EcSSB molecule binding in the (SSB)$_{17}$ mode, extrapolated to 12 pN and normalized over the number of nucleotides$^{99}$. 

66
Figure 2.3.3: Concentration-dependent EcSSB interconversion. (A) Representative curve showing dynamic interconversion between wrapped and unwrapped states regulated through competitive binding. Dashed lines indicate the time points at which the EcSSB concentration is switched from 50 nM to 0 nM, and vice versa. ssDNA extension is initially decreased due to EcSSB binding and wrapping (Bind-wrap transition in blue). As the ssDNA saturates with the EcSSB an increase in the extension is observed (bind-unwrap transition in red). This is indicative of a protein-saturated unwrapped EcSSB-ssDNA conformation. A significant decrease in the extension is observed when the protein solution is exchanged with clean buffer that is indicative of a primarily wrapped EcSSB-ssDNA conformation (Dissociation-Wrap transition in green). Finally, when the clean buffer is exchanged back with protein solution the extension is increased and equilibrated back at the protein saturated unwrapped conformation (B) Δx measured at EcSSB-ssDNA conformation shown in (A). Dotted line indicates the Δx for the (SSB)17 as measured in 111 normalized over the number of nucleotides. (C) Observed rates of the different interconversion transitions shown in (A).
Therefore, when free protein in solution is scarce, the long ssDNA-EcSSB complex mimics the binding of an array of isolated single EcSSB binding events, in a non-competitive manner. The extension change equilibriates at much lower values than 0.08 nm/nt with increasing free protein concentration as the unwrapped state becomes more stable. Interestingly, at 50 nM EcSSB, the highest EcSSB concentration tested in this study, where the majority of the EcSSB must be in the unwrapped, bound state, the bind-unwrap extension-change $\Delta X_{bu} (= \Delta X_{nu})$ equals 0.02 nm/nt. This extension-change is four-fold less than the extension-change associated with the maximal wrapping state, which is the $(SSB)_1^7$ mode. Also, given that in the bound-state occupancy of ssDNA saturated with EcSSB must be much larger than that of the complex in the fully wrapped state, the extension-change associated with a single EcSSB in the bound-state must be at least four-fold smaller than that of an EcSSB in the $(SSB)_1^7$ mode, and therefore must occlude fewer than 4 nt.

**Interconversion rates between wrapped and unwrapped EcSSB-ssDNA modes**

In addition to the transition distances, we also measured the transition rates between the observed distinct states, as seen in Fig. 2.3.3A. As discussed in the previous section, each observed transition results from two primary competing processes, as summarized in our proposed model. The rapid decrease in extension upon initial protein binding ($\Delta X_{bw}$) is due to bimolecular EcSSB binding events followed by subsequent wrapping events (bind-wrap transition). As expected for a two step process, where the second step is followed by bimolecular binding, the bind-wrap rate increases linearly with concentration at lower protein concentrations and saturates at concentrations higher than 10 nM. Therefore the saturated bind-wrap rate ($\sim 1.4 \text{ s}^{-1}$) is an estimate of the fundamental wrapping rate, $k_w$, of an EcSSB bound in $(SSB)_1^7$ at 12 pN (Fig. 2.3.2C). Because there is no analytical solution for such competing processes, we model the
concentration (c) dependence of the bind-wrap rate, $k_{bw}$, with the following expression that estimates the resultant observed rate of the two steps (binding and wrapping) involved.

$$\frac{1}{k_{bw}} = \frac{1}{ck_b} + \frac{1}{k_w} \rightarrow k_{bw}(c) = \frac{ck_b k_w}{ck_b + k_w} = \frac{k_w}{1 + \frac{k_w}{ck_b}}$$

(2.3.2)

The second, bind-unwrap transition is primarily associated with the unwrapping events that are concomitant with binding events. The observed rate for this transition, $k_{bu}$, also saturates at higher protein concentrations (Fig. 2.3.3C). Therefore at saturation, $k_{bu}$ estimates the fundamental unwrapping rate ($k_{uw} \sim 0.015 \text{ s}^{-1}$) of an EcSSB bound in the (SSB)$_17$ state at 12 pN. We model the concentration dependence of the bind-unwrap transition, $k_{bu}$ with a simple two-step binding isotherm, where,

$$k_{bu}(c) = k_{uw} \frac{c}{c + c_{bw}} .$$

(2.3.3)

Here, $c_{bw} \sim 3nM$ is the concentration at which the bound and wrapped states are equally populated at 12 pN. The rate of the third, dissociation-wrap transition ($k_{dw}=0.1 \text{ s}^{-1}$) is associated with wrapping events concomitant with protein dissociation and is independent of the starting protein concentration. (Fig. 2.3.3C). However, the $k_w$ as estimated earlier ($\sim 1.4 \text{ s}^{-1}$) is much larger than the observed $k_{dw}$, indicating that this transition is rate-limited by the dissociation events, and therefore, $k_{dw}$ is an estimate of the dissociation rate. Nevertheless, as described in the next section, the directly-measured dissociation rate from the bound state, $k_{-b}$, is much smaller than the $k_{dw}$, measured here. This suggests a mechanism in which the dissociation of the neighbouring EcSSB in the bound state is stimulated by the active wrapping of EcSSB.
**Force-dependence of EcSSB-ssDNA binding kinetics**

To probe the force dependence of the EcSSB-ssDNA binding dynamics, we repeated the competitive binding measurements at 50 nM EcSSB, as a function of the force exerted on the ssDNA substrate.

**Figure 2.3.4: Force-dependant EcSSB-ssDNA binding dynamics.** (A) Biphasic extension-change of EcSSB-ssDNA binding dynamics as a function of force. (B) ΔX associated with each transition, as shown in Fig. 2.3.3A, as a function of force. (C) Concentration-switch experiment at 20 pN, showing EcSSB dissociation in the absence of free protein in the solution. (D) Force-jump experiment, in which the ssDNA is first incubated with EcSSB at 12 pN. After switching the protein solution with protein-free buffer, the force on the protein-DNA complex is suddenly increased up to 60 pN and brought back to 12 pN. Upon re-introducing the protein solution, the rebind-unwrap transition is observed, indicating that some SSB remained bound even at 60 pN.

We observe the biphasic binding mechanism at all the forces tested in this study: 7 pN, 12 pN, and 20 pN (Fig. 2.3.4). The extension change associated with each transition increases with decreasing force, indicating that the wrapped state becomes more stable as the template tension is lowered (Fig. 2.3.4A-B). The maximal extension change is observed after dissociation-wrap transition at 7 pN (ΔX<sub>bw</sub> (7 pN) ~ 0.13 nm/nt). This extension change is consistent with the previously observed extension change associated with the (SSB)<sub>35</sub> mode, normalized over the
numbers of nucleotides (0.14 nm/nt). Therefore at 7 pN, the wrapped state is primarily in the (SSB)$_{35}$ mode. Nonetheless, $\Delta X_{bu}$ at 7 pN is much smaller than the extension change expected for only the (SSB)$_{17}$ mode, indicating that the bound state remains dominant in the presence of free protein, even at 7 pN. At 20 pN, as the wrapped state becomes less stable, the ssDNA is predominantly occupied by SSB in the bound state, even in the absence of free proteins in the solution. Therefore, after the dissociation-wrap transition, we observe a gradual increase in the extension due to protein dissociation (Fig. 2.3.4C). The interpretation that these are dissociation events is supported by the fact that as we re-introduce the protein solution, we do not observe the rebind-unwrap transition, in which an increase in the extension is expected. Instead we observe a biphasic binding profile that mimics the intial bind-wrap and bind-unwrap transitions (Fig. 2.3.4C), indicating protein binding to free ssDNA as a consequence of dissociation events during the prior step. Furthermore, to test if EcSSB remains bound at forces much higher than 20 pN, we conducted force-jump experiments, starting at 12 pN. Here, during the disociation-wrap transition, the EcSSB-ssDNA complex is stretched from 12 pN up to 60 pN and held for ~10s, before bringing it back to 12 pN (Fig. 2.3.4D). The protein-DNA complex equillibriates at an extension slightly higher than the equillibrium extension of dissociation-wrap transition. However, the extension is still less than that of free ssDNA at this force, indicating that the DNA is still partially occupied by EcSSB. This interpretation is further supported by the observation that, as the protein solution is re-introduced, we observe the rebind-unwrap transition. Therefore, the bound EcSSB mode, although it is less stable and dissociates gradually on a timescale greater than tens of seconds, remains in its original binding mode at forces much higher than 20 pN, without being rapidly removed from the DNA substrate, in contrast to the interpretation of some previously reported studies. This observation provides further evidence that the ssDNA is
minimally wrapped by EcSSB in this mode, allowing it to remain bound at forces much higher than 20 pN.

Simulations of EcSSB-ssDNA binding dynamics
To test the proposed two-step kinetic model, we perform simulations, applying the fundamental rates estimated from the observed results at 12 pN (Fig. 2.3.5). The kinetic scheme as described in equation 2.2.1 yields the following set of differential equations that governs the time evolution of unbound (θ₀), bound (θ_b), and wrapped (θ_w) ssDNA fractions.

\[
\begin{align*}
\frac{d\theta_b}{dt} &= \theta_b k_{-b} - \theta_b c k_b + \theta_w k_w (1 - \eta) - \theta_b \theta_b k_w (\eta^{-1} - 1) \\
\frac{d\theta_b}{dt} &= \theta_b c k_b - \theta_b k_{-b} + \theta_b \theta_w k_w - \theta_b \theta_b k_w \\
\frac{d\theta_w}{dt} &= \theta_b k_{-w} - \theta_b \theta_w k_w^{-1} - \theta_b \theta_b k_w \\
\eta &= \frac{N_B}{N_w}
\end{align*}
\]

(2.3.4)

Where N_B and N_w are the number of nucleotides occluded by the bound and wrapped EcSSB modes, respectively. According to a previous study, (SSB)_{17} is the only wrapped state that is stable at 12 pN, which is also consistent with the extension change we observe in the maximally wrapped state at this force. Therefore, the total extension change as a function of time due to EcSSB binding, ΔX(t), is given by

\[
\Delta X(t) = -\left[ \Delta x_b \Theta_b(t) + \Delta x_w \Theta_w(t) \right].
\]

(2.3.5)

where Δx_b and Δx_w are the extension changes associated with the bound and wrapped states, respectively. Accordingly, we use N_w to be 17 nt and the associated change in extension to be 0.08 nm/nt, as observed in Fig. 2.3.3B. The minimum extension change we observe at 50 nM (0.02 nm/nt) serves as an upper bound to the extension change associated with the bound EcSSB state. Therefore, we chose N_b to be 4 nt and the associated extension change to be 0.01 nm/nt.
The time dependence of ssDNA length from our simulation, shown in Fig. 2.3.3, strongly resemble the dynamics observed in our experiments.

Figure 2.3.5: MonteCarlo simulations of the two-step kinetic model. Simulated data with the parameters estimated from the data shown in Fig.2.3.3 at 12 pN. The fundamental parameters used are, $k_b = 0.15 \text{nM}^{-1}\text{s}^{-1}$, $k_b \sim 0.015 \text{s}^{-1}$, $k_w = 0.11 \text{s}^{-1}$, $\Delta x_w = 0.08 \text{nm/nt}$, $\Delta x_b = 0.01 \text{nm/nt}$, $N_W = 18 \text{nt}$, and $N_b = 4 \text{nt}$.

Discussion
An OB-domain in each subunit allows the EcSSB tetramer to bind ssDNA in different topologies that are identified by the number of nucleotides occluded by the tetramer$^{18,24,100}$. These modes are influenced by the solution conditions such as salt, temperature, and DNA to protein ratio.

Analyses on (dT)$_{70}$ substrates have identified the dominant wrapped modes denoted (SSS)$_{35}$, (SSB)$_{56}$, and (SSB)$_{65}$. Recent single molecule assays have significantly enhanced the understanding of EcSSB function. Stopped-flow assays showed that while distinct modes of EcSSB exist in a dynamic equilibrium, EcSSB can diffuse along the substrate DNA even in its maximally wrapped state, adopting a reptation mechanism. Using an optical tweezers assay, Suksombat et al. showed that, in addition to the solution conditions, template tension also can influence EcSSB binding modes, and they identified an (SSB)$_{17}$ mode that is stable at forces higher than 7 pN. These studies were designed to probe the binding properties of a single protein to extract the fine details associated with its DNA interactions, and therefore information on collective binding is not obtained.
Here, we study the collective binding dynamics of EcSSB to a long (~8000 nt) ssDNA lattice. To describe the biphasic binding mechanism of EcSSB to ssDNA, as observed with force-clamp experiments, we propose a phenomenological two-step kinetic model. According to this model, EcSSB first binds via a bimolecular process to form a minimally wrapped conformation (bound state) that may convert to a wrapped state. Subsequent interconversion between the wrapped and bound states is solely determined by the EcSSB occupancy on the ssDNA lattice. To further test this model, we performed concentration-switch experiments, in which the solution surrounding the ssDNA molecule was rapidly switched between a solution containing a fixed protein concentration and protein-free buffer for long time intervals. We observe that at 12 pN, the EcSSB-ssDNA complex transforms to the maximally wrapped state in protein-free buffer, which is reversible by replenishing the free protein in the solution. By analyzing equilibrium extension changes ($\Delta X$) in the EcSSB-ssDNA complex associated with each of these transitions, we revealed that at the maximally wrapped state all the EcSSB is in the (SSB)$_{17}$ mode at 12 pN. This also implies that as the protein concentration increases, the bound unwrapped state becomes more stable.

Given that the $\Delta X$ of the protein-DNA complex associated with the presence of the highest concentration (50 nM) is four-fold smaller than that of the the maximally wrapped EcSSB-ssDNA complex, and taking into account that as the bound-state becomes more stable, the ssDNA lattice must accommodate significantly more proteins than that of ssDNA in the maximally wrapped complex, our results suggest that the bound but unwrapped state of EcSSB does not wrap the ssDNA, or perhaps just minimally ($<17/4$ nt) wraps ssDNA. This is a surprising result, which is a consequence of collective and competitive binding that was not detectable in previous (dT)$_{70}$ binding assays. Because this binding mode minimally wraps the DNA, EcSSB
remains bound in this state even at forces higher than 20 pN, which we report here for the first time.

By analyzing the concentration-dependence for the bind-wrap, bind-unwrap and dissociate-wrap transitions as observed in the concentration-switch experiments at 12 pN, we estimate the fundamental rates in the proposed two-step kinetic model. By comparing the directly-measured dissociation rate at 20 pN to the dissociation-wrap rate measured at 12 pN, we show that protein dissociation during active wrapping ($k_{dw}$) is ~ten-fold faster than the dissociation rate from the bound, unwrapped state ($k_b$). Since we do not directly observe protein dissociation at 12 pN, we make the assumption that $k_b$ is force-independent, which has been shown to be the case with off rates as measured with other DNA binding ligands\textsuperscript{77-79}.

Taken together, our results describe mechanisms to regulate the density of the ssDNA-bound EcSSB, which have important biological implications for its transient role during genome maintenance and replication. For instance, to date we do not know how a tightly-wrapped protein such as EcSSB dissociates to allow access to a advancing enzyme. Based on our observed results, we propose the following, self regulation of protein density (SRPD) mechanism (Fig. 2.3.6) that can be directly tested with simple experiments suggested in the ‘future experiments’ section. Let us assume that EcSSB initially primarily exists in the intermediately wrapped (SSB)$_{35}$ mode. This assumption is supported by a recent study on human mitochondrial SSB (mtSSB, which is structurally and functionally similar to EcSSB), which showed mtSSB exists in (SSB)$_{35}$ mode on transiently-formed ssDNA during replication\textsuperscript{80}. However, we must also note that EcSSB exists in a dynamic equilibrium between its distinct modes that are able to diffuse along the DNA without dissociation\textsuperscript{24,83}. Therefore, a processing enzyme, for instance, a DNA polymerase, may displace the wrapped EcSSB during synthesis, pushing it forward along the
ssDNA. This process would likely increase EcSSB density as the transient ssDNA is transformed to its ds form.

**Figure 2.3.6: Proposed self-regulation of protein density (SRPD) mechanism.** (#1) Intermediately wrapped EcSSB diffuses as the dsDNA synthesis proceeds. (#2) Increased protein density on ssDNA prompts unwrapping events. (#3) By converting to a further wrapped state, neighboring proteins competitively capture ssDNA segments that are released by unwrapping events. (#4) Wrapping events stimulate the dissociation of the unwrapped SSB. (#5) Unwrapping of the maximally wrapped SSB due to increased protein density recovering the protein-denisty at #1.

This may in turn cause the EcSSBs to unwrap and release segments of free ssDNA, as observed here in the presence of high EcSSB concentration. A neighbouring EcSSB may compete for the free ssDNA to transform into a further wrapped state, ejecting the EcSSB in the bound-state. As the transient ssDNA lattice gets smaller, this dynamic process may continue and self-regulate the protein density to allow the enzyme to proceed while ensuring maximal coverage on the remaining transient ssDNA at any given time.
Chapter 3. Single molecule characterization of LINE-1 ORF1p-nucleic acid interactions

The work described in this chapter is done in collaboration with Furano Lab, NIDDK, NIH, Bethesda, MD

3.1 Background

This section was adapted from the following review article that is currently in press for publication\textsuperscript{65}.


Figure 3.1.1: L1 Retrotransposition cycle and target-primed reverse transcription. The genomic L1 DNA is transcribed and the L1 transcript encodes for two proteins, ORF1p and ORF2p. The two proteins associate with their encoding transcript (cis preference) to form a ribonucleoprotein complex (RNP), which mediates the new L1 insertion via TPRT as shown in the inset. During TPRT, the target site is cleaved and then annealed with the L1 RNA to prime the reverse transcription and extend the first L1 cDNA. Then the second genomic site is cleaved and primed with the first L1 cDNA to synthesize the second L1 DNA. During TPRT ORF2p functions as the endonuclease and the reverse transcriptase. Presumably, ORF1p’s nucleic acid chaperone capabilities mediate strand exchange reactions required to prime the cDNA synthesis and facilitate the nucleic acid rearrangements required during reverse transcription.

The long interspersed nuclear element 1 (LINE1, L1) non-LTR retrotransposon is an autonomously replicating genomic parasite (Fig. 3.1.1) that has been amplifying and evolving in

77
mammalian genomes for >100 Myr and now constitutes 20% or more of certain mammalian genomes\textsuperscript{25,27,28}.

In addition, L1 can also copy non-L1 transcripts into genomic DNA, and as a consequence L1 activity has generated upwards of 40% of the mass of many mammalian genomes\textsuperscript{29,30}. Only a subset of L1 elements actively retrotranspose in the mammalian genomes\textsuperscript{31,138-140}. Currently, 80-100 L1 copies that belong to a single subfamily T\textsubscript{a} is active in human genome\textsuperscript{141} while ~3000 L1 copies of three subfamilies, T\textsubscript{F}, A and G\textsubscript{F}, are active in the mouse genome\textsuperscript{142-144}. Replication and evolution continue in humans\textsuperscript{145-147}, and L1 activity creates genetic diversity and various genetic alterations\textsuperscript{148-151}. Given its detrimental\textsuperscript{152-154} and potential catastrophic effects\textsuperscript{32,155}, explaining its persistence as well as determining a mechanistic understanding of its replication remain formidable biological issues.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure3.1.2.png}
\caption{L1 and ORF1p (A) L1 domain structure. The domain organization of a typical full-length human L1 element. Approximate positions (not to scale) of the 5' untranslated region (5' UTR), ORF1, ORF2 (which includes endonuclease (EN) and reverse transcriptase (RT)), 3' UTR, and the poly A tail are noted. (B) Domain structure and trimeric depiction of ORF1p. The amino acid positions of the N-terminal domain (NTD), coiled coil domain (CC); RNA recognition motif (RRM) and C-terminal domain (CTD) for mouse\textsuperscript{99,145,156,157} and human\textsuperscript{158-160} ORF1p are denoted in red and blue text, respectively.}
\end{figure}
Mammalian L1 elements are 6–7 kb in length, contain a regulatory 5’ UTR, two long open reading frames that encode two proteins ORF1p and ORF2p, and a 3’ UTR (Fig. 3.1.2A). Both ORF1p (Fig. 3.1.2B) and ORF2p are required for retrotransposition. The two proteins associate with their encoding transcript to form a ribonucleoprotein complex (RNP), which mediates retrotransposition (Fig. 3.1.1). L1 replicates by reverse transcription of its transcript. While this is analogous to retroviruses and retroviral-like retrotransposons, L1 replicates by the dramatically different process of target-site-primed reverse transcription, or TPRT (inset of Fig. 3.1.1). In this mechanism reverse transcription of the L1 RNA transcript is primed from a 3’ hydroxyl at the genomic insertion site. Highly conserved endonuclease and reverse transcriptase domains in ORF2p indicate its replicase function in TPRT. In contrast, ORF1p lacks any known enzymatic domains, although it does contain highly conserved non-canonical RNA binding domains and phosphorylation sites, that are required for retrotransposition activity.

Human and mouse ORF1 encode a ~40 kDa protein and early studies showed that they could be isolated as L1 RNP particles associated with their encoding L1 RNA. As presented in detail below, subsequent studies showed that the 40 kDa monomer forms a coiled coil-mediated trimer that binds nucleic acids with high affinity (Fig. 3.1.2B). In vitro studies using bulk biochemical and single molecule assays showed that human and mouse ORF1p (from here on referring to the trimer) act as nucleic acid chaperones, presumably during TPRT, however, at which steps and by what means are unknown. In addition, both human and mouse ORF1p can polymerize in the presence or absence of nucleic acids. Recent structural and theoretical studies on human and mouse ORF1p addressed its complex nucleic acid binding properties, yet we do not understand how they contribute to the molecular mechanism of
L1 retrotransposition. We now review the current understanding of ORF1p in the context of those nucleic acid binding properties that had been obtained with the purified protein.

Therefore, we will not be discussing the extensive and informative cell biological literature on the structure and properties of the L1RNP and its possible interaction with host factors as exemplified by two recently published papers\textsuperscript{174,175}.

**ORF1p structure**

The amino- and carboxy-terminal halves of mammalian ORF1p evolved under dramatically different evolutionary constraints. The carboxy terminal half is highly conserved (Fig. 3.1.3) and contains a non-canonical RNA recognition motif (RRM) and a distinct C-terminal domain (CTD) (Fig. 3.1.2B). Residues in the carboxy-terminal half mediate high affinity RNA binding and nucleic acid chaperone activity\textsuperscript{160,162,168-170,175,176}. In contrast, the sequence of the amino-terminal half can be highly variable, although a coiled coil motif is conserved throughout vertebrate evolution\textsuperscript{153,156,177}. Direct visualization of mouse ORF1p using atomic force microscopy revealed an elongated dumbbell-like structure, consistent with its trimeric conformation\textsuperscript{178} mediated by the coiled coil \textsuperscript{158,160,172}, which is parallel in nature and stabilized by additional inter-chain hydrogen bonds\textsuperscript{158}. Molecular structures for the human C-terminal half\textsuperscript{159,172} and the mouse CTD\textsuperscript{172} domain were solved. The RRM domain of human ORF1p is structured with a typical $\beta_4\alpha_1\beta_4\beta_4$ fold and contains non-canonical RNPI and RNP2 sequences (Fig. 3.1.3) with aromatic side chains that could mediate base-stacking or hydrophobic interactions with nucleic acid substrates. It also contains two highly conserved salt bridges (E165-R215 and E169-R202) that are formed between two loops, L($\beta_1-\alpha_1$) and L ($\beta_2- \beta_3$)\textsuperscript{157,172}. The L ($\beta_2- \beta_3$) loop is intrinsically disordered\textsuperscript{172} and contains the two T/SP motifs (T203P204, and T213P214) out of the four (S18P19, S27P28 in the NTD) highly conserved proline-directed protein kinase (PDPK) targets
in ORF1p. T or S phosphorylatable residues at these respective sites are required for L1 retrotransposition\textsuperscript{166,172}.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{alignment}
\caption{Alignment of mouse and human ORF1p sequences. Consensus sequences of the mouse L1Tf family\textsuperscript{172} and the human Ta1 subfamily of the currently active L1Pa1\textsuperscript{145} family were aligned. The middle L1Tf/Ta1 sequences shows the comparison between the aligned elements. A dot indicates identity, dashes indicate gaps, and lower case indicates conservative amino acid differences. The bottom entry (Identical) shows the 100\% identical positions. The numbers in black refer to the L1Tf, and those in red, the L1Ta1 sequence. The beginning and end of the coiled coil domain, and the starts of the RNA recognition motif (RRM) and C-terminal domain (CTD) are indicated. The alignment is also annotated with the following information: The beginning of truncated M128\textsuperscript{160}, which is largely a monomer at \( \geq 20^\circ\text{C} \); the location of a natural variant in the mouse coiled coil domain (D159H, using the L1Tf numbering with an offset of plus 4 in the mouse alignment only to account for the 4 position gap introduced into the alignment of L1Tf between positions 40 and 45)\textsuperscript{160}; 34,167,169 paired mutations in two adjacent highly conserved arginine residues of the CTD, in human (34,172) and in mouse\textsuperscript{37}. Heptad repeats are highlighted in green and yellow. RNP1 and RNP2 sequences are highlighted in light brown\textsuperscript{170}. Regarding the “phenotype” of these mutations: retro means retrotransposition as measured in a cell culture based retrotransposition assay\textsuperscript{34,159}; RNP means presence of ORF1p in RNP particles isolated in vitro; RNA binding means as performed with purified ORF1p protein in vitro; chaperone means chaperone activity (as described in\textsuperscript{34}).}
\end{figure}
A crystal structure of human ORF1p, which lacked most of its N-terminal half, yielded three distinct orientations consistent with three structural states for the CTD: resting, lifting, and twisting. The π-stacking interaction between Y282 and R155 was shown to primarily influence the dynamics of the CTD. A lifting motion of the CTD was proposed to open the RRM–CTD cleft and expose the basic residues for nucleic acid binding. Subsequent twisting or rotational motions would then wrap the nucleic acid, such that a continuous strand would occupy binding sites at each ORF1p monomer.

Figure 3.1.4: ORF1p polymerization. Schematic representation of possible ORF1p species and their cross-linked products as observed in Callahan et al. The numbers to the left of the cartoons of the cross-linked species indicate their monomer content. 4..n and 5..n indicate higher orders of multimers beyond 3 or 4, respectively. Higher orders of ORF1p multimers of trimers were observed in crosslinking experiments with 1 mM EGS, in the presence of oligonucleotides at 0.05 M NaCl.

The resting position of the CTD render the basic patch residues of the coiled coil domain accessible, which could allow further nucleic acid binding. This binding mechanism is speculative, however, because the structure did not contain nucleic acid. To support the proposed binding mechanism, several mutants were generated and probed in NA binding and retrotransposition studies. For instance, the highly conserved R261 (Fig. 3.1.3), at which substitutions eliminated retrotransposition both in human and mouse ORF1p, was
predicted from the structure to inhibit RNA binding, which was also confirmed by size-exclusion chromatography. Other mutations, such as R220A, abolished retrotransposition but did not inhibit RNA binding. The effect of this and other similar mutations was proposed to be due to a defect in ORF1p flexibility and dynamics based on the structure.

However, it is important to keep in mind that, as mentioned above, the crystal structure was derived from a protein that lacks the NTD and slightly more than half of the coiled coil domain, both of which contain structures critical to ORF1p function. The intrinsically disordered NTD contains highly conserved PDPK phosphorylation sites\textsuperscript{170} and the coiled coil facilitates rapid polymerization of the trimers on single-stranded NA\textsuperscript{166}, two functions that are vital for L1 retrotransposition. Furthermore, the structural confirmation ORF1p may significantly vary in the presence of nucleic acids. Therefore, although the ORF1p C-terminal half structure suggests that flexible inter-domain orientations are functionally important, the exact details of the structure-function correlation may differ for full length protein. Nevertheless, it is clear that multiple domains in ORF1p simultaneously interact with nucleic acids and are also capable of mediating inter-protein interactions\textsuperscript{66} (discussed in the next section, see Fig. 3.1.4). This results in a complex set of possible configurations, which in turn appear to depend strongly on solution conditions. An understanding of these configurations is essential to determine how ORF1p facilitates different stages of retrotransposition.

**Nucleic acid binding properties of ORF1p**

Several early studies demonstrated co-localization of ORF1p and L1 RNA in L1 RNPs. For example, L1 RNA of L1RNP isolated from mouse embryonal carcinoma cells was efficiently cross-linked to protein by a brief exposure to ultraviolet light, indicating that ORF1p and L1 RNA were in close contact\textsuperscript{160}. Similarly, high concentrations of monovalent cations and RNase
dissociated L1 RNPs isolated from human teratocarcinoma cell lines, again indicating co-
localization between the protein and L1 RNA\textsuperscript{167,180}. Co-localization of ORF1p with L1 RNA was
further confirmed when HeLa cells were transfected with an L1 that contained an epitope tagged
ORF1p \textsuperscript{181}. Subsequent quantitative binding studies of ORF1p to NA showed that ORF1p binds
without sequence specificity to both RNA and DNA\textsuperscript{37,66,75,160,169,170,176}.

Electrophoretic mobility shift assays showed that mouse L1 ORF1p binds RNA with positive
cooperativity, providing early evidence of protein-protein interactions between the polypeptides
\textsuperscript{182}. The nature of protein-protein interactions was extensively studied in Callahan et al.\textsuperscript{183} with
human ORF1p purified from insect cells (Fig. 3.1.4). In this study, the cross-linking reagent
EGS, which interacts with primary amines (ex: $\varepsilon$-amino group of lysine), was used to capture
ORF1p interaction products, prior to treatment with 8 M urea, which fully denatures non-
crosslinked ORF1p into its constituent components, \textit{i.e.}, monomers, partial trimers, and multiples
of trimers (Fig. 3.1.4). In 0.5 M NaCl ORF1p is a soluble trimer but is unable to bind nucleic
acids. Optimal nucleic acid binding occurs at 50 mM NaCl, but in the absence of nucleic acids,
ORF1p trimers rapidly form massive polymers that precipitate from solution. When cross linked
with 1 mM EGS the cross-linked polymers are too large to enter a 6% denaturing polyacrylamide
(PAGE) gel. Addition of equimolar oligonucleotides (N$\geq$20) immediately resolved the ORF1p
polymers to trimers or multimers of trimers (depending on the length of the oligonucleotide,
(Fig. 3.1.4). Therefore, ORF1p remains active in its polymeric conformation, its presumed state
in the L1RNP. Partial crosslinking occurs with limiting EGS (0.05 mM) and generates a ladder
of products starting at the size of the $\sim$40 kDa monomer and multimers thereof (Fig. 3.1.4).
Carrying out these experiments with an amino-terminal truncated ORF1p (M128, Fig.2.3)\textsuperscript{160},
which contains just 3.5 heptads of the 14-heptad coiled coil and is largely a monomer at $\geq$ 20 °C,
showed that the amino acid residues that mediate inter-trimer interactions reside in the conserved C-terminal half of the protein.

The ability to form large oligomers in the absence of nucleic acid may provide a mechanism to prevent the diffusion of the newly synthesized protein into the cytoplasm before binding to L1 RNA. However, the relationship between ORF1p translation from the L1 transcript and formation of the L1 RNP with the L1 transcript remains obscure as any such large clusters must dissociate in order to optimally coat the transcript. ORF1p binds RNA and ssDNA with similar affinity, but with an ~8-fold lower affinity to a perfectly matched duplex. In contrast, ORF1p binds mismatched duplexes with about the same affinity as ssDNA, and, in fact, stabilizes it from dissociation\textsuperscript{160}. This occurs at near equimolar or a 10-fold higher molar excess of ORF1p/duplex but at sufficient molar excess, ORF1p dissociated the duplex. The monomeric form of ORF1p (\textit{i.e.}, N-terminal truncated M128p, Fig. 3.1.3) bound nucleic acids with substantially lower affinities, in comparison to the trimer, and could not protect a mismatched duplex from dissociation\textsuperscript{160}, even though M128p contains the intact RRM and CTD domains critical for both high affinity nucleic acid binding and NA chaperone activity\textsuperscript{157,159,160,170}. These results indicate that intra-trimeric interactions play vital roles for the mechanism of nucleic acid binding by ORF1p.

Single molecule DNA stretching studies have been used to quantitatively characterize the affinity of mouse ORF1p for DNA. Fig. 3.1.5A shows a DNA stretching curve in the presence and absence of wild type mouse ORF1p. The width of the transition, shown as $\delta F$ in Fig. 3.1.5A, demonstrates the strong effect of ORF1p on DNA conformation. Transition width is positively correlated with protein binding to DNA, so the width as a function of concentration was used to determine the DNA binding affinity by fitting to a binding isotherm\textsuperscript{171}. Here the protein
concentration dependence of the change in the DNA force-extension profile is used to determine the DNA binding affinity. Below we will discuss the biophysical interpretation of the change in transition width in relation to nucleic acid chaperone activity. The DNA binding affinity measurements demonstrate that mouse ORF1p binding to DNA is in the range of 10 nM in near-physiological solution conditions (50-100 mM Na\(^+\)). However, many ORF1p mutants did not alter binding affinity but strongly affected retrotransposition\(^{169}\). Thus, more complex interactions beyond simple nucleic acid binding determine the effect of mouse ORF1p on retrotransposition, as will be discussed below in the section on nucleic acid chaperone activity.

**Nucleic acid chaperone activity of ORF1p**

The L1 retrotransposon, like retroviruses, replicates by reverse transcription of an RNA intermediate. Reverse transcription of retroviruses is facilitated by the nucleic acid chaperone activity of its nucleocapsid (NC) proteins. This appears to be achieved by lowering the energy barrier of rearranging nucleic acids to achieve maximal base complementarity\(^{170}\). Therefore, in vitro, a nucleic acid chaperone will primarily mediate nucleic acid annealing and strand exchange reactions. NC proteins from multiple retroviruses and retroelements\(^{76,184-186}\) act as nucleic acid chaperones and contain the CCHC zinc finger motif, critical for retroviral replication processes\(^{72,187}\). In the case of HIV-1 NC, nucleic acid chaperone activity is positively correlated with retroviral infectivity\(^{70,73}\). ORF1p of I factor, a LINE-like element of *Drosophila melanogaster*\(^{74}\), also behaves as a nucleic acid chaperone\(^{188}\). Although some retrotransposons such as Ty3\(^{189}\) contain an NC protein, L1 does not. Thus, as mammalian L1 ORF1p does not contain a CCHC motif it likely does not function by the same mechanism as NC proteins. Nevertheless, both mouse and human L1 ORF1p were shown to be nucleic acid chaperones\(^{162,187}\).
To address the mechanism of ORF1p chaperone activity, mutational analyses on mouse ORF1p were performed to correlate *in vitro* nucleic acid binding with retrotransposition. As discussed above, DNA binding affinities for several mutants were determined by single molecule DNA stretching assays using helix-coil transition width as a function of ORF1p concentration. The helix-coil transition in the absence of protein is very narrow (δF is small), due to cooperative conversion of dsDNA to ssDNA, *i.e.*, the requirement for near simultaneous melting of multiple base pairs. Therefore, it was suggested that when an NA-binding protein increases the transition width (transition is less cooperative) it does so by lowering the energetic barrier to rearrangement of nucleic acid secondary structure, thereby promoting melting of small numbers of base pairs.

**Figure 3.1.5: Single molecule characterization of nucleic acid chaperone activity.** A) FECs (extension – solid lines, return – dashed lines) of DNA with no protein (black) and in the presence of 15 nM, mouse ORF1p WT (figure adapted from 160). The transition width, δF is determined by the intersection of the dashed red lines on the figure, which represent ds- and ss- DNA regimes of the ORF1p-DNA complex. B) The relative increases in the transition width, ΔF (ΔF = δF − δF₀ where δF, δF₀ are the measured transition widths in the presence and absence of protein, respectively) for several mouse ORF1p variants (blue) are compared with their retrotransposition activity in an *in vivo* cell-culture assay (red)165,75,169. The values are normalized with respect to the wild type values. Extensive aggregation of single- or double-stranded DNA in comparison with the wildtype protein are denoted with a plus sign.
Thus, measurement of the transition width reflects both DNA binding affinity and its chaperone activity; i.e., that a greater transition width represents greater chaperone capabilities. When fully bound by protein, mouse ORF1p mutants altered the transition width to different degrees, and the extent to which the proteins increased the transition width at saturated protein binding appeared to correlate with retrotransposition activity. This result is shown in Fig. 3.1.5B, which shows the transition width at high protein concentration, near saturated binding, for several mutants \(^{72}\). As described above, HIV-NC protein dramatically increased the transition width \(^{70,72,190,191}\) as also observed with ORF1p. Thus, it was suggested that the mutations reduced ORF1p nucleic acid chaperone activity, which was responsible for reduced retrotransposition \(^{192}\). However, the exact mechanism by which this dramatic increase in helix-coil transition occurs is not clear. A similar effect is also observed in the presence of DNA intercalators such as ethidium \(^{170}\) and ruthenium complexes \(^{78,193}\), which prevent DNA strand separation. Therefore, the observed increase in the helix-coil transition width in the presence of HIV-1 NC was subsequently attributed to the simultaneous destabilization of the duplex while maintaining protein-mediated interactions between the two DNA strands \(^{74,78}\). Below we discuss the ORF1p mutations altered both retrotransposition and their interaction with DNA as measured by single molecule DNA stretching. The mutations are shown in Fig. 3.1.3 and the results from stretching studies are summarized in Fig. 3.1.5B.

Substitutions in the highly conserved R284 (Fig. 3.1.3, note the +4 offset in the mouse ORF1p residue positions as described in the legend to Fig. 3.1.3) in the RRM and RR297-298 and Y318 in the CTD as well as D159 in the coiled coil of mouse ORF1p (Fig. 3.1.2) were analyzed by DNA stretching \(^{75,170,186}\). Alanine substitutions at 284 and 318 (R284A and Y318A) eliminated retrotransposition, while substitutions that conserved charge or hydrophobicity (R284K and
Y318F) were active in retrotransposition. In contrast, KR297-298, KK297-298 or AA297-298 substitutions at RR297-298 eliminated retrotransposition while the RK297-298 mutation remained active, so simply preserving the charge is not always sufficient to retain retrotransposition activity. A D159H substitution at the coiled coil lowered the retrotransposition frequency by a factor of ~15.

As shown in Fig. 3.1.5B, many of the mutants show a direct correlation between reduced transition width and reduced retrotransposition relative to wild type ORF1p (R284A, R284K, Y318A, Y318F, R298K), suggesting that the reduced retrotransposition is due to inhibition of ORF1p chaperone activity. However, there are some outliers, which show similar transition widths to wild type ORF1p but are defective in retrotransposition (D159H, R297K, RR297:298KK). These outliers required further analysis to understand why retrotransposition was inhibited, however we note that D159H is a coiled coil mutant and in section 3.5, we explore the effect of coiled coil mutants of human ORF1p and their effect on retrotransposition.

To further understand the interactions between these additional ORF1p mutants and DNA, the full stretching curve was analyzed. In addition to an increase in transition width, extensive dsDNA aggregation by dsDNA binding proteins can increase the force required to reach the dsDNA contour length (RR297-298 KR, RR297-298KK, and RR297-298AA in Fig. 3.1.5B)\(^{37}\). In a recent study, the dsDNA aggregation by NC proteins from SIV and HIV-1 were quantitatively measured and used to differentiate the nucleic acid chaperone activity of these two proteins\(^{75}\). Strong dsDNA aggregation and wild-type transition widths was observed for RR297-298KR and RR297-298KK. While this shows that these proteins behave differently than wild type ORF1p, we do not know why retrotransposition is inhibited. Although dsDNA aggregation is a strong component of nucleic acid chaperone activity, aggregation can also inhibit chaperone
activity if it slows protein-nucleic acid interaction kinetics\textsuperscript{194}. The reduced retrotransposition for RR297-298AA was attributed to weak RNA binding. In comparison, all of the other mutants studied showed RNA binding affinity similar to wild type ORF1p. Finally, while D159H showed a wild type transition width and wild type levels of dsDNA aggregation, it was defective in retrotransposition. However, upon analysis of the full stretching curve, it was found that the defective activity of the protein is correlated with its extensive ssDNA aggregation. The DNA observed at extensions beyond 0.55 nm/bp is mostly single-stranded (Fig. 3.1.5A) in the salt conditions (50-100 mM Na\textsuperscript{+}) used in the studies discussed here\textsuperscript{61,62,194}. Therefore, a relative decrease in ssDNA extension after the helix-coil transition in the presence of protein is indicative of ssDNA aggregation due to bound protein. This reduced retrotransposition activity resulting from ssDNA aggregation may also be attributed to a reduction in protein-DNA interaction kinetics\textsuperscript{59,169}. Thus, all of the mutants described here that exhibited wild type nucleic acid binding but were defective in retrotransposition showed a difference in ssDNA aggregation, dsDNA aggregation, or helix-coil transition properties in single molecule stretching experiments. These differences in turn may be related to nucleic acid chaperone activity. However, the mechanisms behind these changes were not determined in wild type and mutant mouse ORF1p studies. Therefore, understanding the reason for the observed behavior will require quantitative DNA interaction kinetics studies and characterization of ORF1p oligomerization states as discussed in the next section\textsuperscript{160,194}.
3.2 L1 retrotransposition requires rapid ORF1p oligomerization, a novel coiled coil-dependent property conserved despite extensive remodeling

The work described in this section was adapted from the following publication66.

M. Nabuan Naufer, Kathryn E. Callahan, Pamela R. Cook, Cesar E. Perez-Gonzalez, Mark C. Williams, and Anthony V. Furano. L1 retrotransposition requires rapid ORF1p oligomerization, a novel coiled coil-dependent property conserved despite extensive remodeling. Nucleic Acids Research 44: 281-293 (2016)

Protein purification and ensemble results are the work of Kathryn E. Callahan, Pamela R. Cook, Cesar E. Perez-Gonzalez, and Anthony V. Furano at NIDDK, NIH, Bethesda, MA

Introduction

To address the biochemical consequences of coiled coil evolution we resuscitated an ORF1p that was encoded by the ancestral primate L1Pa5 family, which went extinct about 25 MYA66,153. The emergence of this family coincided with a major episode of amino acid substitutions primarily in the coiled coil that ultimately generated the modern human L1Pa1 family ORF1p (hereafter called 111p). This protein differs from L1Pa5 ORF1p (555p) at 42 positions, 30 of which are in the coiled coil, which accounts for 101 residues of the 381 amino acid protein195 (Fig. 3.2.1A). We also constructed several mosaic proteins that contained mixtures of ancestral and modern coiled coil residues.

Although 555p and 111p were equally active in retrotransposition, one mosaic protein, 151p (9 ancestral coiled coil amino acids), was completely inactive. But the purified proteins were similar in their interactions with NAs including binding affinity and chaperone activity when tested with oligonucleotide substrates (≤ 120-nt) traditionally used in these assays. However, by using single molecule stretching studies with λ phage DNA113,153 we substantially advanced our understanding of ORF1p/NA interactions. In particular, we detected and quantified distinct populations of ORF1p-NA complexes that differed in their dissociation kinetics.
Figure 3.2.1: ORF1p variants. (A) Distribution of amino acid differences between 111p, 555p, and 151p. (B) The retrotransposition reporter. Upon transfection, a full-length L1 transcript is synthesized, and spliced (removes the inactivating intron in the neo gene). Retrotransposition competent elements support subsequent cDNA synthesis of this transcript at a DNA target site and ultimately the insertion of an active copy of the neo gene, which when expressed from its promoter (Pr, in red) generates colonies of G418 resistant cells or foci (see Methods). (C) Retrotransposition activity of ORF1p variants. (D) Western blot of 75 μg of extracts of control HeLa cells or those transfected with vectors that express the indicated ORF1p fused at the carboxy-terminus with the FLAG epitope. These vectors were prepared, and the transfections were carried out as described in Methods. (E) Denaturing gel electrophoresis under reducing conditions of ~200 ng each of 111p, 151p, and 555p purified from insect cells as described in Methods and ref.75.

And thereupon we uncovered a significant difference between the proteins - the retrotransposition incompetent 151p formed stably bound oligomers on ssDNA at an order of magnitude lower rate than the active proteins. Thus, rapid formation of stable NA-bound ORF1p polymers is positively correlated with retrotransposition.
Figure 3.2.2: Binding of ORF1p variants to mismatched duplex oligonucleotide. Binding, protection, and eventual melting of mismatched double stranded 29-nt by the indicated ORF1p as a function of protein concentration. (A) ORF1p variants were incubated with 0.1 nM radioactive mismatched duplex 29-nt for 1 h at 37 °C using the filter-binding assay described in Methods. (B) Autoradiogram of dried polyacrylamide gel of selected samples (heavy tick marks in panel (A) labeled a to e) of the NA binding reaction shown in (A) and the 0 protein controls that were incubated for 1 h at either 0 or 37 °C. Samples were electrophoresed as described in the Methods. (C) Fraction ssDNA of the selected samples relative to zero protein control. These values were determined by quantifying the ds and ss NA species shown in (B) as described in Methods.
Not only is the rate of oligomerization determinative for retrotransposition activity, it is sensitive to the amino acid composition of the coiled coil. Therefore, we discuss these results in the context of both the mechanistic and evolutionary properties of ORF1p.

**Results**

**Comparisons of modern, ancestral and mosaic ORF1p**

Fig. 3.2.1A summarizes the amino acid differences between the modern, mosaic, and ancestral versions of ORF1p: 111p, 151p and 555p. Modern 111p is encoded by the L1.3 element\(^{59}\), a member of the currently active Ta-1 L1Pa1 subfamily\(^{147}\). Ancestral 555p is encoded by the resuscitated ORF1 that we derived from a consensus sequence of 840 L1Pa5 sequences as described in Methods. The mosaic 151p was created by swapping in the ancestral coding region encompassed by the conserved BsmI and BsmFI sites as described in the Methods.

Using the retrotransposition assay depicted in Fig. 3.2.1B we found that 555p and 111p are equally active. However, the 151p mosaic protein cannot support retrotransposition (Fig. 3.2.1C). This result is not due to an inherent inability of 151p to be stably expressed in HeLa cells, as all three are expressed to about the same extent (Fig. 3.2.1D). To investigate the basis of this inactivity we purified 111p, 555p and 151p that had been expressed in baculovirus infected insect cells as described in Methods and ref.\(^{147}\). Denaturing gel electrophoresis (Fig. 3.2.1E) shows that the proteins were essentially homogenous. The bands at ~80 kDa are likely not fully-denatured trimers and the band between the 66 and 55 kDa markers is an apparent staining artifact as it was not seen on other gels (ex: inset in Fig. 3.2.3).
Figure 3.2.2: Binding of ORF1p variants to mismatched duplex oligonucleotide. Binding, protection, and eventual melting of mismatched double stranded 29-nt by the indicated ORF1p as a function of protein concentration. 

(A) ORF1p variants were incubated with 0.1 nM radioactive mismatched duplex 29-nt for 1 h at 37 °C using the filter-binding assay described in Methods. (B) Autoradiogram of dried polyacrylamide gel of selected samples (heavy tick marks in panel (A) labeled a to e) of the NA binding reaction shown in (A) and the 0 protein controls that were incubated for 1 h at either 0 or 37 °C. Samples were electrophoresed as described in the Methods. (C) Fraction ssDNA of the selected samples relative to zero protein control. These values were determined by quantifying the ds and ss NA species shown in (B) as described in Methods.
Figure 3.2.3: NA chaperone activity of ORF1p variants. (A) Schematic of the FRET assay to measure the annealing and exchange phases of NA chaperone activity using the assay conditions described in the Methods. (B) Reactions (40 μl, 0.5 μM ORF1p, 5nM 21-nt mismatched duplex) were incubated at room temperature, irradiated every 0.7 sec at 535 nm, and fluorescence was measured at 590 and 680 nm (solid circles). At the indicated time (vertical arrow) 50 nM perfect complement was added. The relative concentrations of duplex and ORF1p during the annealing phase of these reactions correspond to those that produce maximal protection of the mismatched duplex (caging, Figure 2A, grey rectangle). The annealing and exchange phases data are fit to single and double exponential functions in time (solid lines and Supplementary Table S2) respectively, using the minimization of χ² method.

Interactions of modern, ancestral and mosaic ORF1p with oligonucleotides

Earlier work showed that while purified 111p bound a 29-nt duplex DNA with 1/10ths the affinity as its single-stranded counterpart, it bound a mismatched version of the duplex with the same affinity as the single strand i.e., the protein “sensed” the mismatched duplex as an ssDNA.

This was not due to the protein melting the mismatched duplex and then binding the released single strand. Rather, the protein protected the mismatched duplex from melting (which we termed caging) before eventually melting it at higher molar excesses of ORF1p to
Fig. 3.2.2 shows that 111p, 151p and 555p are indistinguishable in this regard. These reactions were carried out at 0.05 M NaCl, which is optimal for NA binding\textsuperscript{160}. Such complexes would also be a substrate for the strand exchange phase of NA chaperone activity.

Therefore, we compared the activities of the proteins in both annealing and strand exchange with a Förster resonance energy transfer (FRET) based chaperone assay as described in the Methods. We used the relative amounts of protein and NAs indicated by the grey rectangle in Fig. 3.2.2A. As shown in the Fig. 3.2.3, the proteins showed similar activities in both phases of this assay. Thus, neither binding properties nor the nucleic acid chaperone activity account for the inactivity of 151p.

The lengths of the oligonucleotides used in the foregoing assays could only accommodate a single trimer. But, as noted above, 1.0 mM EGS can cross-link multimers of trimers (ex: trimer\textsubscript{2}, trimer\textsubscript{3}) that can assemble on longer oligonucleotides. To determine whether 151p can be differentiated from 111p and 555p on this basis we compared the distribution of trimer and multimers at two concentrations of the proteins on 0.125 μM of a 120-nt oligonucleotide. Fig. 3.2.4 show that the distribution of trimer and multimers is similar for all three proteins at both concentrations. In addition, 0.5 M NaCl essentially inhibited polymer formation by the three proteins as shown previously\textsuperscript{160}. And finally, the inset in Fig. 3.2.4 shows that partial cross-linking (0.05 mM EGS) of the three proteins in the absence of NA at 0.05 M NaCl generated similar monomer\textsubscript{n} ladders when electrophoresed under denaturing conditions. These lanes were obtained from the same gel from which the intervening non-relevant lanes were cropped. The slower migration of the monomer, dimer, and trimer bands in the 111p lane is due to its location at the gel edge where migration is retarded by distortion of the electrophoretic field.
Figure 3.2.4: Trimer oligomerization on 125-nt ssDNA. Silver stained polyacrylamide gel of half of the indicated reactions electrophoresed under reducing and denaturing conditions. Except for lane 4 and the insert, the indicated amounts of ORF1p trimer were incubated in 15 μl with either 0.125 μM of a single stranded-oligonucleotide in 50 mM NaCl or in 0.5 M NaCl for 20 min at RT and then cross-linked for 30 min with 1.0 mM of the bifunctional EGS reagent and then processed for electrophoresis as described in the Methods. The protein in lane 4 and those in the insert, at 0.189 μM, were incubated with 0.015 mM EGS, which partially cross-links the trimer or polymers thereof giving a ladder of products on denaturing gels equivalent in size to monomer – indicated by the numbers to the left of lane 4 and the insert. Each lane of the insert contains the entire sample. Uniform background was reduced by using the ImageJ rolling ball tool which does not alter the contrast of the image (B) Densitometry of the electropherogram. Traces of the indicated lanes were generated using the ImageJ gel profile tool. The band corresponding to the oligonucleotide (120-nt) is indicted with an arrow. The positions of the monomeric and trimeric ORF1p species are indicated.
Complete cross-linking (1.0 mM EGS) under these conditions generates polymers too large to enter the gel \(^{160}\). Thus, none of the foregoing assays were sensitive to the biochemical defect of 151p that renders it inactive in retrotransposition. Therefore, we employed an assay capable of extending analysis of protein/NA interactions beyond those possible using bulk assays with oligonucleotide substrates \(^{160,196}\).

**Single-molecule measurements of ORF1p-ssDNA binding kinetics**

In this assay a double-stranded (ds) \(\lambda\) phage DNA molecule is tethered between two polystyrene beads. One is held in an optical trap while the other is immobilized on a micropipette tip \(^{59}\). The force-extension profile of dsDNA is obtained by gradually moving the fixed bead while recording the extension and the force exerted on the single DNA molecule (Fig. 3.2.5A-F, solid black line) without DNA-binding ligands. The plateau, or the rapid increase of the extension in the narrow 60–65 pN force regime, represents force-induced melting, wherein dsDNA is converted to ssDNA, by either peeling from the ends or forming melted bubbles \(^{197}\). Although the plateau may represent transition from B-form dsDNA into another form referred to as S-DNA \(^{113}\) in high salt conditions (>0.15 M), it is well established that at pH 7.5 and 0.05 M Na\(^+\) (the conditions used here) the plateau represents force-induced melting \(^{59,198}\). If the ssDNA strands could not anneal, the return curve would be the same as the ssDNA curve (Fig. 3.2.5A, purple line). However, here, in the absence of ligands, the DNA strands rapidly anneal indicated by the almost complete reversal of the dsDNA extension curve, which exhibits minimal hysteresis – *i.e.*, the discrepancy between the dsDNA-stretch and return curves (Fig. 3.2.5A, dashed black line). Ligands that bind to ssDNA could inhibit annealing, and by preventing dsDNA formation result in increased DNA length at forces below the melting plateau. This provides an accurate quantitative measure for the ligand-bound ssDNA fraction and was used previously to
characterize the binding of viral restriction factor APOBEC3G to ssDNA. Here we adapted this method to accommodate the more complex binding characteristics of ORF1p to ssDNA.

Figure 3.2.5: Single molecule analysis reveals and quantifies three ssDNA-bound ORF1p populations. (A) Stretch (solid black line) and return (dashed black line) curves of a dsDNA molecule in the absence of ORF1p. Purple line is the FEC of an ssDNA. (B) Return of the ORF1p-DNA complex (blue circles) after incubating an overstretched dsDNA for 360 s in 2 nM 111p. (C) Subsequent stretch (red circles) of the 111p-DNA complex shown in (C). (D) Quantifying the 111p-bound ssDNA fraction (f) bound by ORF1p exhibiting slow dissociation kinetics. The subsequent stretch is fit (red line, f=slow) to a linear combination of dsDNA (grey line, f=0) and the 111p-saturated ssDNA (solid gold, f=1) curves. (E) Quantification of the total fraction (fT) of 111p-bound ssDNA. Dashed blue curve is the linear combination intersecting the force (F0) at which the 111p-DNA complex begins to approach the force regime below the melting plateau, which yields fT. Return data is fit by allowing fast to vary with force in order to correct for the rapid 111p dissociation during the return to find ffast and fint. (F) Summary of the analysis method. The green dashed line is the summed linear combination of fslow and fint. (G) Schematic of the model used to interpret the data from (A-F), representing a single event from each kinetic class of ssDNA-bound ORF1p (not to scale and does not reflect the total fraction in each state). See methods for more details.
To analyze ORF1p-ssDNA interactions we first overstretch the dsDNA up to ~75 pN, converting most of the dsDNA into ssDNA. We then replace the surrounding solution with 2 nM retrotransposition competent 111p or 555p, or the defective 151p.

After incubating the ssDNA with ORF1p for 360 s, we release the force on the DNA and obtain a return curve of the DNA-ORF1p complex (Figs. 3.2.5G, B, blue circles) that reflects the fraction of protein-bound ssDNA. We generate a second curve (Fig. 3.2.5C, red circles) by re-stretching the DNA-ORF1p complex immediately after the initial return. The extension produced by the subsequent stretch is substantially smaller than the extension observed during the initial return (Fig. 3.2.5B, blue circles) but larger than that of dsDNA (Figs. 3.2.5A-F, solid black line). This indicates that while a population of ssDNA-bound ORF1p dissociated during the first return allowing concomitant duplex formation, another population still remains bound, suggesting multiple dissociation kinetics for ORF1p. Further subsequent stretches trace the first subsequent stretch (data not shown for clarity), indicating a ssDNA-bound ORF1p population with negligible dissociation.

At any given force below the melting plateau, the extension of the DNA attained during either the return after incubation or the subsequent stretch represents the sum of dsDNA and ORF1p-bound ssDNA fractions. One can quantify the ORF1p-bound ssDNA fraction (f), by modeling the data as a linear combination (Fig. 3.2.5) of the curves for dsDNA, with no protein bound, f = 0 (Figs. 3.2.5D-F, grey line), and ORF1p-saturated ssDNA curves, for which all the possible binding sites of ssDNA are stably occupied by ORF1p, f = 1 (Figs. 3.2.5D-F, gold line).

The ORF1p-saturated ssDNA is experimentally achieved by incubating the overstretched dsDNA in high (>15 nM) ORF1p concentrations for long times (>30 mins). However, assuming a simple
Linear combination of two kinetic states (slow and fast) yielded a poor fit due to the concave nature of the return curve, an effect that decreases with incubation time (see below and Fig. 3.2.5B). Therefore, we postulated three ORF1p-bound ssDNA fractions, each populated with ORF1p exhibiting distinct dissociation kinetics: fast ($f_{\text{fast}}$), intermediate ($f_{\text{int}}$) and slow ($f_{\text{slow}}$). We hypothesized that the fast fraction is due to the binding of single ORF1p trimers to ssDNA, which equilibrate quickly, exhibiting rapid bimolecular association and dissociation. During the return at forces below that of the melting plateau, rapid dissociation of ORF1p trimers is accompanied by concomitant duplex formation (Fig. 3.2.5G).

![Figure 3.2.6: Representative return curves upon different incubation times with the specified ORF1p variants.](image)

Continuous net dissociation of ORF1p from this fraction during the return would account for the concave nature of the return data (Fig. 3.2.5B). The intermediate fraction could consist of
intermediate-sized ORF1p oligomers (trimer$_2$, trimer$_3$, trimer$_n$), which dissociate when the DNA is completely relaxed. The remaining fraction is bound by ORF1p that does not dissociate from ssDNA during the timescale of our experiments, which we assume constitutes $f_{\text{slow}}$ and likely consists of large protein polymers or aggregates, similar to those observed for APOBEC3G$^{199}$.

We directly obtain $f_{\text{slow}}$ by modeling the subsequent stretch as a linear combination of dsDNA and ORF1p-saturated ssDNA curves (Fig. 3.2.5D, red line and Fig. 3.2.6, bottom panel). Next, we find the linear combination that intersects $F_0$, where $F_0$ is the force at which the return curve begins to approach the force regime below the melting plateau (Fig. 3.2.5F). This linear combination yields the instantaneous total fraction ($f_T = f_{\text{fast}} + f_{\text{int}} + f_{\text{slow}}$) of ORF1p-bound ssDNA that emerged during the incubation (Fig. 3.2.5E, dashed blue line).

![Figure 3.2.7: Direct single molecule measurements of fast and intermediate dissociation time constants.](image)

(A) Representative data for measuring protein dissociation at constant force. Overstretched dsDNA (solid black) is incubated in 2 nM 111p for 360 s. The returning 111p-DNA complex after incubation (open blue circles) is stopped and maintained at a constant force of 43 pN via a force feedback loop for 360 s. (B) The change in extension with time during the constant force feedback loop (Green circles) is fit to a double exponential function of time (solid black). Two time constants $\tau_{\text{int}}$ and $\tau_{\text{fast}}$ represent the characteristic dissociation time constants of the ssDNA-bound 111p populations exhibiting respectively intermediate and fast dissociation kinetics. (C) Variation of the time constants with stopped force for ORF1p variants. Intermediate (circles) and fast (triangles) dissociation time constants are measured as a function of a stopped force ($F_\text{st}$). Green, blue and red data points correspond to: 111p, 555p and 151p respectively. Overall variant averages are fit to an exponential function of force, $\tau(F) = \tau_0 e^{F/\Omega}$, using the minimization of $\chi^2$ method (dashed grey lines), where $\Omega$ is a factor that describes the scale of time constant variation with force. Corresponding fits for $\tau$ and $\tau_i$ yield $\tau_0,\text{fast} = 2.7 \pm 0.4$ s, $\Omega,\text{fast} = 28 \pm 1$ pN and $\tau_0,\text{int} = 57 \pm 4$ s, $\Omega,\text{int} = 75 \pm 10$ pN respectively. Error bars are standard errors for at least three measurements.
The discrepancy between the dashed blue line and the observed return data (Fig. 3.2.5E, blue circles) is due to continuous duplex formation driven by the fast dissociating ssDNA-bound ORF1p, which continuously decreases the fast fraction, $f_{fast}$ during the return. To account for the fast dissociating protein we modify $f_{fast}$ to be varied with a phenomenological force (F) dependence ($\tilde{f}_{fast}(F) = \gamma F^3$, where $\gamma$ is a constant) and model the return curve after incubation (Fig. 3.2.5E, blue line). Collectively our analytical method quantifies the three ORF1p-bound ssDNA fractions $f_{slow}$, $f_{int}$ and $f_{fast}$, in agreement with the observed data. The summary of the analysis and the hypothetical curve that represents the combined fractions of $f_{slow}$ and $f_{int}$ (dashed green line), are shown in Fig. 3.2.5F. Because $F_0$ is determined directly from the data, there are a total of three fitting parameters for all of the data, and each curve is fit to one parameter (see Fig. 3.2.6).

**ssDNA-bound ORF1p exhibits three distinct kinetic states**

We tested the hypothesis of three ssDNA-bound ORF1p populations by directly measuring the dissociation times of ssDNA-bound ORF1p variants. After incubating overstretched dsDNA in 2 nM ORF1p, we stopped the return of the DNA-ORF1p complex at different forces ($F_{at}$=33, 43, and 53 pN) below the melting plateau, which we maintained for 360 s (Fig. 3.2.7A). To maintain constant force the DNA-protein complex was gradually released to compensate for any increase in force caused by duplex formation that would accompany dissociation of ssDNA-bound ORF1p. Hence the temporal decrease in the extension directly measures the net dissociation of the ssDNA-bound ORF1p as a function of time. During the constant force feedback loop the extension reaches a local equilibrium, converging at the subsequent stretch (Fig. 3.2.8A, red circles). This is greater than the extension of dsDNA (Fig. 3.2.7A, black line) at the
corresponding force, and results from the stably bound fraction of ORF1p oligomers that exhibit negligible or slow dissociation from ssDNA ($f_{\text{slow}}$).

![Diagram](image_url)

**Figure 3.2.8:** Single molecule measurements to quantify slow, intermediate and fast ORF1p-bound ssDNA fractions as a function of incubation time. (A) Representative data (open circles) and corresponding fits (solid lines) for returns of ssDNA-ORF1p complexes after incubating an overstretched dsDNA (ssDNA) in 2 nM ORF1p variants for 180 s (pink) and 2880 s (blue). Data are fit to a linear combination of dsDNA (grey line) and ORF1p-saturated ssDNA curves (gold line). (B) ssDNA fractions bound by ORF1p exhibiting slow (blue), intermediate (green), fast (red) dissociation kinetics and the total ORF1p-bound fractions (purple) are quantified as a function of incubation time. The fits of slow, intermediate and total fractions (blue, green, and purple lines, respectively) are single exponential functions of incubation time and the fast fractions are fit (red lines) to a sum of increasing and decaying exponential functions of incubation time. Error bars are standard errors for at least three measurements.

The change in the extension during the constant force feedback loop yielded an exponential function in time with two-time constants ($\tau_{\text{fast}}$ and $\tau_{\text{int}}$) that differed by an order of magnitude (Figs. 3.2.5B-C), in agreement with the fast ($f_{\text{fast}}$) and intermediate ($f_{\text{int}}$) ssDNA- ORF1p kinetic states. Dissociation time constants $\tau_{\text{fast}}$ and $\tau_{\text{int}}$ were similar for the three ORF1p variants at each
stopped force $F_{st}$ (Fig. 3.2.8C). The overall dissociation time constants of all three ORF1p variants averaged over $F_{st}$ are fit to an exponential function of force (Fig. 3.2.7C, dashed grey lines). The corresponding fits yield $\tau_{0, fast} = 2.7 \pm 0.4$ s and $\tau_{0, int} = 57 \pm 4$ s, the zero force fast and intermediate dissociation time constants, respectively. The $\tau_{0, fast}$ results are consistent with the behavior of 111p, 555p, and 151p in the oligonucleotide-based assays shown in Figs 3.2.2-3, as the oligonucleotides used in these assays can only accommodate a single trimer.

**Quantifying oligomerization rates of ORF1p variants on ssDNA**

In order to quantify the three ssDNA-bound ORF1p populations ($f_{slow}(t)$, $f_{int}(t)$ and $f_{fast}(t)$) as a function of incubation time ($t$), we repeated the first experiment by incubating the overstretched dsDNA for different durations in 2 nM ORF1p (Fig. 3.2.8A). For the three ORF1p variants, $f_{slow}(t)$ and $f_{int}(t)$ increase with incubation time and reach saturation following a simple exponential function (Fig. 3.2.8B, blue and green lines). However, $f_{fast}(t)$ Taken together, the results indicate that $f_{fast}(t)$, the ssDNA fraction bound by rapidly dissociating ORF1p, is converted with time to the more stable ssDNA-ORF1p fractions $f_{int}$ and $f_{slow}$. This result supports the hypothesis that, as was the case for APOBEC3G, ORF1p trimers oligomerize to the more stably bound oligomeric forms on ssDNA with increasing incubation time. Therefore, the decay time of $f_{fast}(t)$. To investigate 111p, 555p, and 151p were 356 ± 2 s, 1620 ± 124 s and 18000 ± 8190 s respectively. Thus, the $T_{oligo}$ of 151p, which is completely inactive in retrotransposition, is one to two orders of magnitude higher than both 111p and 555p, which support retrotransposition. This inverse correlation between $T_{oligo}$ and retrotransposition efficiency strongly suggests that a rapid rate of ORF1p oligomerization is essential for retrotransposition.
Discussion
Despite the prominent role of L1 retrotransposons in shaping mammalian genomes and the persistence of L1 activity in most mammals, including humans, we have little mechanistic understanding of the evolutionary and biochemical processes that underlie the success of L1 elements. This is particularly true for ORF1p, a uniquely trimeric NA chaperone that is essential for retrotransposition, and which frequently undergoes major evolutionary changes. Here we present two observations that advance our understanding of its biochemical properties that are relevant to its role in retrotransposition and to the function and evolutionary dynamics of its coiled coil domain.

To address the biochemical consequences of coiled coil evolution we resuscitated the ancestral L1Pa5 ORF1 (encodes 555p), which differs from the currently active L1Pa1 ORF1 (encodes 111p) at 42 positions, 30 of which were located in the coiled coil domain (Fig. 3.2.1D). To assess the effect of these coiled coil substitutions on ORF1p function in retrotransposition, we exchanged different sets of the modern coiled coil amino acids for their ancestral counterparts to generate mosaic ORF1 sequences. We then inserted either the resuscitated L1Pa5 or mosaic ORF1 sequences in place of the modern L1Pa1 ORF1 sequence of a retrotransposition reporter vector and found that whereas 555p and 111p support retrotransposition equally, the mosaic ORF1 sequence (151, encodes 151p) is completely inactive. To determine the biochemical basis for this inactivity we purified and compared the in vitro properties of insect-expressed 111p, 151p and 555p.

Interactions with nucleic acids
All of the corresponding purified proteins were similarly active in their interactions with NAs with respect to binding affinity, stabilization of mismatched duplex, and NA chaperone activity.
when tested with oligonucleotide substrates (≤ 29-nt) traditionally used to assess these activities (Figs 3.2.2 and 3.2.3). These substrates are only long enough to accommodate one ORF1p trimer. Additionally, cross-linking studies showed no differences between the proteins in their assembly of the cross-linkable short oligomers, trimer2 and trimer3, on a 120-mer oligonucleotide. Therefore, these assays were not sensitive to the biochemical defect of 151p.

However, single molecule stretching experiments with λ-DNA did reveal a defect in 151p; namely, this protein polymerized to stably bound oligomers on ssDNA at < 1/10th the rate of retrotransposition-proficient 111p and 555p (Fig. 3.2.7). These experiments identified three populations of ssDNA-bound ORF1p with distinct dissociation timescales: fast (\(f_{\text{fast}}\), seconds), intermediate (\(f_{\text{int}}\), tens of seconds) and slow or negligible (\(f_{\text{slow}}\), Figs. 3.2.5 and Fig. 3.2.8). The \(f_{\text{fast}}\) kinetic state is consistent with association/dissociation of trimers; \(f_{\text{int}}\), or intermediate kinetic state, is populated by more stably bound oligomers of trimers; the negligibly dissociating \(f_{\text{slow}}\) population is composed of presumably large polymers or aggregates\(^{199}\). The dissociation time constants for the populations \(f_{\text{fast}}\) (\(\tau_{\text{fast}}\), fitted value, 2.7±0.4 s) and \(f_{\text{int}}\) (\(\tau_{\text{int}}\), fitted value, 57 ± 4 s) were similar for the three proteins (Fig. 3.2.7C). Therefore, 151p is defective only in the conversion rate of DNA-bound trimers to the stably bound \(f_{\text{int}}\) and \(f_{\text{slow}}\) oligomers. Thus, retrotransposition requires fast conversion of NA bound trimers to more stably bound oligomers.

**A role for the coiled coil in ORF1p oligomerization rate on NAs**

The carboxy-terminal half of ORF1p mediates the inter-trimer interactions responsible for oligomerization\(^{160}\). Fig. 3.2.7 shows that the amino acid substitutions in the 151p coiled coil resulted in a reduced formation rate of stably bound oligomers. This finding indicates that the sequence of the coiled coil can determine the intra-trimer configuration that is conducive to oligomerization, an idea consistent with our conclusion that 0.5 M NaCl inhibits ORF1p
oligomerization through its structural effect on the coiled coil\textsuperscript{150}. Whether all coiled coil mutations that inactivate or strongly reduce retrotransposition, such as the single substitution in the mouse coiled coil\textsuperscript{200}, do so by retarding rapid conversion of ORF1p to stably bound oligomers remains to be determined. However, others showed that the L\textsubscript{93,100,107,113}V set of coiled coil substitutions, which strongly inhibit retrotransposition, decrease the amount of ORF1p incorporated into L1RNP\textsuperscript{201}. This could be explained by a decreased rate of ORF1p oligomerization that we demonstrated here with purified 151p. Notably the substitution at L\textsubscript{107} corresponds to the location of one of the ancestral replacements in 151p.

Because of its length and sensitivity to inactivating mutations, the coiled coil presents a relatively large target for deleterious mutations. Additionally, given the extensive intra- and inter-strand crosstalk within this motif\textsuperscript{202}, the occurrence of compensatory substitutions or other changes that ameliorate rather than reverse the original mutation would not be infrequent. Taken together, these factors could account for the periodic concerted changes in the coiled coil that typify ORF1p evolution.

**ORF1p oligomerization kinetic classes and the L1RNP**

Not only did the single molecule stretching assay provide a biochemical explanation for the inactivity of 151p in retrotransposition but it also revealed aspects of ORF1p/NA interaction that seem directly relevant to the L1RNP retrotransposition intermediate: namely its formation upon translation from the L1 transcript, its stability, and its ultimate dissolution during TPRT, necessary to expose the template for cDNA synthesis and the eventual generation of a new L1 insert.

As ORF1p is a non-specific NA binding protein, cis preference would limit non-productive interactions with non-L1 NAs. The rapid conversion of the fast dissociating population to stably-
bound \( f_{\text{int}} \) and \( f_{\text{slow}} \) would contribute to \( cis \) preference\(^{27,202}\) by preventing the newly synthesized ORF1p from diffusing from its encoding transcript\(^{31,160}\). In addition, its rapid oligomerization to more stably bound ORF1p-RNA complexes would protect the L1 transcript from degradative nucleases or components of the innate immunity pathways such as siRNA and the APOBEC3 family of enzymes\(^ {203,204}\). The fact that the \( f_{\text{fast}} \) population is transient suggests that this kinetic class likely has no role in retrotransposition beyond L1RNP formation.

However, both \( f_{\text{int}} \) and \( f_{\text{slow}} \), which increase for ~20 min at the expense of \( f_{\text{fast}} \), persist unchanged at close to equal amounts for time of the experiment (~50 min). We do not know the molecular distinction between the \( f_{\text{int}} \) and \( f_{\text{slow}} \) populations of ORF1p. However, the \( f_{\text{int}} \) population exhibits several features that would be critical for the L1RNP: they are bound strongly enough to form a stable L1RNP but can be driven off the single strand under force conditions that promote double strand formation. This latter reaction could be considered a proxy for the double-strandedness that would result from cDNA synthesis during TPRT. The dissolution of the L1RNP would make the L1 transcript accessible as a cDNA template and the released trimer NA chaperones could facilitate the various other NA interactions at the integration site that are required to complete the L1 insertion event\(^ {205}\).

**Materials and Methods**

The structures of all of the recombinant DNA clones and ORF1 constructs described here were verified by DNA sequencing.

**ORF1 constructs**

111 - The modern ORF1 (111) sequence was obtained from the highly active L1.3 element\(^ {206}\), a member of the Ta-1 subfamily of the human-specific L1Pa1 family\(^ {147}\) and kindly provided to us by Dr. John Moran on the JCC8 plasmid\(^ {147}\). We subcloned ORF1 as a NotI/Xmal fragment into
MB18 (pUC18 modified to contain BamHI, NotI, XmaI and MluI sites between SphI and EcoRI) to generate p111-mb18. The encoded protein is 111p, the corresponding retrotransposition reporter is p111-rtc (see retrotransposition assays, below).

555 – We resuscitated the ancestral ORF1 (555) sequence of the 25 Myr old extinct L1Pa5 family by first constructing a 60% majority consensus sequence from 840 L1Pa5 ORF1 sequences in the human genome data base (UCSC hg18, NCBI Build 36.1). We converted TpGs to CpG if the alignment contained CpG or CpA at this position, and likewise for any CpA if the alignment contained CpG or TpG. Ambiguities were resolved by reference to the antecedent L1Pa6 or descendant L1Pa4 families. We synthesized the sequence that included the NotI site just 5′ of the ORF1 ATG through the highly conserved BsmFI site (incises the codon for K154, Supplementary Fig. S1) by ligating overlapping, complementary gel-purified 50-nt oligonucleotides. The sequence of the BsmF1 site through XmaI downstream of the TAA was synthesized by Retrogen, Inc. (San Diego, CA). We ligated these sequences (NotI/BsmFI, BsmFI/XmaI) into the NotI/XmaI sites of MB18 to generate 555-mb18. The base sequences between NotI/ATG, and TAA/XmaI are the same in p111-mb18 and p555-mb18. The encoded protein is 555p, and the retrotransposition reporter is p555-rtc (see retrotransposition assays, below).

151 – We generated the mosaic ORF1 sequence (151) by inserting the BsmI/BsmFI fragment of 555-mb18 (Fig. 1B) in the corresponding sites of 111-mb18 to generate 151-mb18. BsmI incises the DNA after the first base, A, for either the R codon in 111-mb18 or the T codon in 555-mb18. Therefore this ligation preserves the ancestral T in the mosaic 151p as shown in Supplementary Fig. S1. The encoded protein is 151p, the retrotransposition reporter is 151-rtc (see retrotransposition assays, below).
Expression vectors

FLAG-tagged 111p, 151p and 555p expression vectors – To compare expression of 111p, 151p, and 555p in HeLa_JM cells we fused the FLAG epitope (DYKDDDDK) to their C-termini. We used PCR with the Phusion® High-Fidelity DNA Polymerase (New England Biolabs) and oligonucleotides F-BamHI-111 and R-Flag-EcorI-111 to recover the ORF1 sequences from templates p111-rtc or p151-mb18, and oligonucleotides F-BamHI-555 and R-Flag-EcorI-555 for p555-mb18 (, #8-11). The 5′ BamHI-Kozak and FLAG-EcoRI-3′ bounded amplicons were cloned into the corresponding sites in pcDNA3.1(+)-puro obtained from the Don Ganem laboratory, UCSF.

ORF1p expression and purification

As described previously for the expression and purification of 111p we inserted the coding sequences for a hexa-histidine (HIS) affinity tag and tobacco etch virus (TEV) protease site (generated by annealing oligonucleotides histev_t and histev_b, into the BsiWI and NotI sites of p555-mb18 or p151-mb18 (i.e., HIS-TEV-ORF1) followed by deletion of the NotI site just 5′ of ORF1. Cloning of the fusions into the pFastBac1 vector (Invitrogen) for baculovirus infection, generation of the respective recombinant baculoviruses using the Bac-to-Bac® based baculovirus expression system (Invitrogen), and infection of Hi5 insect cells, were carried out by the Protein Expression Laboratory, Advanced Technology Program, SAIC-Frederick, as previously described. We purified the proteins from the insect cell pellets as described. The amino termini of all of the ORF1p proteins contain an N terminal glycine followed by the normal initiating methionine that resulted from TEV cleavage step.

Retrotransposition assays
We used HeLa cells (HeLa JM, kindly provided by Dr. John Moran, Univ. Michigan, Ann Arbor) and a retrotransposition reporter, pRTC2, an L1.3-containing version of the retrotransposition reporter plasmid\textsuperscript{160} that we extensively modified as described in detail in the supporting information that accompanies Cook et al.\textsuperscript{34} except that the version used here lacks the puromycin N-acetyl-transferase gene (Fig. IC). HeLa cells were plated in a 6-well dish at 2 x 10\textsuperscript{5} cells/well and within 24 h a mixture of 3 μl FuGene\textsuperscript{\textregistered}6 Transfection Reagent (Roche) and 1 μg of p111rtc, p151rtc, or p555rtc were applied to the cells (per the supplier’s suggestions). After 3 days, 400 μg/ml G418 antibiotic (Gibco) was added to select for G418 resistant foci, and incubation was continued for an additional 10-14 days with media change as needed. The cells were then washed twice with 1X PBS, fixed (2% formaldehyde, 0.2% glutaraldehyde in 1X PBS), and stained with Karyo Max\textsuperscript{\textregistered} Giemsa Stain (Gibco). After 30 min the stain was removed, and the cells were washed repeatedly with 1x PBS until the background was colorless.

**Filter-binding assay**

We employed a previously described modification of a dual membrane filter-binding assay\textsuperscript{166}. Binding reactions (20 mM Tris, 10% glycerol (w/v), 0.5 mM MgCl2, 0.1 mM DTT, and 100 μg/ml BSA, on ice) contained 0.1 nM pre-formed mismatched duplex of d29 and [\textsuperscript{32}P]-d29_c-mm. After adding 1/10th volume of the appropriately diluted ORF1p in 0.5 M NaCl (final [NaCl] = 0.05 M), the reaction was incubated at 37 °C for 1 h whereupon duplicate 5 μl samples were added under vacuum to the membrane layer of nitro-cellulose (binds protein-bound DNA) atop of zeta-probe GT (Bio-Rad, binds free DNA). We washed the membrane stack with binding buffer and determined the radioactivity of the dried membranes using a Fuji FLA-5000 series Image Analyzer (Fuji Medical Systems) and Image Gauge software (version 3.0, Fuji Medical Systems). We fit the fraction bound (FB) [radioactivity bound to nitrocellulose / (radioactivity
bound to nitrocellulose + radioactivity bound to zeta probe) as a function of protein concentration using a logistic function (KaleidaGraph, 4.1): \( y \ (FB) = m_1 + (m_2 - m_1)/(1+(X/m_3)m_4) \), where \( X = [\text{ORF1p}] \); \( m_1 = [\text{ORF1p}] \) at maximal ligand bound; \( m_2 = \) fraction ligand bound at 0 [ORF1p], \( m_3 = [\text{ORF1p}] \) where half of the ligand is bound, i.e., \([\text{ORF1p0.5FB}]\); \( m_4 = \) slope of the binding curve plotted as a function of the natural log of [ORF1p] \(^{160}\). We also determined the fraction of ss- and ds-[\( ^{32} \)P]DNA in some samples: 10 µl sample were diluted two-fold with ice-cold 0.2% SDS, 3% glycerol, 400 ng/µl tRNA, bromophenol blue followed by electrophoresis at 4 °C (6% 29:1 cross-linked polyacrylamide gels, 20 mM HEPES, 10 mM sodium acetate, 200 V). Dried gels were scanned with the Fuji Image Analyzer. A portion of the reaction without ORF1p was kept at 0 °C to serve as a control for melting at 37 °C in the absence of protein. The fraction ssDNA = \([\text{ssDNA} / (\text{ssDNA} + \text{dsDNA})] - (\text{fraction ssDNA at 0 protein})\).

**FRET NA chaperone assay**

This assay was carried out as described in ref.\(^{208}\) using the relative concentrations of ORF1p and NA wherein a mismatched duplex is stabilized (caged), before being eventually melted (Fig. 2A, grey rectangle, and ref.\(^{209}\)). Reactions (20 µl) initially contained 1 µM ORF1p (in terms of monomer), 10 nM 5'-Cy3-ACTGCCAGAGAcTTcCCACAT and 2x FRET buffer (100 mM Tris-Cl, pH 7.4, 100 mM NaCl, 6 mM MgCl2, and 4 mg/ml BSA), incubated at room temperature for ~5 min in a 96-well half-area, low binding black plate (Corning #3993). Annealing was started by injection of 20 µl 10 nM Cy5-ATGTGGAAAATCTCTAGCAGT in water for a final concentration of 500 nM ORF1p and 5 nM of each RNA oligonucleotide in 1x FRET buffer. Cy3 was excited every 0.7 seconds at 535 nm (25 nm band pass), and emissions were read at 590/25 (Cy3) and 680/30 (Cy5). At ~3 min, we added 5 µl of 500 nM of the perfect complement
(, #5, 21r_dna) of Cy5- ATGTGGAAAATCTCTAGCAGT (, #7) and the reading was continued for another 3 min. We used a Synergy2 Microplate Reader running Gen5 data analysis software (Biotek Instruments, Inc.), which had been fitted with a red photomultiplier tube and a xenon flash lamp.

**Protein cross-linking**

As described previously\(^{160}\), 10–20 µL of an ORF1p solution in 20 mM Hepes, pH 7.5, 10% w/v glycerol, 0.05 M NaCl was incubated for 30 min at room temperature in the presence or absence of a 120-nt single-stranded oligonucleotide DNA (, #1, d120_c), and then for an additional 30 min with either 0.05 mM or 1 mM ethylene glycol bis (succinimidylsuccinate) (EGS, Pierce Biotechnology) freshly made in dimethyl sulfoxide (DMSO, Sigma). These respective concentrations partially or completely cross-link the protein. The concentrations of protein, oligonucleotide and EGS are given in the legend to Fig. 3.2.4. The reactions were quenched with one-tenth volume of 1 M Tris-Cl, pH 8.0, subjected to denaturing SDS-PAGE on either Bis-Tris 10% or 4-12% gradient gels (Novex Life Technologies). The protein bands were visualized with silver stain (Pierce).

**Western blots**

Each well of a six-well plate was plated with 6 x 10⁵ HeLa cells, and after 24 h transfected with 1 µg pORF1-Flag vector using 3 µL FuGENE®6 (Promega). After 48 h the cells were washed twice with cold PBS and lysed in cold 50 mM Tris-Cl, pH 7.4, 650 mM NaCl, 1 mM EDTA, 1% Triton X-100, complete EDTA-free protease inhibitor cocktail (Roche), 100 µM leupeptin, and 1 mM PMSF. Sonicated lysates were centrifuged (17,000 x g, 15 min, 4 °C) and 75 µg protein of supernatant protein (Bradford reagent – BioRad) was electrophoresed under denaturing conditions, transferred to nitrocellulose membranes using iBlot® (Invitrogen), which
were then incubated with ANTI-FLAG® M2 monoclonal antibody (Sigma-Aldrich®). Bands were detected with SuperSignal™ West Pico Chemiluminescent Substrate (Thermo Scientific) and exposed on HyBlot EST™ autoradiography film (Denville Scientific, Inc.). Blots were then treated with Restore™ PLUS Western Blot Stripping Buffer (Thermo Scientific) according to manufacturer’s protocol and re-probed with anti-γ-tubulin (Sigma-Aldrich®).

**Single molecule methods**
Each experiment was conducted on a biotinylated bacteriophage λ DNA molecule, tethered from its opposite ends between two streptavidin-coated polystyrene beads in 10 mM HEPES, 50 mM Na+ at pH 7.5. While one bead was fixed on a micropipette tip the other was held in an optical trap. By gradually moving the fixed bead, the force-extension curve of a DNA or DNA-ORF1p complex was obtained. An overstretched dsDNA anneals quickly with minimal hysteresis between the stretch and release curves in the absence of ORF1p. To quantify the ssDNA fractions bound by distinct ORF1p kinetic states, we first overstretched the dsDNA (Force ≈75 pN) to construct an array of ssDNA binding sites and then incubated it in 2 nM ORF1p variant (111p, 555p or 151p) for 360 s. After incubation we obtained the return curve of the resulting DNA-ORF1p complex (Figs. 3.2.5B, E, F and Fig. 3.2.9, blue circles). A second curve is obtained by subsequently stretching the DNA-ORF1p complex immediately after the initial return (Figs. 3.2.5C, D, F and Fig. 3.2.9, red circles). The experiments were conducted in 10 mM HEPES, 50 mM Na+ buffer solution at pH 7.5 with a pulling rate of 400 nm/s and repeated for incubation times, 180, 720, 1440 and 2880 s (Fig. 3.2.7A).

The DNA can be considered ORF1p-saturated when the hysteresis stops increasing with protein concentration or incubation time. The resulting complex is an array of ssDNA in which all the possible binding sites of ssDNA are occupied and stabilized by ORF1p and thus significantly
longer than dsDNA (Fig. 3.2.5A and Fig. 3.2.9, gold line). However, it is shorter than ssDNA
(Fig. 3.2.5A and Fig. 3.2.9, purple line) due to elasticity changes to the bound protein complex.
In unsaturated incubation conditions the resulting complex is a combination of dsDNA and
ORF1p-bound ssDNA sites in which duplex formation is inhibited. Thus, the increase in the
extension of the ORF1p-DNA complex in the force regime below the melting plateau compared
to the extension of a dsDNA in the absence of ORF1p, measures the fraction (f) of the ORF1p-
bound ssDNA. We postulated that the total fraction is bound by a combination of ORF1p
populations exhibiting slow (\(f_{\text{slow}}\)), intermediate (\(f_{\text{int}}\)), and fast (\(f_{\text{fast}}\)) dissociation kinetics. We
quantified these fractions with reference to the dsDNA in the absence of ORF1p (\(f=0\)) and
ORF1p-saturated ssDNA (\(f=1\)) curves, which represent the two extremes. As there is no current
theoretical model that provides the force-extension relationship of these protein-saturated ssDNA
complexes, we fit the saturated data to a second order polynomial using least squares
minimization to obtain the saturated curve (\(b_{\text{sat}}(F)\)) as a function of force, \(F\):

\[
b_{\text{sat}}(F) = \frac{-B + \sqrt{B^2 - 4AC - F}}{2A}
\]

The polynomial coefficients [A, B, C] for 111p, 555p and 151p were found to be [878.51, -489.43, 67.81], [920.38, -543.78, 82.424] and [971.88, -607.21, 98.637] respectively (Figs. 3.2.5-D-F, Fig. 3.2.8 and Fig. 3.2.9, gold line)

The extensible Worm-like chain (WLC) model was used to obtain the theoretical extension
\((b_{\text{sat}}(F))\) of a dsDNA molecule as a function of force (Figs. 3.2.5-D-F, Fig. 3.2.8 and Fig. 3.2.9,
grey line), where\(^{160}\)
\[ b_{ds}(F) = B_{ds} \left\{ 1 - \frac{1}{2} \left( \frac{k_B T}{F P_{ds}} \right)^{1/2} + \frac{F}{S_{ds}} \right\} \] (3.2.2)

The extensible freely-jointed chain (FJC) models the extension of an ssDNA \( b_{ss}(F) \) as a function of force (Fig. 3.2.5A and Fig. 3.2.9, purple line), where\(^6^3\)

\[ b_{ss}(F) = B_{ss} \left\{ \coth \left( \frac{2 P_{ss} F}{K_B T} \right) - \frac{1}{2} \frac{k_B T}{P_{ss} F} \right\} \left[ 1 + \frac{F}{S_{ss}} \right] \] (3.2.3)

The typical parameters for stretch moduli \( S_{ds} = 1361 \text{ pN}, S_{ss} = 720 \text{ pN} \), persistence lengths \( P_{ds} = 45 \text{ nm}, P_{ss} = 0.75 \text{ nm} \) and contour lengths \( B_{ds} = 0.34 \text{ nm/bp}, B_{ss} = 0.55 \text{ nm/bp} \) were used.

At any given force below the melting plateau ORF1p-DNA complex is a combination of dsDNA and ORF1p-bound ssDNA fractions. Thus, the extensions of the ORF1p-DNA complexes \( b(F) \) were modeled as a linear combination of \( b_{sat}(F) \) and \( b_{ds}(F) \), where,

\[ b(F) = (1 - f)b_{ds}(F) + f b_{sat}(F) \] (3.2.4)

to find the ORF1p-bound ssDNA fraction, \( f \). By definition, \( b_{ds}(F) \) and \( b_{sat}(F) \) yield the fits for which \( f \) is 0 and 1, respectively. The subsequent stretch yields the fraction \( f_{slow} \) of the ssDNA-bound ORF1p population that exhibits slow dissociation kinetics. We computed \( f_{slow} \) by fitting the subsequent stretch data to Eq. 3.2.4 (Figs. 3.2.5D, F, red line, Fig. 3.2.9, #1, bottom row).

Assuming that the ORF1p-bound ssDNA fraction that emerged during the incubation remains constant throughout the return after incubation, the return curve should trace the linear combination intersecting \( F_0 \) (Fig. 3.2.5E, dashed blue line, Fig. 3.2.9, #2), where \( F_0 \) is the force at which the return curve begins to approach the force regime below the melting plateau. Thus, using Eq. 3.2.4, we found the linear combination intersecting \( F_0 \) to determine the total ORF1p-bound ssDNA fraction, \( f_T \):
\[ f_T = f_{\text{slow}} + f_{\text{int}} + f_{\text{fast}} \] (3.2.5)

However, the discrepancy between the dashed blue line and the observed return data (Fig. 3.2.5E and Fig. 3.2.9, blue circles) is due to continuous duplex formation at a rate determined by the rapidly dissociating ssDNA-bound ORF1p, which decreases the fast fraction, \( f_{\text{fast}} \) during the return. In order to account for the fast dissociating protein, we modified \( f_{\text{fast}} \) to be varied with a phenomenological force dependence ( \( f_{\text{fast}} \to \tilde{f}_{\text{fast}}(F) \) ) and modeled the return data with a varying total fraction \( \tilde{f}_T(F) \):

\[
\tilde{f}_T(F) = f_{\text{slow}} + f_{\text{int}} + \tilde{f}_{\text{fast}}(F),
\]

where

\[
\tilde{f}_{\text{fast}}(F) = \gamma F^3
\] (3.2.6)

and \( \gamma \) is a constant. By definition

\[
f_{\text{fast}} = \tilde{f}_{\text{fast}}(F_o) = \gamma F_o^3
\] (3.2.7)

Substituting for \( f_{\text{fast}} \), Eq. 3.2.5 can be re-written as

\[
f_T = f_{\text{slow}} + f_{\text{int}} + \gamma F_o^3
\] (3.2.8)

to obtain \( f_{\text{int}} \) in terms of \( \gamma \), where

\[
f_{\text{int}} = f_T - f_{\text{slow}} - \gamma F_o^3
\] (3.2.9)

Substituting for \( f_{\text{int}} \) in Eq. 3.2.6, we obtained \( \tilde{f}_T(F) \), where

\[
\tilde{f}_T(F) = f_T - \gamma (F_o^3 - F^3)
\] (3.2.10)
Since the values for \( F_0 \) and \( f_T \) are already found, by substituting \( f \) of Eq. 3.2.4 with the expression in Eq. Fig. 3.2.8 and fitting the return after incubation data (Figs. 3.2.5E-F, solid blue line, Fig. 3.2.7A, Fig. 3.2.9, #3, top row), we determined \( \gamma \). We then found \( f_{\text{fast}} \) and \( f_{\text{int}} \) from Eq. 3.2.8 and 2.10, respectively. The dashed green line in Fig. 3.2.5F and Fig. 3.2.9(#4) represents the linear combination of the combined fractions \( f_{\text{int}} \) and \( f_{\text{slow}} \).

**Figure 3.2.9:** Single molecule method to quantify fractions of ssDNA-bound ORF1p populations. Grey and gold solid lines are theoretical fits for dsDNA (extensible WLC, Eq. 2.2) and ORF1p (111p)-saturated ssDNA (Eq. 2.1), respectively. Purple line represents the theoretical ssDNA curve obtained by the extensible FJC model (Eq. 2.3). Blue and red open circles are representative data of return after incubating in 2 nM 111p for 360 s and the subsequent stretch, respectively. Data modeling follows the numbered order as described in the Supplementary Methods to find \( f_{\text{slow}}, f_{\text{int}} \) and \( f_{\text{fast}} \). 1. Subsequent stretch data are fit as a linear combination of dsDNA and ORF1p-saturated ssDNA (Eq. 2.4) to find \( f_{\text{slow}} \) (red line) 2. Total ORF1p-bound ssDNA fraction (\( f_T \), Eq. 2.5) is found by the linear combination intersecting \( F_0 \) (dashed blue line), where \( F_0 \) is the force at which the return curve begins to approach the force regime below the melting plateau. 3. The return data are modeled (solid blue line) accounting for the rapidly dissociating ORF1p during the return to find \( f_{\text{fast}} \) and \( f_{\text{int}} \). 4. Dashed green line is the linear combination of the combined fractions \( f_{\text{slow}} \) and \( f_{\text{int}} \).

To directly measure the dissociation time constants, an overstretched dsDNA was incubated with ORF1p for 360 s. During the return after incubation, the ORF1p-DNA complex was stopped and maintained a constant force (\( F_{\text{st}} = 53, 43 \) or 33 pN) for 360 s, via a force feedback loop. The change in the extension during the constant force feedback was recorded every 50 ms and fit to a
double exponential function in time to obtain the characteristic dissociation time constants of the fast and intermediate ssDNA-bound ORF1p kinetic states.
3.3 Single molecule binding dynamics of ORF1p to preformed ssDNA

The following section describes the extended study of ORF1p-ssDNA binding dynamics. The work described here is currently in progress and unpublished.

Introduction

The methods described so far probe the properties of ORF1p’s nucleic acid chaperone activity and the ability to rapidly form stable conformers on ssDNA, both of which have been shown to play a critical role in L1 retrotransposition. Although the question of how exactly either of these properties support L1 activity remains untested to date, it is reasonable to assume while the chaperone properties play a primary role remodeling nucleic acid during TPRT, the oligomerization properties are primarily involved to mediate the packaging of the L1 transcript into an ORF2p containing ribonucleoprotein complex (L1 RNP). To further investigate ORF1p’s nucleic acid packaging properties, we designed experiments described in this section to investigate the conformational aspects of ORF1p-ssDNA binding dynamics. Here, we employ single molecule DNA stretching experiments with preformed ssDNA, instead of force-melted dsDNA that can reanneal at lower forces (<60 pN), as described in section 3.2. This method allows us to directly probe the binding dynamics of ORF1p to ssDNA at a wide range of forces while making it possible to characterize the supramolecular properties of the resulting ssDNA-ORF1p complex. The ORF1p-binding dynamics presented here provide compelling evidence of its ability to wrap ssDNA in multiple conformations. Furthermore, we show that ORF1p forms a highly structured nucleoprotein complex through DNA-mediated protein-protein interactions. This complex likely mimics the L1 RNP, thereby providing mechanistic insights on ORF1p function in L1 retrotransposition.
Results

**ORF1p binds ssDNA in a biphasic mechanism**

To probe the ORF1p-ssDNA binding dynamics, we generate an 8100 nt long ssDNA in situ (Fig. 3.3.1) in an optical tweezers system as described in the legend to the Fig. 3.3.1. Then we perform force-clamp experiments by holding and maintaining a constant force on the DNA, while incubating with ORF1p.

---

**Figure. 3.3.1: Experimental Procedure to probe ORF1p-ssDNA binding dynamics.** (A) Schematic depiction of the experimental method to probe ORF1p-ssDNA binding dynamics. 1. An ~8 kbp dsDNA that is recessed (30 nt) at its 3’ is tethered with its DIG and biotin labeled termini to anti-DIG and streptavidin functionalized beads, respectively. 2. The DNA molecule is held at ~ 50 pN in the presence of T7 DNA polymerase (T7DNAp) for exonucleolysis. 3. A long ssDNA molecule is generated after complete digestion of the bottom strand by T7DNAp. 4. The ssDNA is held at a constant force (30 pN) via a force feedback module and the surrounding solution is exchanged with a solution containing ORF1p. (B) Force extension curves corresponding to the steps described in (A). (C) Change in extension of a ssDNA upon binding to ORF1p as a function of time. ORF1p is introduced to the ssDNA at t=20 s, as denoted from the dashed line. A biphasic change in extension is observed where a rapid decrease is followed by a relatively slow increase in extension.
Thus, the temporal DNA extension change measures the binding dynamics of ORF1p to the preformed ssDNA. Interestingly, we observe a biphasic binding profile at saturating ORF1p concentrations (<3 nM), wherein a rapid decrease in ssDNA extension is followed by a relatively slow increase that equilibrates at an extension less than that of protein-free ssDNA. The biphasic binding mechanism suggests that the ORF1p binds ssDNA in at least with two distinct conformations. Initially the bound ORF1p is converted to a conformation that constricts DNA to a greater extent, which subsequently partially transforms to a protein-ssDNA conformation in which the DNA is less constricted. This process is presumably regulated through ORF1p occupancy on the ssDNA lattice.

**Interconversion of ORF1p binding modes are regulated through ssDNA occupancy**

To test the hypothesis that the ORF1p binding states are primarily regulated through the protein occupancy, we performed ‘concentration-switch’ experiments in which, after the initial biphasic binding extension is equilibrated, the solution surrounding the DNA molecule is switched between protein-free buffer and the protein solution and incubated for long time durations under each condition (Fig. 3.3.2). Surprisingly, after equilibration upon initial binding, subsequent elimination of the free protein in the solution further shortens the DNA. Typically, protein-dissociation in the absence of free proteins is expected to recover the free ssDNA extension (an increase in extension in this case). However, we observe the opposite effect, which is indicative of additional conversion events to the more constricted binding conformation that are concomitant with protein dissociation events. After shortening, the extension reaches an asymptote, indicating negligible protein-dissociation from this state. The additional shortening of the extension is reversed by subsequently switching the solution surrounding the DNA with the protein solution. Taken together, this result is consistent with a competitive binding mechanism,
where the ORF1p binding state interconverts between at least two distinct modes with different binding site sizes, depending on the ssDNA occupancy.

**Figure 3.3.2: Concentration-switch experiments at 30 pN.** The solution surrounding the ssDNA is exchanged between 50 nM protein solution and a protein-free buffer, essentially reducing the free protein concentration down to zero. Dashed lines indicate the time at which the solution is switched. The ORF1p-ssDNA complex is further shortened in the absence of free protein. The extension is recovered by re-introducing the protein solution. Thus, indicating interconversion between distinct protein-ssDNA states, regulated by protein-occupancy on the ssDNA lattice.

**Concentration-dependent interconversion of ORF1p binding modes**

To further investigate the interconversion dynamics, we repeated the concentration-switch experiments as a function of protein concentration (Fig. 3.3.3). To quantify the concentration-dependent interconversion we measure the observed extension changes (Fig. 3.3.3B-C) and the rates (Fig. 3.3.3B-D) associated with each transition. Both the extension changes associated with the first transient shortening ($\Delta X_1$), and the following lengthening that eventually equilibrates ($\Delta X_2$), decrease and asymptotes with increasing concentrations. The extension change observed subsequently upon rinsing out the free protein in the solution ($\Delta X_3$) equilibrates at the initial shortening length observed during the biphasic binding ($\Delta X_3 = \Delta X_1$), thus, indicating that the less-constricted ORF1p binding state becomes more stable as the occupancy of the protein increases in the ssDNA lattice. Moreover, the rates associated with the initial shortening transition ($k_1$), the second lengthening transition ($k_2$), and the subsequent shortening transition
upon rinsing out the free proteins in the solution ($k_3$) increase and saturate with increasing concentrations. The trends of $k_1$ and $k_2$ indicates that these processes are initially rate-limited by the concentration. However, the concentration-dependence of $k_3$ likely indicate the dissociation of high-order protein multimers that forms with increasing protein concentration.

**Figure 3.3.3: Concentration-dependent binding dynamics of ORF1p.** (A) Representative curves denoting the initial biphasic binding mechanism for different ORF1p concentrations. (B) Representative curve for a concentration-switch experiment as described in the text. To quantify the interconversion dynamics, we measure the extension changes ($\Delta X$) and rates ($k$) associated with each transition. (C) $\Delta X$ measured as shown in (B) as a function of protein concentration. (D) Observed transition rates ($k$) for transitions shown in (B).

**Force-dependent interconversion of ORF1p binding modes**

To probe the force dependence of the ORF1p-ssDNA binding dynamics, we repeated the concentration-switch experiments at 30 nM ORF1p as a function of force (Fig. 3.3.4). The extension-changes associated with all transitions decrease with increasing force. Therefore, the constricted ORF1p state becomes less stable with increasing force. Interestingly both $k_1$ and $k_2$ do not depend on force, suggesting that the ORF1p-binding mechanism is not rate-limited by the template-tension.
Secondary compaction mediated through protein-protein interactions

To probe the protein-protein interactions, we evaluate the force-extension cycles and the equilibrium extension changes associated with the ORF1p-ssDNA complex that are incubated at zero force to facilitate secondary protein-protein interactions, as described in the legend to the Fig. 3.3.5. To do this, we first incubate the ssDNA at 30 pN until equilibrium is achieved, and subsequently rinse out the free protein from the solution. Then the protein-DNA complex is brought back to zero-force and incubated for different time durations in buffer.

To allow maximal disruptions after incubation, the protein-DNA complex is stretched and held at 75 pN via a force-feedback module (Figure 3.3.5A-B). As the force is maintained at 75 pN, the extension of the protein-DNA complex increases and eventually equilibrates, indicating that no
further disruption events occur. The slope of the equilibrium force-extension curves of the resulting protein-DNA complex upon incubating at zero-force increases with incubation time, due to a significant reduction of the elasticity of the protein-DNA complex. At the highest incubation time the force-extension curve resembles a polymer with extremely large persistence length, indicating a high degree of secondary compaction due to protein-protein interactions. In addition, the disruptions of this highly compacted protein-DNA complex occur gradually as seen with the extension-time profiles (Fig. 3.3.5B) that are obtained at 75 pN, suggesting that the compacted protein-DNA complex is a well-ordered structure that is mediated by multiple protein-protein interactions. In addition, the disruption events decrease with increasing incubation time, resulting in a highly compacted nucleoprotein complex, that is even stable at 75 pN.

**Figure 3.3.5: Protein-mediated secondary compaction of ORF1p-ssDNA complex** (A) Experimental steps to probe the protein-mediated secondary compactions. A protein-free ssDNA is shown in black circles. (#1) ssDNA is stretched and incubated with 30 nM ORF1p. After the extension is equilibrated the protein solution is exchanged with clean buffer and the protein-DNA complex is brought back to zero force (purple circles). (#2) The protein-DNA complex is incubated at zero-force in clean buffer for a given time period. (#3) The protein-DNA complex is restretched (solid blue circles) up to 75 pN. (#4) A constant force is maintained at 75 pN until the extension is equilibrated. (#5) Return curves after extension achieves equilibrium in #4. Blue, green, red and gold empty circles are equilibrium force-extension curves of the protein-DNA complex obtained by incubating the protein-DNA complex in #2 for, 2 min, 5 min, 15 min and, 30 min, respectively. Time durations are the incubation time between consecutive stretching curves. Note that only one representative stretching curve is shown for step #3. (B) The extension-time profile as observed in step #4 in (A). Dashed black line is corresponding extension of a protein-free ssDNA at 75 pN. Dashed light blue line is the equilibrium extension achieved in #1 in (A) at 30 pN.
Discussion
Here we show that ORF1p binds ssDNA in a biphasic manner where rapid constriction is followed by elongation of the protein-ssDNA complex. Interestingly, the binding conformation of ORF1p depends on the free protein concentration in the solution, indicating a competitive binding mechanism that is regulated through the ssDNA occupancy. This behavior is similar to what we observe with *E. coli* single stranded DNA binding protein (EcSSB), as explained in chapter 2.3. Therefore, the initial biphasic binding of ORF1p to ssDNA is likely due to ssDNA wrapping and unwrapping transitions as observed with EcSSB. This hypothesis is supported by the fact that, like EcSSB (homotetramer), ORF1p also exists as a multimer (homotrimer), and according to structural evidence the trimer possesses surface properties that facilitate a single continuous strand of nucleic acid to wrap around the protein and occupy binding surfaces on each monomer. However, unlike EcSSB, ORF1p can rapidly form stable oligomers on ssDNA, and thereby increase the complexity of the supramolecular protein-DNA states.

For instance, in contrast to the *E. coli* binding dynamics (see chapter 2.3), the equilibrium extension change, as well as the rate at which this change occurs upon rinsing out free proteins in the solution, increases with protein concentration (Fig. 3.3.3D). The formation of higher-order protein multimers with increasing concentration may describe the concentration dependence of $k_3$. This is also consistent with the ability of ORF1p to rapidly oligomerize on ssDNA (see section 3.2). This hypothesis implies that higher-order multimer formation is facilitated when the ORF1p is in its unwrapped-bound state, which is favored at high protein-occupancy. In this scenario, the low-order ORF1p multimers stimulate the dissociation of the higher-order multimers by converting to the wrapped state. Thus, the protein-DNA interaction must be much
stronger in the low-order state (and thereby able to compete and wrap more ssDNA), which in turn suggests that, as the multimers in the unwrapped state grow in size, their interaction with DNA becomes weaker. This hypothesis may have interesting biological implications, providing a possible mechanism for ORF1p dissociation during TPRT.

We also show that the ssDNA-bound ORF1p complex slowly transforms into a highly stable conformer at low forces due to protein-protein interactions. However, the protein-protein interaction in this case is mediated by the wrapped ORF1p state at zero-force. Therefore, these inter-protein interactions are likely mediated through DNA and may differ structurally and functionally from the multimers that are formed in the unwrapped state. However, this hypothesis requires further testing with additional experiments, suggested in the future work section.

Overall, this study probes the nucleic acid packaging properties of ORF1p, which likely are involved in the formation of the ORF1p-mediated L1 RNP, a critical intermediate in L1 retrotransposition. Although a complete model for the observed-results is not available, this study has already revealed novel information on ORF1p-ssDNA binding dynamics. A detailed analysis to dissect the underlying mechanisms in the results we observe is currently in progress.
Chapter 4. Single-molecule characterization of nucleocapsid proteins from HIV and SIV

The work described in this chapter is conducted in collaboration with Judith G. Levin at NICHD, NIH, Bethesda, MD and Karin Musier-Forsyth and Ioulia Rouzina at The Ohio State University, Columbus, OH

4.1 Background

Figure 4.1.1: Replication cycle of a retrovirus (figure from\textsuperscript{210} reprinted by permission). (A) Viral entry into the host cell proceeds with the following steps: binding to a specific receptor on the cell surface; membrane fusion; release of the viral core and partial uncoating; reverse transcription; transit through the cytoplasm and nuclear entry; and integration into cellular DNA to give a provirus. (B) Viral exit involves the following steps: transcription by RNA polymerase II (RNAPII); splicing and nuclear export of viral RNA; translation of viral proteins, Gag assembly and RNA packaging; budding through the cell membrane; and release from the cell surface and virus maturation.

Retroviruses represent a family of single-stranded RNA viruses that incorporate their genome into the host cell via reverse transcription (Fig. 4.1.1). They are classified into seven genera based on genome complexity and virion morphology, namely, alpharetroviruses,
betaretroviruses, gammaretroviruses, deltaretroviruses, epsilon retroviruses, lentiviruses, and spumaviruses.

AIDS, a devastating disease that emerged in the late 20th century, is caused by two lentiviruses: HIV-1 and HIV-2. Early on, there was intense interest in the origin of these viruses and the AIDS pandemic, which by 2014 led to infection of 76.2 million people (UNAIDS 2014 estimates). Molecular and phylogenetic analyses of fecal samples collected from the forest floor, primarily in southern Cameroon, demonstrated that HIV-1 infection of humans resulted from cross-species transmission of a chimpanzee simian immunodeficiency virus (SIVcpz), a recombinant generated from two distinct monkey SIV lineages (reviewed in ). HIV-2 was transmitted to humans by an SIV from sooty mangabeys (SIVsm). Interestingly, SIVmac, a strain used in non-human primate model systems (including the present work), was acquired unexpectedly by transmission of SIVsm from sooty mangabeys to rhesus macaques at the California National Primate Research Center, where both groups of animals were housed.

Figure 4.1.2: Schematic of HIV-1 virion organization. Nucleocapsid (NC) proteins are shown in blue.
Like other retroviruses, SIV and HIV-1 have a nucleocapsid protein (NC), a small basic structural protein containing two zinc-binding domains, i.e., zinc fingers (ZFs), each with the invariant CCHC motif and connected by a short basic flexible peptide. NC is generated by viral protease (PR)-mediated cleavage of the Gag precursor protein during virus maturation. For HIV-1, it has been shown that the NC domain in Gag is essential for specific recognition of the Psi packaging element in genomic RNA (gRNA) and tRNALys3 primer placement on gRNA.

Retroviral NC proteins are nucleic acid (NA) chaperones, i.e., they remodel NA structures to facilitate formation of the most thermodynamically stable conformations (reviewed in refs. 45,47,238,239). This activity is critical for ensuring specific and efficient reverse transcription, including initiation, as well as the minus- and plus-strand transfer reactions. For example, in the minus-strand transfer step, NC facilitates annealing of the complementary repeat regions (R, r), which contain the highly structured transactivation response element (TAR) in gRNA and its minus-strand DNA complement, respectively.
4.2 Differential contribution of basic residues to HIV-1 nucleocapsid protein’s nucleic acid chaperone function and retroviral replication

The work described in this section was adapted from the following publication\textsuperscript{74}


Protein purification and ensemble results are the work of Mithun Mitra, Robert J. Gorelick, and Karin Musier-Forsyth. Theoretical work was done by Ioulia Rouzina. Single molecule analyses were done by Hao Wu, myself, Micah J. McCauley and Mark C. Williams.

Introduction

NA aggregation, destabilization and rapid protein-NA interaction kinetics\textsuperscript{71,243-246} are key requirements for nucleic acid chaperone activity. In vitro studies suggest that the ability of NC proteins to aggregate NA is due to their cationic character, and this activity is largely independent of their specific zinc-finger structures\textsuperscript{245,247,248}. In contrast, the other major component of the chaperone activity, NA destabilization, primarily requires properly folded zinc fingers\textsuperscript{249,250}. In the case of HIV-1 NC, this activity depends on the preferential binding of the zinc fingers to unpaired NA bases\textsuperscript{251,252}. The duplex destabilization activity differs significantly between different retroviral NC proteins, as well as related retrotransposons\textsuperscript{185,187,253,254}

The detailed relationship between HIV-1 NC’s structure and its ability to aggregate and destabilize NAs is not completely understood. For example, although the effective cationic charge of HIV-1 NC, defined as a negative slope of the log-log dependence of the $K_d$ versus salt concentration, is approximately $+3.5$\textsuperscript{255-258}, the total number of positively charged residues on this protein is 15, with only 4 negatively charged residues. Thus, it is unclear how and to what extent specific HIV-1 NC charged residues participate in nonspecific versus specific NA
binding. The role of specific basic residues in other chaperone activities, such as NA aggregation, is also unknown.

Figure 4.2.1: Sequence of WT HIV-1 NC (NL4-3 isolate) and variants investigated in this work. A: wild type; B: K3A/R7A/R10A/K11A/K14A (pentamutant); C: R7A/R10A/K11A (N-terminal trimutant); D: K14A/K20A/K26A (Zinc finger 1 trimutant); E: R29A/R32A/K33A/K34A (Zinc finger linker mutant).

To clarify the interplay between specific basic residues and HIV-1 NC’s NA chaperone function, we use several complementary in vitro approaches. Ensemble assays are used to quantify NA binding, aggregation and annealing activities of wild-type (WT) and mutant HIV-1 NC proteins. These results are compared with the results obtained from single molecule DNA stretching experiments. We find that cationic HIV-1 NC variants are defective in their overall NA binding affinity, aggregation and strand-annealing activities, but retain significant NA stacking capability.
at sufficiently high concentrations. These results are in contrast to previous work showing that aromatic residue variants are completely defective in stacking with NA bases and have greatly reduced single-stranded DNA (ssDNA) binding affinity. Cell-based assays showed that all of the cationic residue mutations investigated lead to reduced virus infectivity, which correlated strongly with measurements of TAR RNA and DNA annealing as well as the capability to intercalate DNA at high force in single molecule stretching experiments.

**DNA stretching in the presence of WT HIV-1 NC**

Fig. 4.2.1 shows the sequences of WT HIV-1 NC and the basic residue variants studied in this work. Single-molecule DNA stretching studies were carried out as previously described.

![Figure 4.2.2: FEC of DNA in the presence-HIV-1 NC](image)

As shown in Fig. 4.2.2, in the absence of protein, little force is required to stretch the DNA to its B-form contour length, which is about 0.34 nm/bp. As the contour length is approached, the force increases dramatically, reflecting the elasticity of the DNA helix. As the DNA is further stretched, it undergoes a force-induced overstretching transition from double-stranded DNA...
(dsDNA) to ssDNA. During this process, the DNA extension increases from 0.34 nm/bp to 0.6 nm/bp at an approximately constant force of 60 pN. The overstretched state of DNA can be either completely strand-separated with one or both strands under tension, or stretched to an unwound but still ds structure, depending on solution conditions, DNA sequence and pulling rate. Under conditions of relatively low solution ionic strength as used here, or in the presence of ssDNA binding proteins, the DNA becomes ss during the transition. In contrast, in the presence of proteins or ligands that preferentially bind to both DNA strands, the two strands may never separate. In the absence of protein, as the DNA is released back to lower extension, the force-extension curve is almost completely reversible, showing little hysteresis.

In the presence of WT HIV-1 NC, the overstretching transition becomes dramatically sloped, as the DNA starts to elongate beyond its B-form contour length from 20-30 pN (Fig. 4.2.2). The transition in this case is poorly defined and extends over a broad range of forces between 20 pN and 90 pN, leading to an apparent force-extension slope up to 350 pN/nm/bp (Fig. 4.2.2). A similar effect on DNA stretching was previously observed in the presence of intercalators such as ethidium, several ruthenium complexes, and threading- and bis-intercalators. In contrast to NC, these intercalators bind preferentially to dsDNA and increase the overall stability of duplex DNA, and this is usually reflected in an overall increase in the DNA stretching force. An example of a destabilizing intercalator is the anticancer drug Actinomycin D (Act D), which was recently characterized by DNA stretching. The stretching curves in the presence of Act D strongly resemble those observed in the presence of NC. This observation, when combined with our recent observation that the two DNA strands do not separate upon overstretching in the presence of NC, suggests that NC acts as a weak intercalator, which upon binding holds the two DNA strands together while simultaneously destabilizing the duplex form. Thus, the
significant DNA elongation observed in the presence of NC at forces greater than 20 pN results from the stacking of some NC residues between dsDNA bases, resulting in an intercalation process, similar to that observed for Act D\textsuperscript{267}, although it occurs at a much faster rate. In the case of HIV-1 NC, this duplex elongation likely occurs upon intercalation of the aromatic residues (Phe\textsubscript{16} and Trp\textsubscript{37})\textsuperscript{267,268}, similar to the stacking with bases that was observed in the NMR structures of HIV-1 NC bound to HIV-1 stem loops SL\textsubscript{2} and SL\textsubscript{3}\textsuperscript{252,253,258,269-276}. Here, we will use the value of the slope as a primary quantitative characteristic of NC-DNA binding, which reflects the ability of these aromatic residues to optimally stack with the DNA bases.

**DNA stretching with HIV-1 NC basic residue mutants**

Presented in Fig. 4.2.3 are the stretch and release curves obtained in the presence of HIV-1 NC basic residue mutants. All of these curves differ significantly, showing high sensitivity to even a few amino acid substitutions.

In order to summarize the protein-DNA interaction information contained in these data, we use two quantitative parameters describing the stretch-release cycle for the protein-DNA complex: the slope of the overstretching transition, $S$, and the scaled hysteresis in the stretch-release cycle, $H$, which is derived from the area between the stretch and release curves\textsuperscript{277}. $S$ is measured as the slope of the tangent line to the force-extension curve at the transition midpoint of the extension ($\sim0.48$ nm/bp). The protein-free slope, $S_0$, is approximately 20 pN/nm/bp, and the protein-saturated maximum slope, $S_{\text{max}}$, reflects the protein’s ability to intercalate into dsDNA upon saturated binding at high forces. A small transition slope close to $S_0$ indicates a force-induced DNA strand separation. In contrast, $S_{\text{max}}>>S_0$, reflects strong intercalation by the protein at high force. The theoretical maximum amount of hysteresis would be achieved if DNA could be stretched as dsDNA and released as ssDNA. We define the scaled hysteresis, $H$, as the actual
area of the protein-DNA stretch-release cycle divided by this maximum possible value. If the protein intercalates DNA, this may result in small H because it prevents the two strands from separating when being stretched.

If the protein does not intercalate DNA and the strands are separated by force during overstretching, then small H indicates that the strands rapidly anneal on the time scale of stretching. Larger values of H may occur if the DNA intercalates but the intercalation is slow, as is the case for Act D. If there is no intercalation and the strands separate during stretching, large

Figure 4.2.3: DNA stretching (solid line) and relaxation (dotted line) in the presence (red) and absence (black) of HIV-1 NC cationic mutants: (A) 200 nM N-terminal pentamutant [K3A/R7A/R10A/K11A/K14A]; (B) 200 nM N-terminal trimutant [R7A/R10A/K11A]; (C) ZF-1 cationic mutant [K14A/K20A/K26A]; (D) Zinc finger linker mutant [R29A/R32A/K33A/K34A].
hysteresis may result from protein binding to ssDNA and subsequent slow dissociation. We will use these observed features of DNA stretching curves below to probe the effects of cationic mutations on NC-DNA interactions.

**Effect of HIV-1 NC basic residue mutations**

Presented in Fig. 4.2.3A is a typical DNA stretch-release cycle in the presence of an HIV-1 NC variant with five N-terminal basic residues changed to alanine (K3A/R7A/R10/K11A/K14; pentamutant) (see Fig. 4.2.1B). Surprisingly, the dsDNA stretching curve in the presence of 200 nM pentamutant resembles the curve in the absence of protein. The fact that the transition slope remains close to the protein-free value (Fig. 4.2.3A) suggests that the protein does not intercalate DNA and the DNA is melted by force in the presence of the pentamutant NC, consistent with the large hysteresis observed upon DNA release. The large hysteresis also suggests strong and relatively slow binding of the pentamutant to ssDNA. It is initially surprising that the pentamutant NC is unable to intercalate DNA, as there is still an aromatic residue on each of the two zinc fingers. However, it has been shown that the strength of DNA intercalation by proteins depends strongly on interactions with amino acid side chains that are outside the intercalative wedge of the protein, and this must be the case for NC as well²⁷⁸. In addition, it is clear that the intercalative ability of the pentamutant is weakened but not eliminated, as the DNA stretching slope increases with concentration.

Fig. 4.2.3B shows the result of the DNA stretch-release curves in the presence of an N-terminal NC variant with only three of the basic residues mutated to alanine (R7A/R10A/K11A; N-terminal trimutant). This variant displayed a greatly increased $S$ relative to the pentamutant, but a reduced $H$, suggesting efficient DNA intercalation by the N-terminal trimutant, similar to that observed for WT NC. Two additional basic residue mutants containing three
(K14A/K20A/K26A; zinc finger 1 trimutant) or four (R29A/R32A/K33A/K34A; zinc finger linker mutant) changes displayed features suggesting weakened intercalative binding by these proteins as well as partial DNA melting by force, as determined by the smaller slopes and larger hysteresis observed compared to WT or N-terminal trimutant NC (Figs. 4.2.3C and 3D).

Figure 4.2.4: Concentration-dependence of transition slope and hysteresis (A) Slope of the transition versus concentration for WT HIV-1 NC and cationic mutants. Note that the 5 nM data points for the zinc finger 1 linker mutant and the N-terminal tri-mutant are coincident. (B) Hysteresis ratio versus concentration curves for WT HIV-1 NC and cationic mutants.

A summary of the $S$ and $H$ parameters calculated for all the cationic NC mutants investigated as a function of protein concentration is presented in Fig. 4.2.4, respectively. For all mutants, the $S$ parameter increases with concentration, suggesting that the ability to intercalate is strongly enhanced by binding that is facilitated by basic residues, and this enhancement is different depending on the location and number of the residues. The $H$ parameter also increases with concentration for all basic residue mutants studied, in contrast to WT NC (Figs. 4.2.4B). This result reflects the loss of the ability of all of these mutants to intercalate DNA, along with defects in their abilities to aggregate ssDNA and promote strand annealing after separation by force.
Based on these results, we can conclude that changes to any cluster of basic residues in the N-terminus, N-terminal zinc finger, or linker region of HIV-1 NC significantly impacts at least one of these characteristics, which are associated with NA chaperone activity.

The fact that all of these mutations alter DNA interactions is notable, as the net effective charge of WT HIV-1 NC binding to NA is only about +3.5, which is much smaller than the total number of cationic residues (+15)\textsuperscript{255,278} distributed throughout the protein. The effective charge implies that approximately 3.5 Na\textsuperscript{+} ions are released from the NA upon HIV-1 NC binding\textsuperscript{253,256}. The fact that mutation of several different subsets of cationic residues along the protein sequence affects the binding implies that a majority of the cationic residues participate in direct interactions with the NA phosphates, but to different extents. Thus, the extent to which mutations contribute to HIV-1 NC-DNA interaction defects does not completely correlate with the total number of altered basic residues (3, 4 or 5). The quantity most reflective of the ability of NC to intercalate DNA, S, is greatest for the least number of residues changed (N-terminal trimutant) and lowest for the greatest number of residues changed (pentamutant). The N-terminal trimutant and the zinc finger 1 trimutant exhibit different slopes within uncertainty, despite having the same number of basic residues mutated. These observations illustrate the differential contributions of the cationic mutations to the ability of NC to intercalate. These results are consistent with the effects observed for cationic mutations on high mobility group (HMG) proteins and other intercalating protein\textsuperscript{270}.

**Ensemble studies**

Results from fluorescence anisotropy (FA) binding studies, TAR RNA/DNA annealing assays, and aggregation assays are summarized in Table 4.2.1.
To investigate to what extent the *in vitro* measurements correlate with *in vivo* activity, cell-culture based studies were performed. Table 4.2.2 summarizes the effects of the NC basic residue mutations expressed in HIV-1 virions on gRNA packaging, and single- and multiple-round infectivity.

<table>
<thead>
<tr>
<th>HIV-1 NC Variant</th>
<th>Micro-TAR RNA&lt;sup&gt;a&lt;/sup&gt; (nM)</th>
<th>(TG)&lt;sub&gt;4&lt;/sub&gt; DNA&lt;sup&gt;a&lt;/sup&gt; (nM)</th>
<th>%TAR RNA aggregated&lt;sup&gt;b&lt;/sup&gt;</th>
<th>TAR RNA/DNA annealing rate&lt;sup&gt;c&lt;/sup&gt; min&lt;sup&gt;-1&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>315 ± 48</td>
<td>5 ± 1</td>
<td>94.2 ± 1.2</td>
<td>&gt; 6</td>
</tr>
<tr>
<td>N-terminal trimutant</td>
<td>1,720 ± 129</td>
<td>27 ± 2</td>
<td>84.7 ± 3.1</td>
<td>1.19 ± 0.56</td>
</tr>
<tr>
<td>ZF1 trimutant</td>
<td>591 ± 138</td>
<td>19 ± 10</td>
<td>93.8 ± 1.5</td>
<td>0.51 ± 0.38</td>
</tr>
<tr>
<td>ZF linker mutant</td>
<td>552 ± 20</td>
<td>82 ± 54</td>
<td>79.3 ± 9.5</td>
<td>0.60 ± 0.41</td>
</tr>
<tr>
<td>Pentamutant</td>
<td>2,440 ± 370</td>
<td>62 ± 20</td>
<td>85.4 ± 2.5</td>
<td>0.09 ± 0.07</td>
</tr>
</tbody>
</table>

Table 4.2.1: Binding, aggregation, and annealing parameters measured for WT HIV-1 NC and basic residue mutants.  
<sup>a</sup>Apparent equilibrium dissociation constants, $K_d$, for binding to NA oligomers measured at room temperature in 50 mM NaCl by FA as described in Materials and Methods.  
<sup>b</sup>Percent TAR RNA aggregated by 10 μM protein in a solution containing 15 nM radiolabeled TAR RNA and 45 nM TAR DNA at 37 °C in 50 mM NaCl.  
<sup>c</sup>TAR RNA-DNA annealing rate in the presence of 10 μM protein under the same solution conditions used in the aggregation assay.

<table>
<thead>
<tr>
<th>NC Protein</th>
<th>RNA packaging&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Single-round infectivity&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Relative multiple-round H9 infectivity&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT</td>
<td>100</td>
<td>100</td>
<td>1.0</td>
</tr>
<tr>
<td>N-terminal trimutant</td>
<td>51 ± 19</td>
<td>7.4 ± 1.6</td>
<td>(3.8 ± 3.3) × 10&lt;sup&gt;-3&lt;/sup&gt;</td>
</tr>
<tr>
<td>Zinc finger 1 trimutant</td>
<td>31 ± 14</td>
<td>8.8 ± 2.4</td>
<td>(5.0 ± 4.4) × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
</tr>
<tr>
<td>Zinc finger linker mutant</td>
<td>61 ± 11</td>
<td>3.2 ± 0.7</td>
<td>(7.2 ± 4.5) × 10&lt;sup&gt;-5&lt;/sup&gt;</td>
</tr>
<tr>
<td>Pentamutant</td>
<td>3.0 ± 0.3</td>
<td>0.016 ± 0.007</td>
<td>≥ (7.1) × 10&lt;sup&gt;-6&lt;/sup&gt;d</td>
</tr>
</tbody>
</table>

Table 4.2.2: Properties of WT HIV-1 NC and basic residue mutants measured in cells.  
<sup>a</sup>Determined by normalizing genome quantities to equivalent RT activities and reported as % of WT. Results from at least two separate experiments with standard deviations reported.  
<sup>b</sup>Either HCLZ or TZM-bl cells were used for these analyses. Titers were determined by taking the BCFU/mL and dividing by WT, corrected for input virus (based on exogenous template RT activity). Results are from at least four separate experiments. Errors are standard deviations.  
<sup>c</sup>Determined by taking the minimum dilution that gives rise to a spreading infection over 8 weeks (average of at least three infection experiments), normalized for equivalent exogenous template RT activities. Titers of mutants reported relative to a WT infection. Errors represent the standard error of the mean. The uncertainties are close to the mean due to the choice of dilutions used.  
<sup>d</sup>Three independent analyses were performed with the undiluted sample of the pentamutant being negative in each assay.

**Discussion**

In this work, we study the effect of HIV-1 NC cationic mutants on protein-NA interactions and retroviral replication. All of the basic residue mutations that were tested reduced micro-TAR
RNA binding affinity, with the pentamutant showing the greatest reduction in affinity, as expected. However, the N-terminal trimutant exhibits the strongest reduction in binding relative to the number of residues changed. In contrast, the strongest reduction in binding affinity to (TG)$_4$ DNA comes from the zinc finger linker mutant and the pentamutant, with the linker mutant having the strongest effect per residue mutated. Therefore, the ability of NC to bind short NA sequences decreases with the number of neutralized basic residues, but the positions of these residues are also important. These results suggest that the cationic residue mutations on the N-terminal domain most strongly alter the binding affinity to non-specific NA sequences, represented by the micro-TAR RNA, while mutations on the zinc finger linker have a greater effect on the protein’s ability to stack with NAs, as required for optimal binding to (TG)$_4$. Although all of the mutants aggregated NA at high concentration, they were less effective aggregating agents at the lowest concentration tested (0.88 µM) and the percent aggregated even at 10 µM was somewhat less for all of the mutants except the zinc finger 1 mutant.

The single molecule studies also revealed defects in DNA interactions upon basic residue mutation. For all of the mutants, the ability of NC to alter the slope of the DNA stretching curve was significantly compromised, but the slope change was recovered to some extent by increasing the concentration of protein used. The ability to alter the stretching slope was weakest for the pentamutant, followed by the zinc finger linker mutant, the zinc finger 1 trimutant, and finally the N-terminal trimutant, which was closest to WT NC activity. However, while these effects increase with the number of residues mutated, the zinc finger 1 and N-terminal trimutants exhibit significantly different effects, supporting the conclusion that the position of the residues that are mutated is critical. Similarly, the amount of hysteresis measured for the mutants also increases with the number of residues mutated, with the hysteresis area ratio at 200 nM being greatest for
the pentamutant. However, in this case the zinc finger 1 trimutant is almost as defective as the pentamutant, and the other mutants are closer to WT NC. Overall, the single molecule studies show that each set of mutated residues decreases NC’s ability to optimally interact with DNA to varying extents, depending on the location of the mutated residues.

To determine the extent to which the results of these *in vitro* measurements reflect the ability of NC to facilitate viral replication in cells, we calculated the correlation coefficient between the *in vitro* single molecule, binding, and annealing measurements and measurements of RNA packaging, single round infectivity, and multiple round infectivity in cells (Table 4.2.3). Because cationic residue mutations are expected to primarily reduce electrostatic binding interactions, it would be reasonable to expect the cell-based measurements to result in a negative correlation coefficient with $K_d$. Surprisingly, however, the measurements in cells correlate only moderately with binding affinity to microTAR or $(TG)_4$ DNA, with the absolute value of these correlation coefficients less than the 90% confidence interval for two of the three measurements, and less than 95% confidence for correlation with all three. These measurements confirm that there is not a strong correlation between *in vitro* binding affinity and measurements in cells, although removal of the wild type data point results in a strong negative correlation between the binding affinity of mutant NC to $(TG)_4$ DNA and single round infectivity. The lack of strong correlation between the binding of the cationic NC mutants and the extent of RNA packaging and replication defects observed in cells is consistent with the hypothesis that the aromatic zinc finger residues play the primary role in gRNA selection and chaperone function, while the cationic residues play only a secondary role in these functions. This result is also consistent with the hypothesis that the chaperone activity of NC occurs *in vivo* when NC concentrations are in excess of the $K_d$ values observed here even for the most binding-defective cationic NC mutants.
For example, cumulative evidence suggests that HIV-1 reverse transcription is mechanistically linked to capsid uncoating\textsuperscript{279-288} and the early steps of reverse transcription are likely to occur within an intact capsid core.

<table>
<thead>
<tr>
<th>Correlation</th>
<th>RNA packaging</th>
<th>Single-round infectivity</th>
<th>Log (multiple round infectivity)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K_d ), microTAR RNA</td>
<td>-0.74</td>
<td>-0.53</td>
<td>-0.48</td>
</tr>
<tr>
<td>( K_d ), (TG)(_4) DNA</td>
<td>-0.42</td>
<td>-0.65</td>
<td>-0.69</td>
</tr>
<tr>
<td>RNA/DNA annealing rate</td>
<td>0.85</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>DNA stretching hysteresis</td>
<td>-0.91</td>
<td>-0.52</td>
<td>-0.76</td>
</tr>
<tr>
<td>Transition slope</td>
<td>0.89</td>
<td>0.81</td>
<td>0.96</td>
</tr>
</tbody>
</table>

\textbf{Table 4.2.3: Correlation coefficients} are determined by using values from Table 4.2.1 for \( K_d \) and Table 4.2.2 for measurements in cells. Single molecule transition slope and hysteresis were evaluated at 200 nM NC for the mutants and 20 nM for WT NC. To consider the exponential effect of multiple rounds of replication, we used the logarithm of the infectivity for correlation calculations. Positive results mean that the two quantities are correlated, and negative results mean inverse correlation. Assuming a one-tailed distribution, for five measurements the correlation must be greater than 0.805 for 95\% confidence in the correlation, and 0.687 for 90\% confidence\textsuperscript{289}. Therefore, correlations greater than 0.69 should be considered very strong for this number of measurements.

The ability of NC to facilitate TAR RNA/DNA annealing appears to correlate very strongly with RNA packaging and replication measurements based on Table 4.2.3, consistent with the importance of overall NA chaperone activity for viral replication\textsuperscript{290}. However, because the results of the TAR RNA/DNA annealing measurements for wild-type protein are significantly different than those of the mutants, this leads to an anomalously high correlation between annealing and measurements in cells. When the wild type NC is not considered in the correlation measurements, the correlation coefficients become 0.75, 0.66, and 0.99 for correlation of annealing rate with RNA packaging, single round infectivity, and the logarithm of multiple round infectivity, respectively. This suggests a moderate to strong correlation between TAR RNA/DNA annealing and replication measurements in cells. Similarly, the ability of NC to prevent DNA strand separation by force, as reflected in small hysteresis, also correlates very strongly with RNA packaging, and moderately with multiple round infectivity (Table 4.2.3,
DNA stretching hysteresis). The hysteresis measurements do not change significantly when influential data points are removed. In addition to RNA/DNA annealing, the other strong correlation between *in vitro* measurements and measurements in cells is in the slope of the single molecule DNA stretching curve. All three measurements in cells correlate very strongly with the slope $S$, even when influential data points are removed. As discussed above, this change in slope reflects the ability of the protein to elongate dsDNA without allowing the strands to separate. This elongation at high forces represents a measurement of DNA intercalation by NC, which only occurs at high forces. However, the correlation with replication measurements suggests that this optimal ability to intercalate is a critical component of NC’s NA chaperone activity. Similar stacking interactions between the F16 and W37 residues of NC and ss bases were recently characterized by NMR$^{277}$. These interactions resemble the hemi-intercalation observed between the aromatic rings of the dsDNA intercalator Actinomycin D and ssDNA bases$^{277}$. While the observed intercalation is too weak to occur frequently at zero force when NAs are fully double-stranded, the intercalation measured in these single molecule experiments likely becomes very important in cases of locally unstable elements of NA secondary structure, such as duplexes containing mismatches, loops, or bulges. Thus, these combined experiments demonstrate a strong correlation between *in vitro* NA chaperone activity and cellular replication measurements, illustrating the importance of specific HIV-1 NC basic residues for these processes.

**Materials and Methods**

**Ensemble studies**

The generation of plasmids, mutagenesis and recombinant protein production is described in detail elsewhere$^{74}$. Recombinant NC proteins, 55 amino acids in length were expressed and purified as described$^{291,292}$. The mutant NC amino acid sequences examined in either viruses or
as purified NCp7 are presented in Fig. 4.2.1. Detailed information and methods on ensemble measurements in this study using fluorescence anisotropy binding studies, TAR RNA/DNA annealing assays, aggregation assays, and cell culture-based assays is available elsewhere

**Single molecule DNA stretching studies**

In the single-molecule stretching experiments performed essentially as described, two laser beams are overlapped to trap one streptavidin-coated bead. A second streptavidin-coated bead is attached to a glass micropipette. Bacteriophage λ DNA, which was labeled with biotin on both 5’ ends as described, is caught between the beads due to strong non-covalent bonds between biotin and streptavidin. All stretching experiments were performed at a constant pulling rate of ~100 nm/s in 10 mM HEPES, 50 mM Na+, pH 7.5 buffer. After attachment of one DNA molecule, buffer was used to rinse out the other DNA molecules, and solutions containing specific protein concentrations were flowed around the DNA to investigate protein effects on DNA stretching curves.
4.3 Mechanistic differences between HIV-1 and SIV nucleocapsid proteins and cross-species HIV-1 genomic RNA recognition

The work described in this section was adapted from the following publication76.


Protein purification and ensemble results are the work of Klara Post, Erik D. Olson, Robert J. Gorelick, Karin Musier-Forsyth, and Judith G. Levin. Theoretical work was done by Ioulia Rouzina.

Introduction

The simian origin of HIV-1, SIV NC shares 53% amino acid sequence identity with the HIV-1 protein (Fig. 4.3.1). Early work demonstrated that both SIV from *Macaca nemestrina* (SIVmne) and HIV-1 NCs had very similar NA binding properties in studies with model oligo- and polynucleotide substrates289. Moreover, the NMR solution structure of SIVl’hoest NC (residues

13-51) showed that the overall structures of SIV and HIV-1 NC are also very similar, despite several amino acid sequence differences in the ZFs and structural differences in the flexible linker293. The ability of SIV NC to coordinate Zn2+ is required for efficient replication in cell-based assays294,295, Gag processing296,297, proper core condensation and NC protein stability296, as well as RNA packaging, although to a lesser extent than for MLV and HIV-1295. In contrast to HIV-1 NC, ZF2 of SIV NC appears to be slightly more important than ZF296,297. Interestingly, compared with the HIV-1 protein, the chaperone activity of HIV-2 NC is not as robust, likely due, at least in part, to the shorter HIV-2 N-terminal basic region296.

This section discusses the in-depth analysis of the chaperone activities of SIV and HIV-1 NCs in the context of biologically relevant reactions: the minus-strand transfer step in reverse transcription and selective binding to the Psi packaging element.
**Figure 4.3.1: Sequence and structural features of HIV-1 and SIV NC proteins.** (A) Schematic diagrams of NC proteins: HIV-1 NL4-3 NC and SIVmne NC. Basic residues are colored blue, acidic residues are colored red, the CCHC residues that coordinate the Zn$^{2+}$ ions in the ZFs are colored gray, and the aromatic residue in each finger is underlined. The numbering is based on the sequence of the mature NC protein in each case. (B) Sequence alignment of HIV-1 and SIV NC proteins. Coloring and underlining are the same as in (A). The boxes indicate the sequence comprising each ZF.

Using a variety of biochemical and biophysical (e.g., single molecule DNA stretching and dynamic light scattering) approaches, we show that the slightly higher activity of HIV-1 NC in the annealing reaction in minus-strand transfer can be explained by the greater aggregation activity of the HIV-1 protein relative to SIV NC. Salt-titration assays show that both NC proteins have a similar balance of electrostatic and specific binding contacts with NAs and both can distinguish HIV-1 Psi RNA from non-Psi sequences. However, neither protein is capable of specific binding to the putative SIV RNA packaging signal tested here. Overall, while most of the NA binding and chaperone activities of HIV-1 and SIV NC are similar, our analysis reveals mechanistic differences that provide unique information regarding the replication strategies of HIV-1 and SIV.
Results

Comparison of SIV and HIV-1 NC proteins and predicted secondary structures of TAR and Psi RNAs

The sequences of the SIV and HIV-1 NC proteins are compared in Fig. 4.3.1. Both proteins are highly basic and have two ZF domains containing conserved CCHC motifs and aromatic residues. However, HIV-1 NC is more basic than SIV NC over a wide range of pH; at pH 7, for example, the estimated charge for HIV-1 is 11.2 and for SIV, 10.2. Additionally, ZF1 of SIV NC has a Trp residue, whereas the corresponding amino acid in HIV-1 NC is Phe. This difference is expected to be minor, since mutation of F16 to W has little effect on HIV-1 NC NA chaperone activity and intravirion reverse transcription and no effect on infectivity[^298].

During the course of virus replication, the HIV-1 NC protein interacts with structured RNA elements present at the 5′ end of the viral genome: the TAR stem-loop (SL), which is at the extreme 5′ end of R in gRNA (Fig. 4.3.2A) and is involved in the minus-strand transfer step of reverse transcription (reviewed in refs.[^47,73]); and the Psi region, composed of three SLs, which contributes to the dimerization and packaging of gRNA and has been studied extensively[^228,299]. The SL structures include SL1, which contains the dimerization initiation site (DIS) and two bulges; SL2 containing the major 5′ splice donor site; and SL3, which is important for packaging viral RNA (Fig. 4.3.2B). The earlier studies showed that while SL1-3 are all necessary for efficient gRNA encapsidation, SL1 and SL3 play a larger role than SL2. The RNA sequence that constitutes the SIV Psi element has not been studied to the same extent as that of HIV-1. The available information suggests that SIV and HIV-1 Psi share a similar secondary structure and that this region is also critical for SIV gRNA packaging[^300-302] (Fig. 4.3.2D).
Figure 4.3.2: Sequence and mfold-predicted secondary structure of TAR and Psi RNA constructs used in this study. (A) HIV-1 TARpolyA, (B) HIV-1 Psi. (C) SIVmac TAR. (D) SIVmac Psi. In all cases, numbering refers to the nt position in gRNA. The box in (A) indicates HIV-1 TAR RNA. The boxes in (B) and (D) indicate the ΔDIS mutation, where DIS loop residues are replaced with a GNRA-type tetraloop (GAGA) to eliminate dimerization.
Figure 4.3.3: Kinetics of minus-strand annealing with SIV and HIV-1 substrates in the presence of SIV and HIV-1 NC proteins. (A) Reconstituted system used to assay minus-strand annealing and transfer. The diagram shows the acceptor RNA with a portion of U3 and the R sequence at the 3′ end of the viral genome annealed to (-) SSDNA with the complementary r sequence and a portion of u5, complementary to the U5 sequence. For the SIV substrates, the nt lengths of u5, R/r, and U3 sequences are as follows: u5, 20 nt; R/r, 176 nt; and U3, 52 nt. For the HIV-1 substrates, the lengths are: u5, 34 nt; R/r, 94 nt; and U3, 54 nt. The asterisk indicates that the (-) SSDNA is labeled at its 5′ end with ³²P. Annealing of the complementary R regions is indicated by vertical lines. The U3 sequence serves as the template for RT-catalyzed extension of annealed (-) SSDNA. The final DNA transfer product is 248 nt (SIV) or 182 nt (HIV-1). The diagram is not drawn to scale. (B-1 and B-2) and (C-1 and C-2). Reactions were incubated with SIV (B-1 and B-2) or HIV-1 substrates (C-1 and C-2) and different concentrations of SIV NC or HIV-1 NC for 30 min at 37 °C and analyzed as described in Materials and Methods. The percent (%) annealed product was plotted against time of incubation. Error bars represent the standard deviation (SD) from three or more independent experiments.

In the binding and small angle scattering (SAXS) experiments described below, we used an HIV-1 TARpolyA construct (Fig. 4.3.2A) to more closely match the longer and more complex SIV TAR structure (Fig. 4.3.2C).³⁰³,³⁰⁴ Note that although the overall predicted secondary structures of HIV-1 and SIV Psi are similar (Fig. 4.3.2B and D, respectively), SIV Psi is 40 nucleotides (nt) longer than HIV-1 Psi, with significantly longer predicted SL1 and SL3 stem regions.

Minus-strand annealing and strand transfer activities of SIV and HIV-1 NCs
In view of the critical role of NC chaperone activity in reverse transcription, we investigated the activity of SIV and HIV-1 NCs in model systems that recapitulate the reactions required for minus-strand transfer (Fig. 4.3.3A). In the initial step, a DNA oligonucleotide representing (-) SSDNA (derived from sequences complementary to the 5′ end of the genome) is annealed to an RNA transcript representing the acceptor RNA (derived from sequences at the 3′ end of the genome) (Annealing). The annealed DNA is then extended by reverse transcriptase (RT) to give the transfer product (Minus-Strand Transfer). Note that in our systems, the minus-strand transfer assay includes the annealing step as well as DNA elongation.

Figure 4.3.4: Kinetics of SIV minus-strand transfer in the presence of SIV and HIV-1 NC proteins. Reactions were incubated with the indicated concentrations of SIV or HIV-1 NC for 60 min at 37 °C and were analyzed as described in Materials and Methods. (A) Representative gels showing DNA species present in reactions with 1.25 µM SIV or HIV-1 NC. The transfer product (T) and (-) SSDNA are indicated to the left of the gel image and these were the only two bands that appeared on the gel. Note that self-priming products were not formed under the conditions used for these assays. Lane C shows the migration of (-) SSDNA in the absence of other reactants. (B) and (C). The % strand transfer product formed was plotted against time of incubation for reactions with SIV NC (B) or HIV-1 NC (C). Error bars represent the SD from three or more independent experiments.

To compare the annealing activities of the two NCs (Fig. 4.3.3), we evaluated reactions containing SIV substrates and either SIV NC (B-1) or HIV-1 NC (B-2), as well as reactions with
HIV-1 substrates and SIV NC (C-1) or HIV-1 NC (C-2). As the NC concentration was increased, the extent of annealing was also increased. For example, when 1.25 µM NC was used to measure SIV annealing, the percent annealed product was ~60% with both NCs, but at 5.0 µM NC, plateau values of almost 80% were reached (Fig. 4.3.3B-1 and B-2). In general, at the 30 min end point, the percent annealed DNA was very similar for reactions with SIV or HIV-1 NC; in some cases the values obtained with HIV-1 NC (with either substrate set) were slightly higher than the values observed with SIV NC, but the difference was never greater than ~1.4-fold (e.g., reaction with HIV-1 substrates and NC at 0.65 µM: compare data in C-1 (31%) and C-2 (42%)).

In contrast, comparison of the rates of annealing showed small, but more significant differences between the activities of the two NCs (Table 1A). Thus, with the SIV substrates, the rate of annealing with HIV-1 NC was ~3-fold higher than the rate with SIV NC, whereas with the HIV-1 substrates, the difference was ~2-fold. SIV NC appeared to be somewhat more active with the HIV-1 substrates, but HIV-1 NC had the same activity in both systems.

Minus-strand transfer was tested next, initially in reactions with SIV substrates and either SIV or HIV-1 NC (Fig. 4.3.4). The gel images clearly show bands corresponding to the transfer product and (-) SSDNA (Fig. 4.3.4A). As was observed for annealing, the end point values for SIV minus-strand transfer were similar for reactions with SIV (Fig. 4.3.4B) or HIV-1 (Fig. 4.3.4C) NC. For example, at 1.25 µM NC, the percent strand transfer with SIV NC was 42% and with HIV-1 NC, it was 51%. At 5 µM, the plateau values were 57% (SIV NC) and 70% (HIV-1 NC). Again, there was a small, but more significant difference in the reaction rates with the two NCs. With 1.25 µM HIV-1 NC, the rate was 2.3-fold greater than with SIV NC. Not surprisingly, minus-strand transfer with the HIV-1 substrates was slightly more efficient than with the SIV
substrates, but in this case too, the end point values with 1.25 µM NC were very similar (51%, SIV NC; 64%, HIV-1 NC). The rate of the reaction was 2.6-fold higher with HIV-1 NC.

**Single molecule DNA stretching experiments**

To further understand differences between HIV-1 NC and SIV NC, we tested the force-extension (stretch and return) curves of single DNA molecules as a function of protein concentration for both proteins. As shown in Fig. 4.3.5A, the shapes and qualitative characteristics of the force-extension curves were very similar. To quantify these characteristics, we calculated the transition slope, which reflects the degree of intercalative binding by NC to DNA, and the hysteresis area ratio, which reflects the amount of strand separation, as a function of concentration 73,186,240.

The slope of the force-extension curve is measured near the midpoint of the transition by averaging over the extensions between 0.4 and 0.5 nm/base pair (bp). To determine the hysteresis area ratio, we find the linear combination of the worm-like chain (WLC) model, which describes double-stranded DNA (dsDNA), and the freely-jointed chain (FJC), which describes single-stranded DNA (ssDNA), that intersects the highest extension of the data. This makes it possible to obtain a value for the hysteresis area ratio that is independent of how far the DNA is stretched for a particular curve. The results of these quantitative analyses as a function of concentration are shown in Fig. 4.3.5B and C. The transition slope measurements suggest that SIV and HIV-1 NC have similar binding affinities in the nM range, as shown by their nearly identical equilibrium dissociation constants: $K_d = 5.5 \pm 0.4$ nM and $4.2 \pm 0.4$ nM, respectively (Fig. 4.3.5B). The overall maximum transition slope is slightly higher for SIV NC, consistent with its slightly stronger intercalative binding relative to HIV-1 NC. Analysis of the hysteresis as a function of concentration also shows that for the lowest concentrations tested, HIV-1 NC
binding resulted in a larger hysteresis area ratio compared to SIV NC. At all other concentrations tested, the proteins behaved in a very similar manner (Fig. 4.3.5C).

**Figure 4.3.5:** (A) Force-extension curves for dsDNA stretch (solid lines) and return (dashed lines) with no protein and in the presence of 30 nM SIV NC or HIV-1 NC. (B) and (C) Dependence of the measured transition slope (B) and hysteresis area ratio (C) on protein concentration for HIV-1 NC and SIV NC. The lines in panel B are fits to a simple binding isotherm, revealing equilibrium dissociation constants $K_d = 5.5 \pm 0.4 \text{ nM}$ for SIV NC and $K_d = 4.2 \pm 0.4 \text{ nM}$ for HIV-1 NC. Error bars are standard errors for three or more measurements.

The primary difference between HIV-1 NC and SIV NC, as observed in single molecule DNA stretching experiments, can be seen upon close examination of the force-extension curves at low forces and extensions (Fig. 4.3.6A). To stretch dsDNA at extensions below the dsDNA contour length of 0.34 nm/bp in the presence of protein, higher forces are needed relative to the “DNA only” sample. This additional force at low extensions is referred to as the DNA compaction force ($F_c$). The magnitude of the $F_c$ reflects the ability of the protein to attract dsDNA, which normally results in DNA aggregation in the absence of applied force$^{186,307}$. To quantify this compaction force, we used the method described in the legend to Fig. 4.3.6A. The results showed that the compaction force for HIV-1 NC is ~2-fold higher than that of SIV NC at both 30 nM and 60 nM concentrations (Fig. 4.3.6B).

The additional compaction force for HIV-1 NC relative to that of SIV NC, averaged over both concentrations and weighted by uncertainty, is $1.3 \pm 0.4 \text{ pN}$. This corresponds to a difference in
compaction energy of $0.11 \pm 0.03 \text{ k_BT/bp}$. Thus, for a 10 kbp dsDNA molecule, similar to the length of the HIV-1 genome, the additional compaction energy for HIV-1 NC is $1100 \pm 300 \text{ k_BT}$, or $640 \pm 170 \text{ kcal/mol}$, which is a very large energy difference for a molecular process.

Figure 4.3.6: (A) Method for calculating the compaction force ($F_c$) induced by protein-DNA interactions. Inset shows stretch (solid lines) and return (dashed lines) curves for dsDNA in the absence of protein and in the presence of near saturated (30 nM) HIV-1 NC protein. $F_c$ is calculated in the low force-extension regime denoted within the gray box in the inset and magnified in the main figure. The DNA only extension curve is fit to the WLC model. The force difference ($F_c$) between the return curve in the presence of high protein concentration and the DNA only stretching curve is averaged over measured extensions less than 0.31 nm/bp to obtain $F_c$. (B) $F_c$ for SIV NC and HIV-1 NC for concentrations of 30 nM and 60 nM. Error bars are standard errors for three or more measurements.

Analysis of HIV-1 and SIV NC NA aggregation properties by dynamic light scattering (DLS)

The ability of retroviral NC proteins to aggregate NAs is important for NC’s NA chaperone function, as well as for formation of the ribonucleoprotein complex containing gRNA that is located within the mature virion core \(^{192,308-310}\). We used DLS to characterize the NA aggregate size generated by HIV-1 and SIV NCs in the presence of SIV Psi RNA (Fig. 4.3.7). In addition to the reactions with NC, a no NC control was included. The average size of the NA aggregate formed in the absence of NC was $0.74 \pm 0.02 \text{ nm}$ (n=4) in diameter and ranged from 0.54 to 1.1
nm, consistent with the lack of aggregation under these conditions. In contrast, HIV-1 NC generated aggregates with a mean diameter of 642 ± 60 nm (n=3) and a range from 164 to 1484 nm. These values agree with previous reports using different NA substrates \(^{309,311,312}\)

Interestingly, the range of NA aggregates produced by SIV NC was found to be only 106 to 1106 nm, with an average size of 448 ± 65 nm (n=5), which is smaller than the corresponding NA aggregates produced by HIV-1 NC. Taken together, these data suggest that HIV-1 NC is a more effective aggregating agent than SIV NC, consistent with the \(F_c\) measurements (Fig. 4.3.6B).

**RNA Binding properties of HIV-1 and SIV NC proteins**

Retroviral NC proteins interact with NAs using both specific and non-specific modes of binding \(^{312}\). To evaluate the RNA binding properties of SIV and HIV-1 NCs, we examined the interaction of these proteins with four RNA constructs: HIV-1 TARpolyA; SIV TAR; HIV-1 Psi; and SIV Psi (Fig. 4.3.2). Since HIV-1 TARpolyA and SIV TAR sequences have been shown to be largely dispensable for selective gRNA packaging in HIV-1 and SIV virions, respectively\(^{230,301,302,313,314}\), we used these non-Psi RNAs to assay non-specific binding.

Fluorescence anisotropy (FA) salt-titration assays have previously been shown to be capable of distinguishing the relative contribution of specific vs. non-specific or electrostatic interactions for any given NC-RNA binding event\(^{228,230}\). Briefly, the FA signal emitted by fluorescently-labeled RNA is measured at fixed protein and RNA concentrations, while the NaCl concentration is varied. As the salt concentration increases, less NC is able to bind RNA, resulting in a decrease in the FA signal. Thus, a protein-RNA complex that is more dependent on electrostatic interactions than specific contacts, dissociates at lower salt concentration relative to a complex that is characterized by specific non-electrostatic binding interactions. To quantify the results, the
data are fit to an equation (see Methods), which yields the parameters \( K_{d(1M)} \) and \( Z_{eff} \). \( K_{d(1M)} \) represents the \( K_d \) of the protein-RNA interaction when all electrostatic contacts have been screened out and only specific ones remain (e.g., hydrogen bonding or aromatic stacking interactions). \( Z_{eff} \) represents the number of Na\(^+\) ions displaced from the RNA upon protein binding, which corresponds to the number of electrostatic contacts made between the protein and the RNA.

We performed FA salt-titration assays using the four SIV and HIV-1 RNAs shown in Fig. 4.3.2 and the corresponding NC proteins. HIV-1 NC binding to HIV-1 TARpolyA and Psi substrates was characterized by \( K_{d(1M)} \) values equal to \( 1.2 \times 10^{-4} \) M and \( 4.0 \times 10^{-6} \) M, respectively, and \( Z_{eff} \) values equal to 2.4 and 1.2, respectively (Fig. 4.3.8). The significant difference (~30-fold) between the \( K_{d(1M)} \) values for the HIV-1 NC-Psi and TARpolyA interactions is in general agreement with our previous report and shows that NC binds more specifically to Psi RNA relative to TARpolyA\(^{315}\). SIV NC binding to HIV-1 TARpolyA and Psi RNAs yielded \( K_{d(1M)} \) values of \( 8.2 \times 10^{-5} \) M and \( 3.9 \times 10^{-6} \) M, respectively, and \( Z_{eff} \) values of 2.1 and 1.3, respectively (Fig. 4.3.8). These values are very similar to the values obtained with HIV-1 NC. Comparable binding affinities of HIV-1 and SIV NC to SL structures in HIV-1 Psi have also been reported as “unpublished results” in ref.\(^{230}\). However, the current data also indicate that both proteins bind HIV-1 Psi RNA in a more specific, non-electrostatic manner than TARpolyA RNA.

In contrast, HIV-1 NC was found to interact with SIV TAR and Psi RNAs with very similar \( K_{d(1M)} \) values of \( 8.1 \times 10^{-5} \) M and \( 1.0 \times 10^{-4} \) M, respectively, and \( Z_{eff} \) values of 2.6 and 3.6, respectively (Fig. 4.3.8). SIV NC was also unable to effectively discriminate between SIV TAR and Psi RNA interactions with \( K_{d(1M)} \) values of \( 6.2 \times 10^{-5} \) M and \( 7.2 \times 10^{-5} \) M, respectively, and \( Z_{eff} \) values of 2.4 and 3.0, respectively. These results suggest that neither HIV-1 nor SIV NC
interacts with the SIV Psi RNA construct used here with greater specificity than with a non-Psi sequence and also show that regardless of the RNA substrate tested, both HIV-1 and SIV NCs interacted with very similar $K_{d(1M)}$ and $Z_{eff}$ values (Fig. 4.3.8).

**SIV RNA interactions with HIV-1 Gag**

We next wanted to establish whether specific Psi RNA recognition in the SIV system requires a Gag polyprotein. In earlier work, it was demonstrated that HIV-1 GagΔp6 binds to HIV-1 Psi RNA with even greater specificity than the NC domain alone\textsuperscript{293}. It was therefore of interest to test HIV-1 Gag’s binding to SIV Psi versus TAR RNA (Fig. 4.3.8).

In accord with previous results\textsuperscript{230}, HIV-1 Gag clearly exhibited specific binding to HIV-1 Psi versus TARpolyA with $K_{d(1M)}$ values of $1.8 \times 10^{-4}$ M and $1.3 \times 10^{-1}$ M, respectively (Fig. 4.3.8A). Similarly, $Z_{eff}$ values of 5.0 and 8.7 for binding to HIV-1 Psi and TARpolyA, respectively, (Fig. 4.3.8B) were in good agreement with the values obtained in the earlier study\textsuperscript{230}. In contrast, when we tested HIV-1 Gag with SIV TAR and Psi RNAs, the $K_{d(1M)}$ and $Z_{eff}$ values were very similar: $K_{d(1M)}$ values of $1.2 \times 10^{-1}$ M and $4.7 \times 10^{-2}$ M, respectively, and $Z_{eff}$ values of 8.0 and 8.8, respectively (Fig. 4.3.8). Thus, as we found with the NC proteins, HIV-1 Gag was also unable to differentiate the SIV Psi RNA construct from the non-Psi TAR RNA.

**SAXS reveals overall shape of SIV Psi RNA**

We considered the possibility that the NC proteins and HIV-1 Gag are not able to specifically bind SIV Psi RNA due to structural differences between HIV-1 and SIV Psi RNAs. The structure of the HIV-1 Psi construct used for the current work was previously characterized by SAXS\textsuperscript{230} and we now applied this approach to SIV Psi. To ensure that the RNA was monomeric, the wild-type (WT) DIS loop was replaced with a GNRA-type GAGA tetraloop (ΔDIS mutation, Fig. 4.3.2D). SIV Psi-ΔDIS was purified by size exclusion chromatography (SEC) prior to analysis.
by SAXS. The SEC trace and subsequent analysis by electrophoresis in a native gel (performed concomitant with SAXS data acquisition) confirmed that SIV Psi-ΔDIS was predominantly monomeric.

The SAXS envelope generated for SIV Psi-ΔDIS was compared to that of HIV-1 Psi-ΔDIS. As previously reported, all stem-loops of HIV-1 Psi-ΔDIS are solvent exposed and clearly discernable in the envelope with apparent co-axial stacking between SL1 and SL3231. In contrast, the SIV Psi RNA appears more globular and there is no apparent co-axial stacking between the stem-loops. The SAXS data are therefore consistent with the conclusion that SIV Psi-ΔDIS adopts an altered global fold relative to HIV-1 Psi-ΔDIS.

Discussion

The goal of the present study was to obtain a detailed comparison of the NA chaperone activities of SIV NC and the more extensively characterized HIV-1 protein. This issue is of great importance in view of the essential role of the NC protein in retrovirus replication231,316 and the widespread use of simian model systems for studies on HIV-1 pathogenesis, vaccine development, and drug resistance317, and more recently, the development of HIV-1 eradication and cure strategies318. Here, we focus on two important events in the virus life cycle: the minus-strand transfer step in reverse transcription that is required for synthesis of a full-length copy of the viral RNA genome; and selective gRNA packaging directed by the Psi structural element. Differences in the global folds of putative SIV and HIV-1 Psi RNA sequences were uncovered in this study, although SIV and HIV-1 NC proteins exhibited similar behavior when interacting with each of these RNAs. Our results are consistent with the earlier NMR structural study of SIV NC319, as well as with previous functional analysis294, which also came to the conclusion that HIV-1 NC is highly similar to the SIV protein. However, we also demonstrated that despite an
overall functional similarity, quantitative differences in NA aggregation and compaction capability distinguish the two proteins, which may be relevant to the infection process.

In our initial approach, we examined the minus-strand transfer reaction, since the rate-limiting step, i.e., annealing of the complementary R regions, is strongly dependent on NA chaperone activity to transiently destabilize the TAR structures and facilitate efficient NA binding. Assays of annealing or annealing plus DNA elongation (Fig. 4.3.3-4) showed that the SIV NC-mediated reactions occur at a 2- to 3-fold slower rate than with HIV-1 NC, although the extent of product formation after incubation for 30 or 60 min is only slightly elevated in the presence of HIV-1 NC. To understand the physical basis for this behavior, we performed single molecule DNA stretching and DLS experiments.

In single molecule stretching determinations, both HIV-1 and SIV NC showed very similar NA binding affinities and qualitative interactions with DNA, consistent with their very similar domain structure (Fig. 4.3.5). However, a more quantitative analysis revealed a significant difference in the compaction force, $F_c$, generated by these two proteins at extensions less than the DNA contour length (Fig. 4.3.6). The $F_c$s induced by SIV NC were 2-fold lower than those of HIV-1 NC, leading to a very large difference in the DNA compaction energy for the two proteins. Since the N-terminal domain is believed to be a primary determinant for HIV-1 NC’s aggregation properties, this lower $F_c$ could be due to the shorter N-terminal domain of SIV NC, which has one less basic residue, relative to that of HIV-1 NC. In addition, the ZF linker domain of HIV-1 NC has a much higher charge density (five basic residues) than SIV NC (three basic residues), and this is also likely to contribute to the stronger aggregation ability of the HIV-1 protein. DLS measurements to assess NA aggregate size showed that SIV NC produces NA aggregates with a smaller average size and smaller size distribution than HIV-1 NC (Fig. 4.3.7),
in excellent agreement with the single molecule stretching data. Thus, the observed difference in NA chaperone activity of HIV-1 NC relative to SIV NC, although modest, is likely due to stronger aggregation and electrostatic interaction properties of the HIV-1 protein. Moreover, single molecule DNA compaction energy measurements suggested that the differences between the NA interactions of HIV-1 NC and SIV NC may be amplified for NA chaperone functions involving longer NAs.

The smaller slope of the DNA stretching curves observed in the presence of HIV-1 NC compared to SIV NC (Fig. 4.3.5B), reflects less optimal intercalative binding to the DNA duplex. We hypothesize that this may be due to the presence of a Trp residue in ZF1 of SIV NC instead of the Phe present in HIV-1 NC, and that this subtle difference may lead to stronger stacking and intercalation. Interestingly, feline immunodeficiency virus (FIV) NC binding resulted in an even smaller transition slope than HIV-1 NC at all concentrations tested\textsuperscript{186}. FIV NC also has one aromatic residue in each ZF, but the aromatic residue in ZF2 is located on the opposite side of the finger relative to that on SIV NC and HIV-1 NC. These data suggest that both SIV and HIV-1 NC intercalate more strongly than FIV NC, consistent with the non-optimal location of the aromatic amino acid in ZF2 of the FIV protein\textsuperscript{186}.

We also showed using a FA salt-titration binding assay that HIV-1 and SIV NCs interact with a very similar degree of electrostatic vs. specific binding contacts, independent of the RNA examined. Importantly, like HIV-1 NC, SIV NC was capable of interacting with HIV Psi RNA using a more specific binding mode (i.e. lower $K_{d(1M)}$) relative to a non-Psi RNA (HIV-1 TARpolyA) (Fig. 4.3.8). This finding is consistent with the previous observation that SIV proteins are capable of packaging and transducing HIV-1 gRNA\textsuperscript{186}, although a separate study reported that HIV-1 gRNA packaging by SIV GagPol occurred at a reduced efficiency and HIV-
1 gRNA transfer to SIV target cells was not observed\textsuperscript{320}. A recent study showed that both HIV-2 NC and HIV-2 GagΔp6 preferentially bind HIV-2 Psi RNA\textsuperscript{321}. However, the affinity of Gag for the Psi element is greater than that of NC, reflecting contributions from both the NC and MA domains in Gag. Interestingly, the mature HIV-2 MA protein also has NA chaperone activity, but unlike mature HIV-2 NC, is unable to distinguish Psi and non-Psi RNAs. Another recent report consistent with our results found that the dimerization properties of HIV-1 and SIV 5′ leader RNAs are determined by their DIS sequence and not by the identity of the NC protein (HIV-1 vs. SIV) used to induce dimerization\textsuperscript{322}. This led the authors to conclude that the HIV-1 and SIV NC proteins are functionally equivalent in their ability to promote RNA dimerization.

In contrast to binding data with HIV-1 Psi, neither SIV NC nor HIV-1 NC interacted with the putative SIV Psi RNA in the more specific binding mode and the interaction profiles were not readily distinguishable from that of either HIV-1 TARpolyA or SIV TAR (Fig. 4.3.8). HIV-1 Gag was also unable to discriminate between SIV Psi versus TAR RNAs, even though it bound HIV-1 Psi with high specificity. This result is surprising in light of reports showing that HIV-1 Gag/GagPol can package and propagate SIV gRNA\textsuperscript{323,324}. Taken together, these observations suggest that it may be the NA sequence that we have selected as “SIV Psi”, rather than an inability of SIV NC to make specific NA interactions, which is responsible for the lack of observed specificity.

It is important to note that the minimal Psi packaging element has not yet been unambiguously identified in HIV-1\textsuperscript{228,321} or SIV\textsuperscript{233,301,302,313}, as it has for other retroviruses such as MLV\textsuperscript{314,325,326} or Rous sarcoma virus (RSV)\textsuperscript{326,327}. While we derived both HIV and SIV Psi constructs used in this work from gRNA regions that have been shown in genetic experiments to be necessary for efficient genome packaging, additional sequences have been proposed to play a role in HIV-1 \textsuperscript{328}
and SIV 299 gRNA encapsidation. Despite the strong secondary structural homology between HIV-1 and SIV Psi329, additional RNA sequences may be required for SIV Psi to fold into a well-defined packaging signal in vitro. In fact, the SAXS envelope of SIV Psi RNA appears less well defined and differs significantly from that of HIV-1 Psi. This is in contrast to MLV Psi, which has an overall fold that resembles that of HIV Psi303. Alternatively, we cannot rule out the possibility that SIV Gag is required for specific binding to the SIV Psi element. Indeed, we have previously shown that the MA domains of HIV-1330 and RSV230 Gag enhance the specificity of binding to their cognate Psi RNAs.

In summary, the functional similarities between HIV-1 and SIV NC proteins are highlighted in the present work by their ability to interact specifically with HIV-1 Psi RNA and to effectively discriminate HIV-1 Psi vs. non-Psi RNAs such as HIV-1 TARpolyA and SIV TAR, providing additional mechanistic insight into inter-species genomic RNA packaging. Nevertheless, despite the high structural and functional homology, our studies also clearly reveal subtle differences in the NA chaperone functions of HIV-1 and SIV NC proteins that can be explained by differences in their NA aggregation capabilities and DNA compaction energies.

Materials and Methods

DNA oligonucleotides and pIDTSMART vectors were purchased from Integrated DNA Technologies (IDT) (Coralville, IA, USA). [γ-32P]ATP was obtained from PerkinElmer (Shelton, CT, USA). HIV-1 RT was purchased from Worthington (Lakewood, NJ, USA). T4 polynucleotide kinase, proteinase K, SUPERaseIN, and Gel Loading Buffer II were bought from Life Technologies (Foster City, CA, USA). E271 loading dye base was obtained from AMRESCO LLC (Solon, OH, USA). The Ambion MEGAscript T7 kit was purchased from Life Technologies. The sequences of the HIV-1 acceptor RNA and (-) SSDNA, as well as the
TAR and Psi RNAs were derived from HIV-1 NL4-3 (GenBank Accession no. AF324493). The corresponding SIV NAs were derived from SIVmac239 (GenBank Accession no. M33262), which was obtained from Dr. Ronald Desrosiers through the AIDS Reagent Program, Division of AIDS, NIAID, NIH.

**Recombinant NC and Gag proteins**

HIV-1 recombinant NC proteins were expressed in *E. coli* BL21 (DE3) cells and purified as described previously. Essentially the same procedures were used to prepare the SIV NC proteins. Briefly, the DNA regions encoding the 52-amino-acid sequences from SIVmne (Genbank accession no. M32741) or SIVmac239 were cloned into the pET32a expression vector (Novagen, Inc., Madison, WI, USA), expressed in *E. coli*, cleaved from the thioredoxin fusion partner, and purified as described. Note that SIVmne and SIVmac239 NC proteins are identical except for the amino acids at positions 27, 38, and 40: the Mne residues are T27, Q38, and G40, respectively, whereas the Mac239 residues are A27, K38, and D40, respectively. In addition, there is functional identity at position 4, with K for Mne and R for Mac239. The experiments presented here were performed with SIVmne NC. SIVmac239 NC was used for some of the initial minus-strand annealing and strand transfer experiments; however, the results obtained with either SIV NC were the same within uncertainty (data not shown). The charge of each NC protein over a range of pH was calculated using the protein calculator at http://protcalc.sourceforge.net/. The HIV-1 Gag protein lacking the p6 domain (HIV-1 GagΔp6, also referred to simply as “Gag”) was purified as previously described. The concentration of NC in solution was determined by measuring the absorbance at 280 nm using the extinction coefficients 5,680 M⁻¹ cm⁻¹ and 11,560 M⁻¹ cm⁻¹ for HIV-1 and SIV NCs, respectively, and for Gag, using the extinction coefficient 63,090 M⁻¹ cm⁻¹.
Synthesis of viral RNA transcripts

The HIV-1 acceptor RNA (RNA 148) was prepared as described previously. The sequence of the SIV acceptor RNA (RNA 228) consisted of the 52 3′ terminal nt in the unique 3′ region (U3) (beginning at nt 10,184) to the last nt in R (nt 10,411). It was prepared by performing a PCR reaction using the SIV p239SpE3′ plasmid (GenBank Accession no. M33262) obtained from Dr. Ronald Desrosiers through the AIDS Reagent Program, Division of AIDS, NIAID, NIH. The dsDNA product was run on a 2.5% agarose gel and then gel purified prior to transcription with T7 RNA polymerase using the MEGAshortscript T7 kit. The SIV acceptor RNA product was subjected to electrophoresis in a 6% polyacrylamide denaturing gel and then gel purified prior to use in the minus-strand annealing and strand transfer assays.

For the salt-titration and SAXS experiments, DNA template sequences encoding the T7 RNA polymerase promoter sequence followed by the sequences for HIV-1 viral RNAs (TARpolyA and Psi-WT) were prepared as described. DNA template sequences encoding the SIV RNAs (TAR and Psi-WT) cloned into pIDTSMART vectors were obtained from IDT. The HIV-1 and SIV Psi variants with DIS mutated to a GNRA tetraloop sequence (ΔDIS) were generated from the Psi-WT plasmids using the QuikChange Lightning Site-Directed Mutagenesis kit (Agilent Technologies, Santa Clara, CA, USA). RNAs were prepared by in vitro transcription and purified as previously described. Additional non-native G residues were added to all RNAs to facilitate efficient T7-mediated in vitro transcription: HIV-1 TARpolyA contains one additional G residue, HIV-1 Psi contains two additional G residues, SIV TAR contains two additional G residues, and SIV Psi contains two additional G residues. Purified RNAs were fluorescently labeled at their 3′ ends with fluorescein-5-thiosemicarbazide (Invitrogen, Carlsbad, CA, USA) as described. The RNA concentrations in solution were determined by measuring the
absorbance at 260 nm, using the following extinction coefficients: HIV-1 TARpolyA, 935,693 M\(^{-1}\) cm\(^{-1}\); HIV-1 Psi-WT, 973,073 M\(^{-1}\) cm\(^{-1}\); HIV-1 Psi-ΔDIS, 926,348 M\(^{-1}\) cm\(^{-1}\); SIV TAR, 1,113,248 M\(^{-1}\) cm\(^{-1}\); SIV Psi-WT, 1,318,838 M\(^{-1}\) cm\(^{-1}\); and SIV Psi-ΔDIS, 1,281,458 M\(^{-1}\) cm\(^{-1}\).

The extent of labeling with fluorescein was determined by measuring the absorbance at 495 nm and \(\varepsilon_{495} = 85,000\) M\(^{-1}\) cm\(^{-1}\).

**Minus-strand annealing assay**

A 196-nt DNA (DNA 196) (SIV) or a 128-nt DNA (DNA 128) (HIV-1) representing (-) SSDNA (0.2 pmol), labeled at its 5' end with \(^{32}\)P\(^{339,340}\), was incubated at 37 °C with 0.4 pmol of acceptor RNA (RNA 228, SIV; RNA 148, HIV-1) in buffer containing 50 mM Tris-HCl (pH 8.0) and 75 mM KCl and the indicated concentrations of SIV or HIV-1 NC (final volume, 20 µl). Each substrate was tested with the same concentrations of SIV and HIV-1 NC. However, higher concentrations of both NCs were used for the SIV substrates, which were significantly longer than the HIV-1 substrates. The standard reaction was scaled up as needed and 15-µl portions were removed at intervals between 1 and 30 min. Reactions were terminated by addition of sodium dodecyl sulfate to a final concentration of 1% (vol/vol). The mixtures were placed on ice for 5 min and then extracted once with phenol/chloroform. Four µl of loading dye mix containing 12.5% glycerol (vol/vol) and 1x E271 loading dye base were added to 10 µl of the aqueous layer and an 8-µl portion was loaded onto a native 6% polyacrylamide gel prepared with a 4% stacking gel. Analysis of the gel data and calculation of the percent (%) annealed DNA were performed as described previously\(^{341}\). Note that to obtain efficient annealing and minus-strand transfer in the SIV system, the ratio of acceptor RNA to (-) SSDNA normally used in our HIV-1 system \(^{305,342}\) was increased from 1:1 to 2:1. For comparison, identical conditions were used for the HIV-1 system.
**Minus-strand transfer assay**

Reaction mixtures containing reaction buffer (50 mM Tris-HCl (pH 8.0), 75 mM KCl, 1 mM dithiothreitol (DTT), 0.2 pmol (-) SSDNA (DNA 196, SIV; DNA 128, HIV-1) labeled at its 5′ end with $^{32}$P, 0.4 pmol acceptor RNA (RNA 228, SIV; RNA 148, HIV-1), and SIV or HIV-1 NC as specified, were incubated for 5 min at 37°C. HIV-1 RT (1 pmol) and 0.5 units SUPERaseIN were then added and the entire mixture was incubated for another 5 min at 37°C. Reactions (final volume, 20 µl) were initiated by addition of 100 µM each of the four dNTPs, and 1 mM MgCl$_2$. The standard reaction was scaled up as needed. Incubation was at 37°C and 10-µl portions of the reaction mixture were removed at the indicated times. Reactions were terminated by addition of 4 µl of Gel Loading Buffer II. Polyacrylamide gel electrophoresis in 6% denaturing gels and PhosphorImager analysis were performed as described previously$^{336}$. The % strand transfer product formed was calculated by dividing the amount of transfer product by the total signal in the gel lane and multiplying by 100$^{342}$.

**Single molecule DNA stretching experiments**

A biotinylated bacteriophage λ DNA molecule was tethered in between two streptavidin-coated polystyrene beads, torsionally unconstrained from their opposite ends. One bead was held in an optical trap, while the other was immobilized on a micropipette tip attached to a flow cell placed on a translational piezoelectric stage$^{247}$. By gradually moving the fixed bead while recording the extension and the force exerted on the single DNA molecule, the force-extension profile of a dsDNA in the absence of protein was obtained. The buffer surrounding the DNA molecule was then exchanged for a solution of fixed HIV-1 or SIV NC protein concentration to obtain the force–extension curves in the presence of protein at a 100 nm/s pulling rate. The experiments
were conducted in 10 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), 50 mM Na\(^+\) buffer solution at pH 7.5.

**DLS measurements**

DLS experiments were performed using 100 nM SIV Psi RNA in buffer containing 50 mM HEPES (pH 7.5), 5 mM DTT, 1.3 mM MgCl\(_2\), and 20 mM NaCl buffer. NC (1.2 µM) was added to the reaction mix and incubated at room temperature for 30 min prior to DLS measurement on a Zetasizer Nano-ZS instrument (Malvern Instruments Ltd, Malvern, Worcestershire, UK). Data were analyzed using the Dispersion Technology Software provided by the manufacturer, and sizes were plotted as volume percent vs. particle size. The average size of the aggregate population produced was calculated by taking the product of the aggregate volume at each particle size sampled and averaging over the total volume of the population. The average diameter was calculated as the mean of 3-5 measurements with the standard error indicated.

**FA salt-titration binding assays**

The salt-titration binding assays were performed essentially as previously described\(^{115,230}\). Briefly, a fixed concentration of either HIV-1 or SIV NC (400 nM) was incubated with refolded RNA (10 nM) in increasing NaCl concentrations (30 mM to 750 mM) together with 20 mM HEPES (pH 7.5), 20 µM Tris-(2-carboxyethyl)-phosphine (PP), 5 mM 2-mercaptoethanol, and 1 mM MgCl\(_2\). RNAs were refolded in 50 mM HEPES (pH 7.5) by heating at 80°C for 2 min and then at 60°C for 2 min, followed by addition of 10 mM MgCl\(_2\) and incubation on ice for at least 30 min. The reactions were incubated at room temperature in the dark for 30 min and then FA was measured using a SpectraMax M5 plate reader (Molecular Devices, Sunnyvale, CA, USA). Gag salt-titration assays were performed using the same protocol, except that 20 nM RNA was used and the reaction buffer also contained 2 mM Tris-HCl, pH 7.4. To correct for the effect of
increasing NaCl on RNA anisotropy independent of protein binding, separate salt-titration assays of the RNA in the absence of protein were carried out with every trial. The no protein control values were then subtracted from the data obtained for protein-containing reactions. The corrected data were then analyzed as described\textsuperscript{230,315}. Briefly, the dissociation constant $K_d$ varies as a function of Na\textsuperscript{+} ion concentration as follows:

$$K_d = K_{d(1M)} \cdot [Na]^{Z_{\text{eff}}}, \quad (4.2.1)$$

In Eq. 4.2.1, $K_{d(1M)}$ is the dissociation constant of the RNA-protein interaction at 1 M NaCl when all electrostatic charges have been screened out, thereby reflecting the strength of the non-electrostatic binding contacts. $Z_{\text{eff}}$ represents the number of electrostatic contacts involved in the interaction. Substituting Eq. 4.2.1 into the binding isotherm as previously described\textsuperscript{315} allows determination of the two parameters, $K_{d1M}$ and $Z_{\text{eff}}$.

**Preparation of RNA for SAXS analysis**

SIV Psi-\textDelta DIS RNA (450 µg) was refolded as described above, except that an additional step of incubation at 37 °C for 5 min was added between the addition of MgCl\textsubscript{2} and incubation on ice. The folded RNA was then purified via SEC on a 24-ml Superdex 200 10/300 GL Increase column (GE Healthcare, Little Chalfont, Buckinghamshire, UK) in running buffer containing 150 mM NaCl, 50 mM HEPES (pH 7.4), 1 mM MgCl\textsubscript{2}, and 3% glycerol (wt/vol) at a flow rate of 1 ml/min. Peak fractions containing the desired RNA were pooled and concentrated to 70-90 µl using an Amicon 0.5-ml 10K molecular weight cutoff spin concentrator (EMD Millipore, Bellerica, MA, USA). Sample concentrations ranged from 3.0-3.7 µg/µl. The SEC running buffer was used to serially dilute the RNAs, yielding three sample concentration ranges (3.0-3.7 µg/µl, 1.5-1.9 µg/µl, and 0.75-0.93 µg/µl). An aliquot of the SEC buffer used to purify the RNAs was saved for use in SAXS for buffer subtraction.
SAXS data acquisition and analysis

Samples were shipped in 96-well plates (Axygen Scientific, Union City, CA, USA) at 4 °C to the 12.3.1 SIBYLS beamline at the Advanced Light Source (Lawrence Berkeley National Lab, Berkeley, CA, USA)\textsuperscript{230,343}. Scattering data were acquired and buffer subtraction was performed by the SIBYLS beamline staff as described\textsuperscript{344}. Subsequent data analysis and \textit{ab initio} envelope generation was performed largely as previously described\textsuperscript{345}. Briefly, the SAXS data collected at different exposure times for a given concentration of sample were examined separately and exposures with clear evidence of radiation damage were discarded. High quality exposures for each RNA concentration dilution were merged and then analyzed using Guinier analysis\textsuperscript{231} to calculate the radius of gyration ($R_g$) and the extrapolated scattering intensity at zero scattering angle ($I_0$) using the program PRIMUS\textsuperscript{343}. Kratky analysis\textsuperscript{346} was also performed for each RNA concentration dilution to confirm the extent of folding. If the $R_g$ was found to increase upon increasing RNA concentration (indicative of concentration-dependent effects) or if the Kratky plots suggested that the RNA was not well folded, the data were not analyzed further. If samples passed these quality control analyses, then the data sets from the three concentration dilutions were scaled and merged into a single curve. The inter-electron P(r) functions were calculated using the program GNOM\textsuperscript{343}. The maximum inter-electron distance ($D_{\text{max}}$) was varied until the P(r) decayed smoothly to zero and the experimental data fit well. The $D_{\text{max}}$ was increased by 2 Å increments up to 15 Å above the $D_{\text{max}}$ initially selected by the GNOM program. Then for each of these $D_{\text{max}}$ values, five \textit{ab initio} envelopes were generated in fast mode with no symmetry restraints imposed using the ATSAS suite of programs as described\textsuperscript{231,346}. The average $\chi^2$ fit of the five envelopes to the experimental data was determined and the $D_{\text{max}}$ condition that gave the best fit was chosen for further analysis. Using this $D_{\text{max}}$, 20 \textit{ab initio} envelopes were generated using the same protocol as described above, and the $\chi^2$ fits and reproducibility (NSD) values...
were calculated. These 20 envelopes were averaged into one envelope, which was then packed with at least 20,000 “dummy atoms” and used as the starting point for an additional 24 \textit{ab initio} envelope calculations, generated in jagged mode with no symmetry restraints imposed. These envelopes were averaged to generate the final envelope and their $\chi^2$ fits and NSD values were determined.
Chapter 5: Conclusions and future work

Conclusions

Single molecule force spectroscopy is a well-established technique that utilizes mechanical manipulation on biological systems to dissect molecular mechanisms. Here, single molecule DNA stretching studies are employed to probe the complex nucleic acid interactions of replication proteins from *E. coli*, HIV-1 and SIV retroviruses, and the L1 retrotransposon, using optical tweezers.

Information on the detailed molecular mechanism of DNA replication by the *E. coli* DNA pol III not only sheds light on the most fundamental interaction in process of life, but also may aid in the rational development of novel targeted antibiotics. We characterize the catalytic functions of DNA pol III in the context of its core subassembly using single molecule force spectroscopy. While revealing key characteristics of its activity, we significantly enhance the understanding of the mechanisms by which the catalytic polymerization and proofreading are regulated. Taken together, our results support a model in which the pol and exo activities of pol III core are effectively independent and the stability of the primer-template junction determines the selection between α and ε binding to the 3’ end of the primer strand. Once pol III binds in a pol- or exo-competent conformation, enzymatic activity proceeds with relatively high velocity of 10-100 nt/s. Despite these high catalytic rates of the pol- and exo- domains, the processivity of pol III core remains low at 10-20 nt, as both catalytic events occur via short ~0.2 s bursts. These bursts are interrupted by pol III core dissociation from the primer-template junction. Re-initiation of exo activity requires rebinding of pol III to the primer-template junction from solution, followed by the slower melting of ~2 bp at the primer terminus. This is then followed by the much faster
transfer of the 3’ end of the primer strand to the pol III exo site. The force dependence we measure for exo initiation closely matches the previously measured temperature dependence of the same process\textsuperscript{346}. This result supports a model in which mismatch recognition during proofreading is determined by primer template duplex end stability, rather than a duplex defect recognition model\textsuperscript{111}. Overall, the independent nature of the pol and exo states as well as frequent dissociation of pol III from its substrate are expected to allow stronger regulation of these processes by other factors involved in \textit{E. coli} replication.

Although not considered a part of the multiprotein replisome, \textit{E. coli} single-stranded DNA binding protein (EcSSB), plays crucial roles during DNA replication. As the replication fork proceeds, the long stretches of transiently formed ssDNA in the lagging strand is rapidly sequestered and protected by EcSSB, amongst other potential functions it possesses. EcSSB is a homotetramer that can wrap ssDNA with multiple topologies. Here, we study the competitive binding of EcSSB to a long ssDNA molecule. For the first time, we reveal the biphasic binding mechanism of EcSSB that regulates its own modes depending on the ssDNA occupancy in the binding lattice. This study also reveals the existence of a minimally wrapped EcSSB mode (<<8 nt) that remains stable even when the substrate DNA is held at 20 pN. We investigate the competitive interconversion between the wrapped and unwrapped modes of EcSSB as a function of protein concentration. When free protein in the solution is scarce, EcSSB primarily exists in the (SSB)\textsubscript{17} wrapped state. However, as the free protein concentration increases the unwrapped mode becomes dominant. Furthermore, we show that while EcSSB binding events facilitate the unwrapping of wrapped SSB-ssDNA modes, active wrapping of ssDNA by EcSSB stimulates the dissociation of the bound but unwrapped EcSSB. Finally, we propose a general two-state kinetic model, that satisfactorily describes the observed results.
L1 retrotransposon is the dominant transposable element in mammals that is responsible for the make of more than 40% of the human genome. L1 retrotransposition is a unique replication process of which a complete understanding is vague, especially with respect to ORF1p, one of the two proteins it encodes. Here we employ single molecule DNA stretching assays on human and primate ORF1p to determine its DNA interaction kinetics in the context of its previously measured oligomerization properties\textsuperscript{66,112}. Using equilibrium ORF1p binding to a force-induced ssDNA lattice we show that ORF1p forms stably bound oligomers on ssDNA, with more stably bound oligomers at longer incubation time. Three ORF1p variants are compared; modern human ORF1p, its resuscitated primate ancestor and a mosaic variant of the modern protein in which 9 of the 30 substitutions in the coiled coil domain retained their ancestral state. Although these proteins form equally stable trimers, behave indistinguishably in bulk binding and \textit{in vitro} nucleic chaperone assays, the mosaic 151p is completely inactive in an \textit{in vivo} retrotransposition assay. Single molecule analysis is used to determine the rate of ORF1p oligomerization on ssDNA. The results show that both retrotransposition-competent ORF1p proteins (i.e, the modern human protein and its resuscitated ancestral counterpart) rapidly form stably-bound oligomers on single-stranded DNA. In contrast, the retrotransposition-defective mosaic ORF1p forms such stably bound oligomers at a rate at least 10-fold slower. This property is a function of the coiled coil sequence, presumably achieved by positioning the carboxy-terminal halves of ORF1p in a way that is conducive to the inter-trimer contacts that mediate oligomerization. Other than trimer formation per se (which is necessary but clearly is not sufficient for retrotransposition), this study reveals the first functional role of the ORF1p coiled coil domain in L1 retrotransposition.

In an extended study we probe ORF1p binding dynamics to long preformed ssDNA. Surprisingly, ORF1p qualitatively mimics the biphasic binding mechanism of EcSSB discussed above.
However, unlike EcSSB, ORF1p binding is stable at much higher forces (>>20 pN). Similar to EcSSB, ORF1p also exists as a multimer (trimer) and hence this result provides compelling evidence for the ability of ORF1p to wrap the ssDNA with multiple topologies. Moreover, we show that inter-trimer interactions of ORF1p further compact ssDNA, resulting in a well-structured and packaged ssDNA-ORF1p complex. This study reveals the mechanism by which ORF1p packages its own transcript to form the L1 RNP, an essential intermediate of L1 retrotransposition, thereby significantly enhancing the understanding of ORF1p function in L1 replication cycle.

Dissecting the molecular mechanisms involved in the retroviral life cycle is a key to developing effective anti-viral drugs. Retroviruses carry their genetic code in the form of an RNA strand, which is reverse transcribed into DNA before it is integrated into the cellular genome. The necessary nucleic acid rearrangements to achieve this process are facilitated by the retroviral NC proteins. To elucidate the functional roles of the basic residues in the N-terminal domain, the N-terminal zinc finger, and the linker region between the two zinc fingers in the HIV-1 NC, we characterized the binding properties of NC mutants with respect to the wildtype protein. A pentamutant with 5 N-terminal basic residues substituted to alanine (K3A/R7A/R10/K11A/K14), an N-terminal trimutant with only 3 basic residues substituted to alanine (R7A/R10A/K11A), and two additional basic residue mutants that containing 3 (K14A/K20A/K26A; zinc finger 1 trimutant) or 4 (R29A/R32A/K33A/K34A; zinc finger linker mutant) alanine substitutions were characterized. With respect to the wildtype protein, all the mutants were compromised in terms of the ability to increase the overstretching transition force or the stretching slope of the DNA, which is known to positively correlate with nucleic acid chaperone capabilities. The ability to alter the stretching slope was weakest for the pentamutant, followed by the zinc finger linker mutant, the zinc finger 1 trimutant, and finally the N-terminal trimutant, which was closest to
wildtype NC activity. However, while these effects increase with the number of residues mutated, the zinc finger 1 and N-terminal trimutants exhibit significantly different effects, indicating the criticality of the positional roles of these basic residues.

Furthermore, we characterized the nucleic acid chaperone properties of the NC protein from SIV, the origin of HIV-1. Despite an overall functional similarity, quantitative differences in nucleic acid aggregation and compaction capability distinguish the two proteins, which may be relevant to the infection process. Single molecule analysis showed that the compaction force for HIV-1 NC was ~2-fold higher than that of SIV NC. The additional compaction force for HIV-1 NC relative to that of SIV NC, averaged over both concentrations and weighted by uncertainty, was $1.3 \pm 0.4$ pN. This corresponds to a difference in compaction energy of $0.11 \pm 0.03$ kBT/bp. Thus, for a 10 kbp dsDNA molecule, similar to the length of the HIV-1 genome, the additional compaction energy for HIV-1 NC is $1100 \pm 300$ kBT, or $640 \pm 170$ kcal/mol, which is a very large energy difference for a molecular process. In conclusion, we demonstrate that HIV-1 NC is a slightly more efficient NA chaperone protein than SIV NC.
Future work

**Pol III activity in the presence of its processivity clamp, β:** The processivity of pol III core is strongly enhanced with its processivity clamp, β, a ring-shaped dimer that tethers the protein complex to the DNA template\(^{160}\). It will be interesting to test how β influences the activity of the proofreading subunit, ε. In addition, mechanochemical characterization of the pol III core- β complex, and pol IV, a DNA damage response polymerase, will provide a means to study these polymerases together to probe the polymerase switching mechanisms.

**Characterization of EcSSB mutants:** Although we observe that EcSSB exhibits dynamic interconversion between its distinct binding modes depending on ssDNA occupancy, the underlying molecular mechanism of this competitive binding behavior is not completely understood. Previous studies\(^{21,22}\) have shown that the C-terminal intrinsically disordered loop (IDL) influences inter-tetramer interactions. It will be informative to study the ssDNA binding dynamics of the C-terminal truncated/mutated EcSSB (among other mutants) to test if inter-tetramer interactions influence competitive interconversions.

**Characterizations of SSBs from other organisms:** SSBs are found in many organisms and known to serve common functional roles, although they are structurally different. For instance, the eukaryotic RPA is a heterotrimer\(^{101}\) and the SSB from *Thermus aquaticus* (Taq SSB) is a homodimer\(^{347}\). Characterization and comparison of the ssDNA-binding dynamics between different SSBs such as RPA, and Taq SSB, will shed light to the overall SSB function.

**Experiments to test SRPD:** The self-regulated protein-density mechanism proposed in this dissertation (chapter 2.3) can be tested with a growing RecA-filament in an EcSSB-bound ssDNA (in the absence of free EcSSB). Alternatively, a DNA polymerase such as T7DNAp can
be used instead of RecA to displace EcSSB. Optical tweezers incorporated with fluorescence capabilities (C-Trap) will provide a great advantage for these studies by allowing one to visualize the changes in protein-density on ssDNA.

**Experiments to test the functional roles of different ORF1p-ssDNA binding states:** To test the functional roles of ORF1p binding states, one can test if the formation of higher order multimers reduces the interconversion dynamics of the ORF1p states. This can be probed by investigating the interconversion dynamics as a function of ssDNA incubation time with ORF1p, at constant force.

**ssDNA binding dynamics with ORF1p variants:** Characterization of ssDNA binding dynamics with ORF1p variants will allow one to correlate its mechanistic nucleic acid interactions to its *in vivo* activity. In addition, these experiments will provide information on structural properties that affect its ssDNA-binding characteristics.

**ORF1p-nucleic acid interactions in the presence of A3 proteins:** The mechanism by which the A3A, A3B and A3F proteins exhibit inhibitory activity against LI retrotransposon elements is poorly understood. A3 proteins are also known to bind ORF1p. Probing the nucleic acid-binding properties of ORF1p in the presence of A3 proteins will provide insights on whether these inhibitory effects are mediated by A3-ORF1p interactions.

**ssDNA binding dynamics of retroviral NCp and Gag polyprotein:** Single molecule studies of retroviral nucleocapsid protein (NC) with preformed ssDNA will provide means to isolate and characterize its ssDNA binding properties. In addition, similar experiments with HIV-1 Gag polyprotein will shed light into its nucleic acid packaging properties.
Publications

1. **Naufer, M. N.,** Morse, M., Rouzina, I., Beuning, P. J., and Williams M. C., Self-regulation of Single-Stranded DNA Wrapping dynamics by E. coli SSB mediated through competitive binding (in preparation)

2. **Naufer, M. N.,** Morse, M., Rouzina, I., Furano, A.V., and Williams, M.C., LINE1 ORF1p wraps and packages single-stranded DNA (in preparation)

3. Morse, M., **Naufer, M. N.,** Rouzina, I., Chelico, L., Williams, M. C., APOBEC3G binds single-stranded DNA in multiple steps and conformations to search and deaminate viral DNA (in preparation)


10. **Naufer, M. N.,** Callahan, K. E., Cook, P. R., Williams, M. C., Furano, A. V., L1 retrotransposition requires rapid ORF1p oligomerization, a novel coiled coil-dependent property conserved despite extensive remodeling. Nucleic Acids Research (2015)

Appendices

Appendix A: UTC, a complete solution for optical tweezers data acquisition

I developed the Universal Tweezers Central (UTC), named by Mark Williams, a LabWindows data acquisition software based on a previously designed software by Mark Williams and Micah McCauley. Feedbacks and suggestions from Mike Morse significantly contributed to improve the image acquisition module. For More information and source code visit: https://github.com/nabuan/UTC

The significant improvements in the new software includes the digital control of the piezo stage, and the image acquisition capabilities achieved with multi-threading techniques.

Figure A1. Main controller panel of the UTC. The figure shows a screen shot during a DNA intercalation experiment. The Main control panel is used to setup all the data acquisition parameters such as extension-step sizes, delay between each steps, force-clamp information. In addition the image acquisition model can be used to set up the image capturing frequency. Force-clamp experiments can be manipulated real-time to switch between forces. Before saving the data acquired, the user can imput information about an experimental run on the ‘save details’ module (right) such as, the protein type and concentration and comments, that will automatically be saved along with all the data acquisition parameters, as supplemental detail information for a given run.

The software drives a piezo electric stage (nPoint) using pulse frequency modulation and interprets the piezo position using a differential quadrature encoder. Digitally interfacing the piezo electric stage improved the spatial resolution of the optical tweezers from 0.8 nm to 0.1 nm. In addition, simultaneous image capturing allowed us to correct for the long-term thermal drifts, which in turn made it possible to precisely capture subtle extension-changes involved in protein-DNA interactions. UTC contains 4 panels; the main controller, extension-time, camera controller, and camera view. The main controller (Fig. A1) allows the user to configure an experimental run while displaying real-time data acquisition information such as piezo position.
and laser deflection. The extension-time panel displays the real-time extension of the tethered DNA as a function of time. The camera-control panel allows the user to switch between the two cameras and control the camera attributes, such as gain. And the camera view panel displays the field of view of the flow-cell channel. The software utilized multi-threading capabilities to drive the camera in a separate thread to ensure that the image acquisition is independent from the rest of the modules. The user can choose to acquire images simultaneously with stage-position and laser deflection data, at a desired frequency. The metadata of the acquired images are stored in the raw data file with time-stamps, which allows us to compare extension information acquired from the stage and the images.

![Extension-time panel of UTC](image)

**Figure A2. Extension-time panel of UTC** Extension and time are plotted in y and x axis, respectively. The snapshot shown here was taken during an experiment with DNA intercalators. This simple and newest addition continuously displays the extension as a function of time during force-clamped experiments. Extension-time information updates the user on the current kinetic state and thereby allowing to manipulate parameters during the data acquisition.

![Camera controller and view](image)

**Figure A3. The camera controller (right), camera view (left) of UTC.** The camera controller allows the user to configure multi-threading priority parameters and switch between the two cameras.
Appendix B: iFextA, a complete automated solution incorporated with image processing for high-precision optical tweezers data calibration

I developed a MATLAB-based data calibration software (Fexta) in 2014 to fully automate the post data processing. Later with the enormous support from Rohan Gala I significantly improved the algorithms and incorporated the image analyses module. The software was continuously improved with invaluable feedback and ideas from Mark Williams, Micah McCauley and Mike Morse. For more information and source code visit: https://github.com/nabuan/Fexta2016 and https://github.com/nabuan/imFexta

Raw data acquired from UTC contains information on the positions of the stage ($x_s$), and the laser deflection ($x_d$) due to the bead displacement on the optical trap. Therefore, the raw data requires to be calibrated in order to obtain the absolute end-to-end distance between the tether points of the DNA molecule and to convert the laser deflection into pN. The extension-force profile of a dsDNA molecule is well-behaved in a given monovalent salt concentration\(^{348}\). Here the program utilized the raw data of a dsDNA molecule and the bead displacement as a function of laser deflection; information that essentially is a measure of the trap-stiffness, and the expected force-extension profile of a dsDNA in a given monovalent salt concentration, to find offsets in force and extension, and conversion factor to convert $x_d$ in A.U. to pN. The empirical force induced melting midpoint\(^{64}\) at a given [Na\(^+\)] and the theoretical extensible worm like chain model (WLC) are the constraints used to calibrate data obtained for a dsDNA molecule.

First, the midpoint of the plateau region in the stiffness corrected dsDNA curve is determined using a heuristic algorithm. This value of laser deflection, $x_d^{melting}$, corresponds to a known value of force, $F^{melting}$. This is used to determine an offset by which $x_d$ is shifted.

$$F^{melting} = x_d^{melting} + y_{offset}$$

The extensible WLC model describes the extension ($b_{ds}$, in nm/bp) of a dsDNA as a function of force:

$$b_{ds}^{WLC}(F) = B_{ds} \left( 1 - \frac{1}{2} \left( \frac{k_B T}{FP_{ds}} \right)^{1/2} + \frac{F}{S_{ds}} \right)$$

Here, $S_{ds}$ is the stretch modulus (1361 pN), $P_{ds}$ is the persistence length (45 nm), and $B_{ds}$ is normalized contour length (0.34 nm/bp).

$b_{ds}$ is related to $x_s$ in the double stranded portion of the data through an unknown $x_{offset}$ as follows:

$$b_{ds} = \frac{x_s - x_d}{N_{bp}} + \tilde{x}_{offset}$$

In addition, we assume that $F$ is related to $x_d$ through a single stretch parameter $c$ as follows:

$$F = c \left( x_d + y_{offset} \right) F^{melting} + F^{melting}$$
Figure B1: A representative screen shot from a data analyses example from iFexta. The analyses process is done in the following steps. (1) chose the file containing the raw stiffness data (2) input the number of bp in the dsDNA and the salt concentration at which the dsDNA data was obtained. Chose ‘Analyze DNA’. The program will then process the calibration, display the calibration parameters and the calibrated dsDNA (blue circles in the right panel). (3) The user may input the data obtained from the same DNA to calibrate the subsequent experimental data conducted on the same DNA molecule. In this example the exonculeolysis data of T7DNAp is analyzed. The force-extension is shown in red on the left panel, and the extension-time (blue) and the force-time (green) are displayed in the right panel in blue.

These parameters are determined by assuming that a portion of the data matches the WLC model. Therefore, we can formulate the following optimization required to find the offsets in extension ($x_{offset}$), force ($y_{offset}$) and the stretching parameter ($c$).

$$y \rightarrow F - F_{melting} = c \left( x_d + y_{offset} - F_{melting} \right)$$

$$\arg \min_{x_{offset}, c} \left\| \frac{x_d - x_r}{N_{bp}} + \tilde{x}_{offset} \right\| - B_{ds} + \frac{B_{ds}}{2} \left( \frac{k_B T}{(c + F_{melting}) P_{ds}} \right)^{\frac{1}{2}} - \frac{B_{ds}}{S_{ds}} \left( c + F_{melting} \right) \right\|$$

$$\arg \min_{x_{offset}, c} \left\| \frac{x_d - x_r}{N_{bp}} + \tilde{x}_{offset} \right\| - B_{ds} + \frac{B_{ds}}{2} \left( \frac{k_B T}{(c + F_{melting}) P_{ds}} \right)^{\frac{1}{2}} - \frac{B_{ds}}{S_{ds}} \left( c + F_{melting} \right) \right\|$$

$$\frac{B_{ds}}{S_{ds}} \left( \frac{c}{c + F_{melting}} \right)$$

186
Because the calibration parameters remain the same for a given tethered DNA molecule, we use these values to calibrate the subsequent data acquired using the same DNA molecule. The software detects and separates the force-extension data, and extension-time data, and provide the user with the option to automatically save those as a text file in the same folder where the raw data is located.

**Image analysis**

The raw extension data is the position of the stage, which we calibrate using the force-extension profile of a dsDNA molecule to find the offset and obtain the end-to-end extension between the DNA tether points. However, this offset is prone to change over time, due to thermal drift of the micropipette tip. Information from the image of the two beads can be directly used to measure the absolute distance between the bead centers and thereby determine the end-to-end distance of the tethered DNA molecule. This allows us to perform, long-term experiments with a tethered DNA molecule. For instance, to probe ssDNA-binding dynamics, one needs to first digest a strand to prepare the ssDNA substrate before performing the experiment which may consume a considerable amount of time between obtaining the FEC of a dsDNA and generating an ssDNA to perform the experiment and thus the stage offset determined by the dsDNA curve may no longer be accurate. In addition, during force-clamp experiments, the pipet-drifts may introduce artifacts in the data that are impossible to detect without the information of the absolute end-to-end extension. Therefore, I developed the image analysis module to determine the absolute DNA-extension to deconvolve the data from possible artifacts. Comparison between the extension-changes measured from images and obtained from the stage calibration allows us to eliminate artifacts from drifts while preserving the high-resolution data obtained from the stage.

The bead centers are determined by a cross correlation method that uses a segment of the image containing each bead as a template. Therefore, the bead detection does not depend on the variables such as bead-size or image intensity. The image analysis module finds two parameters to calibrate the image information. First by comparing corresponding step sizes of the stage (in nm) and the images (in pixels), the pixel to nm conversion factor is calculated. This factor is determined by the pixel resolution and the field of view size of the camera, and therefore will not change with DNA molecules. Because the cross-correlation method provides the distance between the two bead centers, we are also required to find the total offset, which is the sum of the two distances from the center of a given bead to the DNA tether point on that bead. This offset will change for each DNA-tether but will not change for a given captured DNA molecule. Because, the extension profile of the dsDNA is accurately calibrated, we use this information to find the offset to the tether points.
**Figure B. Image analysis module.** An example image correction of the extension-time profile obtained from a DNA intercalation experiment. The template images containing each bead, as chosen by the user, are shown on the top left panel. The bottom left shows the bead detection. The 'calibration with DNA' function will cross-examine the image data with stage data for the same dsDNA molecule and determine the pixel to nm conversion factor and the offset due to DNA tether points as displayed in the corresponding text fields. The blue curve on the right shows a long-term experimental data of extension-change as measured by the stage. The red data shows the absolute end-to-end distance as determined by the image information. The discrepancy between the data is indicative of long term drift, which is shown in with yellow data points. The stage data is then corrected for the drift.
Appendix C: pBacgus-11, 3′ recessed, digoxigenin and biotin labeled 8 kbp DNA construct

The protocol for the linearized DIG and Biotin labeled 8.1 kbp DNA construct was designed in collaboration with the Beuning Lab. The plasmid was a generous donation by Borja Ibarra and was amplified by Dave Murison in the Beuning lab.

The 8.1 kbp is a relatively shorter (5-fold than the traditionally used 38.5 kbp λ DNA) construct that enhances the S/N ratio of the polymerase trajectories as measured in the optical tweezers system. The same construct also allows us to efficiently and reliably generate ssDNA in situ through exonuclease digestion, and thereby to probe protein binding-dynamics with preformed ssDNA. Here, a biotinylated oligonucleotide and a multiple-DIG containing dsDNA handle is ligated to the terminals of a linearized pBACgus-11 vector.

**Generating DIG handle:**

1. Amplification of pUC19 (NEB, #N3041S, 50 μg, 1000 μg/mL) region between nucleotides 337-434 using Taq DNA Polymerase (Thermo, #EP0402, 500 U, 5 U/μL). *** Digoxigenin-11-dUTP (DIG-dUTP) used must be alkali-stable (Sigma Aldrich, #11093088910, 25 nmols, 25 μL, 1 mM) ***

   **Primers:**
   - pUC-DIGPrimer1a GGGATGTGCTGCAAGGCGATTAAG (forward)
   - pUC-DIGPrimer2 CTGCAGGTCGACTCTAGAGGATC (reverse)

2. Perform two amplifications for two linearized/gapped pBACgus-11 reactions. Mix reagents for 50 μL PCR amplification in thin-walled PCR tubes:

   - 19.2 μL Nuclease-free water
   - 5.0 μL 10x Taq Buffer (either KCl+ or (NH₄)₂SO₄+)
   - 2.0 μL 5 mM dATP, final [dATP] = 200 μM
   - 2.0 μL 5 mM dCTP, final [dCTP] = 200 μM
   - 2.0 μL 5 mM dGTP, final [dGTP] = 200 μM
   - 1.3 μL 5 mM dTTP, final [dTTP] = 130 μM
   - 3.5 μL 1 mM DIG-dUTP, final [DIG-dUTP] = 70 μM
   - 4.0 μL 25 mM MgCl₂, final [MgCl₂] = 2 mM
   - 0.5 μL pUC19 Vector (diluted 1:1000), 1ng/μL, [pUC19] = 10 pg/μL
   - 5.0 μL Forward Primer (5μM), [final] = 500 nM
   - 5.0 μL Reverse Primer (5μM) [final] = 500 nM
   - 0.5 μL Taq Polymerase
B. Mix well, spin down, run protocol [AMPLIFYD] on white PCR

<table>
<thead>
<tr>
<th>Step</th>
<th>Temp (°C)</th>
<th>Time</th>
<th>Number Cycles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Denaturation</td>
<td>95</td>
<td>3 min</td>
<td>1</td>
</tr>
<tr>
<td>Denaturation</td>
<td>95</td>
<td>30 sec</td>
<td></td>
</tr>
<tr>
<td>Annealing</td>
<td>63 (T_m-5)</td>
<td>45 sec</td>
<td>35 cycles</td>
</tr>
<tr>
<td>Extension</td>
<td>72</td>
<td>1 min</td>
<td></td>
</tr>
<tr>
<td>Final Extension</td>
<td>72</td>
<td>10 min</td>
<td>1</td>
</tr>
</tbody>
</table>

C. Perform PCR Clean-up on amplification reaction to remove polymerase, nucleotides, and primers. Elute in 40 μL Elution Buffer.

Preparing gapped, linearized pBACgus-11 DNA:

1. Linearize pBACgus-11 plasmid with double-digest using BamHI-HF enzyme (NEB, #R3136S, 10,000 units, 20 U/μL) and SacI-HF (NEB, #R3156S, 2,000 units, 20 U/μL).

A. Add to two thin-walled PRC tubes:
   - 162 μL Nuclease-free water
   - 20 μL 10x CutSmart
   - 2 μL pBACgus-11 #1 (4 μg/μL, total ≈ 8 μg)
   - 8 μL SacI-HF Enzyme (20 units/μL, Total = 160 Units)
   - 8 μL BamHI-HF Enzyme (20 units/μL, Total = 160 Units)

B. Split the 200 uL solution into 4 50 uL solutions in PCR tubes.
C. Mix the 50 μL solutions gently by flicking. Spin down.
D. Run PRC protocol [BamSac] on white thermocycler 37 °C for 16 h
   Hold @ 4 °C.

E. Perform PCR clean up all-in-one on completed reactions: Elute in 60 μL Elution Buffer.
**Digest of DIG amplification product:**

1. Amplification products should be cleaved using BamHI-HF restriction enzyme (NEB, #R3136S, 10,000 units, 20 U/μL) to create sticky ends complementary to the 5’ on linearized pBACgus-11. Products of amplification reaction using Primer1a (forward) and Primer2 (reverse) should be 121bp. After BamHI digest, fragments should be 98bp + 4nt overhang, and 19bp + 4nt overhang.

   A. Transfer to 40 μL of cleaned amplification reaction to thin-walled PCR tube and add:
      5 μL  10x CutSmart Buffer
      2 μL  BamHI-HF Enzyme (40 U)
      3 μL  Nuclease-free water
   B. Spin down and incubate in 37 °C water bath for 90 mins.
   C. Perform PCR Clean-up on digest reactions:
      Remove restriction enzyme and smaller, cleaved DNA fragment
      Elute in 15 μL *Elution Buffer*.

**Ligation of DIG-handle and biotinylated oligonucleotide to gapped, linearized pBACgus-11**

1. Functionalize gapped, linear pBACgus-11 with digoxigenin-labeled PCR handle (DIG-handle) and biotinylated oligonucleotide (synthesized by IDT, resuspended in 1x TE Buffer to 100 µM). Ligate using T4 DNA Ligase (NEB, #M0202S, 20,000 Units, 400 U/μL).

   A. Transfer cleaned, gapped, linear pBACgus-11 reaction and half of cleaned DIG-Handle digest reaction to thin-walled PRC tubes:
      15 μL  Cleaned, gapped, linear pBACgus-11
      15 μL  Cleaned DIG-Handle digest reaction
      1 μL  100 μM SacI Biotinylated Oligonucleotide (100 pmols)
      4 μL  10x T4 DNA Ligase Buffer
      2 μL  T4 DNA Ligase Enzyme (800 Units)
      3 μL  Nuclease-free water
      Mix by flicking and spinning down.
   B. Incubate over night at 16 °C in thermocycler. Elute and store in 40 μL *Elution Buffer*. 
   
191
Appendix D: Permission details for reprinting

Figure 2.1.1: DNA polymerase III holoenzyme (dimer form) at a replication fork.

License Number: 4550361429850
License date: Mar 15, 2019
Licensed Content Publisher: Springer Nature
Licensed Content Publication: Nature Structural & Molecular Biology
Licensed Content Title: Single-molecule studies of fork dynamics in Escherichia coli DNA replication

Figure 4.1.1: Retroviral Life cycle.

License Number 4550351240752,
License date Mar: 15, 2019
Licensed Content Publisher: Springer Nature
Licensed Content Publication: Nature Reviews Microbiology
Licensed Content Title: Studies of endogenous retroviruses reveal a continuing evolutionary saga
References


195


Naufer, M. N., Callahan, K. E., Cook, P. R., Perez-Gonzalez, C. E., Williams, M. C. & Furano, A. V. L1 retrotransposition requires rapid ORF1p oligomerization, a novel coiled


Godet, J., Ramalanjaona, N., Sharma, K. K., Richert, L., de Rocquigny, H., Darlix, J. L., Duportail, G. & Mely, Y. Specific implications of the HIV-1 nucleocapsid zinc fingers in


Para


