UNIVERSITY OF CALIFORNIA SAN DIEGO

Studies of Motor-Driven Viral DNA Packaging by Single DNA Manipulation with Optical Tweezers

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Physics

by

Youbin Mo

Committee in charge:

Professor Douglas E. Smith, Chair

Professor Alexander Groisman

Professor Judy E. Kim

Professor Zhaowei Liu

Professor Mark L. Paddock

2021

The Dissertation of Youbin Mo is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

I dedicate this dissertation to my parents Yinfu Mo, Juanming Guo.

EPIGRAPH

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way—in short, the period was so far like the present period, that some of its noisiest authorities insisted on its being received, for good or for evil, in the superlative degree of comparison only.

'A Tale of Two Cities' Charles Dickens

莫愁前路无知己，天下谁人不识君?

*《别董大》*高适

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge my parents Yinfu Mo and Juanming Guo, and my family. They provided a lot of support and encouragement during I grew up. As the first tutor in my life, my parents taught me to live enthusiastically and to work actively. When I was struggled with difficulty, their word was always the best placebo which supported me to fight to it. Besides, I would like to acknowledge my grandparents who encouraged me to complete the doctoral degree in UC San Diego.

I would like to acknowledge Professor Douglas E. Smith for his support as the advisor of my Ph.D. career and the chair of my committee. His guidance has proved to be invaluable. I really appreciated the freedom in Smith's lab that we could develop our own projects. Doug also helped me on publishing my research sparing no effort.

I would like to acknowledge our collaborators Prof. Venigalla B. Rao and Dr. Neeti Ananthaswamy in the Catholic University of America, and Stephen C. Harvey in the University of Pennsylvania. They offered us the material used for phage packaging experiment. And they also provided suggestions and comments for our publications.

I would like to acknowledge our lab members Damian del Toro, Nicholas Keller and Mariam Ordyan. They taught me the operation of optical tweezers and shared their knowledge of viral packaging experiments. Without them I could not conduct any data measurements. And I would like to acknowledge Mounir Fizari and Kristina Koharchik who helped me collected the data.

I would like to acknowledge my friends Songyue Qian, Huan Hu and Bin Ren who brought a lot of happiness to my life. And I would like to acknowledge Martha Stacklin who helped me adapting to the life in San Diego.

And the most important person I would like to acknowledge is my girlfriend Chenliang Zhang (Sherry). It has been almost 6 years since we fell in love in 2015. We have seen the Aurora in Alaska and the Inca ruins at Cancun, together. These wonderful and sweet memories are the most beautiful experience in the past 7 years. You encouraged me to learn new stuff and try a new direction of my career. I cannot find my destination of my life without you. Although we are living in two countries and we have not seen each other in person for many years, my heart is always with yours.

Chapter 2, in full, is a reprint of the material as it appears in the journal Nucleic Acids Research. Mo, Y., Keller, N., delToro, D., Ananthaswamy, N., Harvey, S. C., Rao, V. B., & Smith, D. E. (2020). Function of a viral genome packaging motor from bacteriophage T4 is insensitive to DNA sequence. *Nucleic Acids Research*, 48(20), 11602-11614. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in the journal Frontiers in Molecular Biosciences. Mo, Y., Fizari, M., Koharchik, K., & Smith, D. E. (2021). Determining Trap Compliances, Microsphere Size Variations, and Response Linearities in Single DNA Molecule Elasticity Measurements with Optical Tweezers. *Frontiers in Molecular Biosciences*, 8, 93. The dissertation author was the primary investigator of this paper.

Chapter 5, in part is currently being prepared for submission for publication of the material by Youbin Mo and Douglas E. Smith. We thank Damian del Toro for his collaboration in conducting measurements for this project. The dissertation author was the primary investigator and author of this material.

VITA

2013	Bachelor of Science, Xiamen University
2016–2017	Teaching Assistant, Department of Physics University of California San Diego
2017–2021	Research Assistant, University of California San Diego
2018	Candidate for Doctor of Philosophy, University of California San Diego
2021	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Mo, Y., Keller, N. A., & Smith, D. E. (2020). High GC Content DNA does not Affect Phage T4 DNA Packaging--Test of a Scrunchworm Model for Packaging Motor Function. *Biophysical Journal*, *118*(3), 188a (Conference Abstract).

Mo, Y., Keller, N., delToro, D., Ananthaswamy, N., Harvey, S. C., Rao, V. B., & Smith, D. E. (2020). Function of a viral genome packaging motor from bacteriophage T4 is insensitive to DNA sequence. *Nucleic Acids Research*, *48*(20), 11602-11614.

Smith, D. E., Mo, Y., Keller, N., Ananthaswamy, N., Harvey, S. C., & Rao, V. B. (2021). Function of a Viral Genome Packaging Motor from Bacteriophage T4 is Insensitive to DNA Sequence. *Biophysical Journal*, *120*(3), 36a (Conference Abstract).

Mo, Y., Fizari, M., Koharchik, K., & Smith, D. E. (2021). Determining Trap Compliances, Microsphere Size Variations, and Response Linearities in Single DNA Molecule Elasticity Measurements with Optical Tweezers. *Frontiers in Molecular Biosciences*, *8*, 93.

FIELDS OF STUDY

Major Field: Physics (Biophysics)
Studies in Biological Physics and Applied Physics
Professor Douglas E. Smith

ABSTRACT OF THE DISSERTATION

Studies of Motor-Driven Viral DNA Packaging by Single DNA Manipulation with Optical Tweezers

by

Youbin Mo

Doctor of Philosophy in Physics

University of California San Diego, 2021

Professor Douglas E. Smith, Chair

DNA packaging is a major step in the process of assembly of many viruses, including many dsDNA bacteriophages (bacterial viruses). An ATP-powered molecular motor translocates a dsDNA molecule into the viral prohead shell where it is confined to near-crystalline density. Our studies concern efforts to understand how this multi-subunit nanoscale motor functions to transfer ATP chemical energy to mechanical work as needed to translocate

DNA against resisting load forces. In this dissertation, optical tweezers, a method for real-time, high force, and small displacement measurements on single DNA molecules, was employed to conduct a series of experiments on bacteriophage T4, φ29 and Lambda DNA packaging mechanisms. Improved instrument calibration methods and novel data analysis methods were also introduced. We used measurements of DNA elasticity in the linear, high-force regime to determine trap compliance more-reliably, simultaneously with other calibration parameters. New methods for pause and slip detection, data segmentation, and motor velocity determination were also explored. A major focus was investigation into whether substrate DNA sequence affects motor function. Evidence against a 'B-A Scrunchworm' model was obtained by showing T4 translocation dynamics are insensitive to A-philic DNA sequences. Preliminary data was also obtained supporting the same result for the φ29 motor. For T4, which exhibits large motor velocity fluctuations and pausing/slipping, more general evidence against sequence-dependent motor function was obtained from analyses looking for correlations across many events recorded with the same or different sequences. Phage lambda packaging was measured with improved resolution and with translocation interrupted by a putative ADP release inhibitor (Na_3VO_4) previously found with φ29 to cause isolated bursts of translocation steps ~10 bp in size. For lambda, preliminary evidence was obtained suggesting a different quantized translocation size of ~5 bp and a distinct packaging-slip-pause behavior. Lastly, preliminary studies of the T4 motor found no significant slowing with a high concentration of added phosphate, suggesting that phosphate release following ATP hydrolysis

is essentially irreversible, consistent with prior reported findings on φ29 packaging. Preliminary T4 measurements with Na_3VO_4 revealed significantly slowed packaging but quantized translocation steps could not be detected because all translocation events were interrupted by variable size slips.

Chapter 1
Introduction of Bacteriophages and Dual-Trap Optical Tweezers

1.1 Double-stranded DNA Bacteriophages

Double-stranded DNA (dsDNA) Bacteriophages are viruses that have double stranded genomic DNA and reproduce themselves by infecting the host bacteria cells.[1] They have a general structure where the genomic DNA is contained inside a protein shell called a ‘head’ or ‘capsid’, often icosahedral in shape. (Fig. 1.1A). The diameter of the phage capsid is in order of ~40-100 nm. In some cases, the capsid has a slightly different and smaller shape prior to DNA packaging and this structure is called a ‘prohead’ or ‘procapsid’. Viral dsDNA genomes are typically ~20-200 kilo base-pairs, or ~7 to 70 μm, meaning that the genome length is typically hundreds times the diameter of the virus body. A phage’s lifecycle starts from when it attaches itself to the membrane of the host cell and injects its dsDNA into the cell cytoplasm.[2, 3] The injected viral DNA undergoes transcription, translation and replication processes for preparation of the assembly of new viruses. An alternate pathway occurring with some phages in some conditions is the lysogenic cycle in which the phage DNA integrates into the host cell genome and is replicated, but viruses are not produced; as such this pathway is not directly relevant to our studies (Fig. 1.1B) The procapsid shell assembles first and then fully packages one copy of viral DNA, and after the motor dissociates it usually then assembles a ‘tail’ structure, to become a mature bacteriophage. After packaging, host cell membrane lysis is triggered such that the fully

packaged viruses release to infect other host cells, which starts a new lifecycle of phage replication.

Fully packaging the entire viral genomic DNA is the vital step in determining whether a phage can start a new lifecycle to replication. During the packaging stage, a single DNA molecule is packed into a prohead to a very high density of ~0.5 mg/ml, (approaching crystalline density levels, with the remainder of the volume being water and salt/buffer ions). This process is driven by a packaging motor which uses ATP chemical energy to do mechanical work to translocates DNA against forces resisting DNA packaging arising from entropy loss, DNA bending rigidity, and electrostatic self-repulsion forces. The power density of the package motor has been estimated to be as high as ~5000 kW/m^3 which is as twice as high a typical car engine. A topic of major interest is to understand the detailed mechanism by which this motor interacts with and translocates DNA.[4-6]

Bacteriophages T4, Lambda and φ29 are the model virus systems we use in our lab because they are relatively easy to manipulate and share many similar features (such as a high degree of similarity in the structures or predicted structures of the ATPase motor proteins), but also because there are notable interesting differences to explore (such as differences in genome lengths and capsid sizes and shapes, and since T4 and Lambda lack an unusual RNA element that the φ29 motor has).

1.1.1 Background on Phage T4

Bacteriophage T4 was selected as one of the model systems to conduct DNA packaging measurements primarily because it has fairly reliable activity for successful recording of

packaging events in the optical tweezers assay. Also, T4 serves as an alternative system to the much-studied φ29, in order to potentially discern universal features about motor function from system-specific ones. The T4 system notably differs from φ29 in that the motor complex does not have an RNA subunit, translocation occurs with only on one type of protein subunit (the 'large terminase subunit' (gp17), and has an average translocation rate four times higher than φ29.

The T4 icosahedral head is composed of gene products gp23 and gp24 that encapsidates a 172 kbp concatemeric dsDNA genome.[7-9] In our measurements we usually only package much shorter DNA substrates of only ~10-25 kilobases. When studying basic motor function issues this is an advantage because the large DNA capacity of the T4 head results in a situation where the density of the packaged DNA remains sufficiently low that there are no significant 'internal' forces resisting DNA packaging, which can otherwise be difficult to estimate and can otherwise cause the motor velocity to decrease with increasing amount of DNA packaged. Hence, in our studies the total load force on the motor is only the 'external' force we apply by stretching the DNA with optical tweezers. Gene product gp20 forms a dodecameric portal protein ring which forms the channel into the prohead to which the motor attaches and through which to the DNA is translocated (and later ejected). The T4 motor 'terminase' complex, comprised in the full in vivo system of a ring of multiple gp16 (18kDa) and gp17 (70kDa) proteins is directly connected with gp20. The 'large subunit' gp17 is alone sufficient for responsible for ATP hydrolysis and DNA translocation and forms a pentameric ring of subunits. In practice, the in vitro T4 optical tweezers packaging assay does not include the gp16 protein because it is unnecessary and, in fact, interferes with efficient initiation of packaging in the in vitro assay.[10]

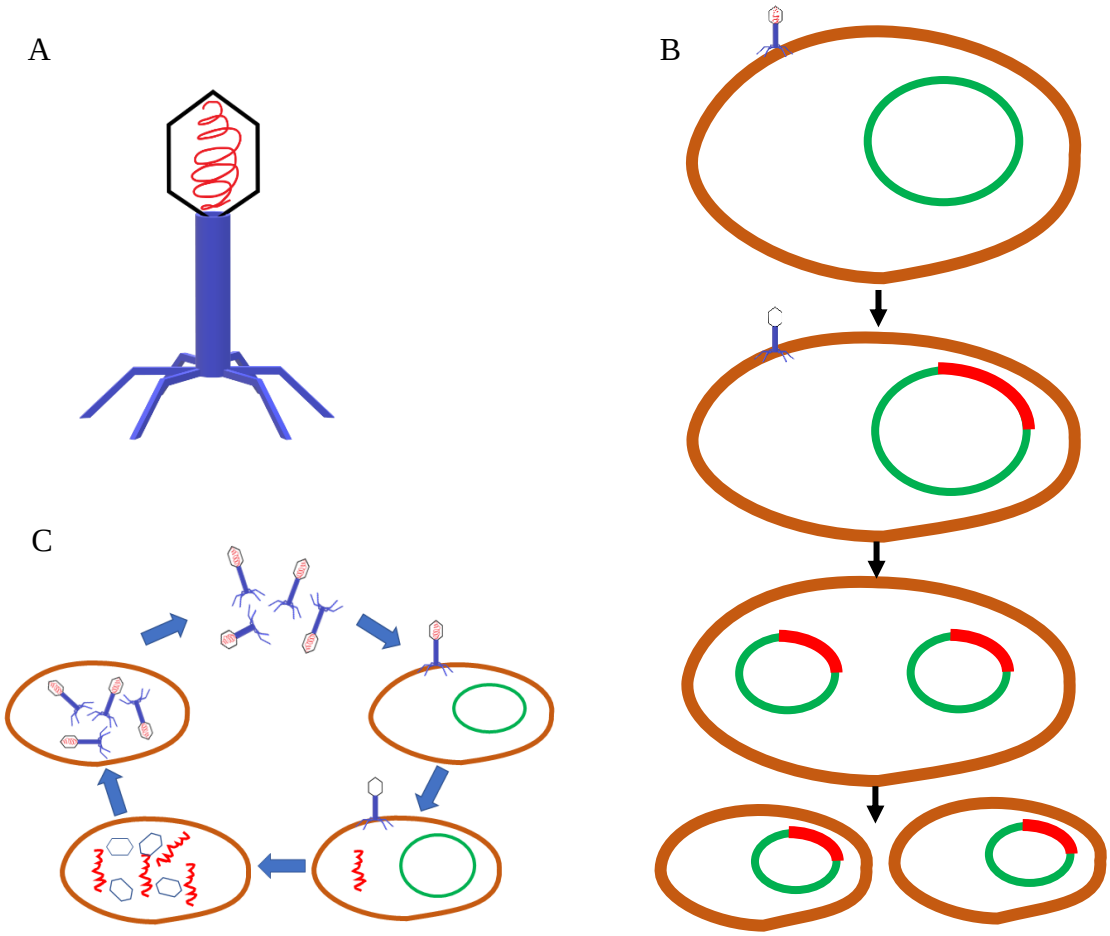

Figure 1.1 Schematic illustrations of (A) Bacteriophage and (B) Lysogenic and (C) Lytic lifecycles.

Another feature of the T4 packaging terminase/motor complex is the 'endonuclease' function. In vivo replication of the viral DNA results in a concatemer string of multiple genomes. The T4 motor introduces an endonucleolytic cut to this concatemer of the genomes to initiate packaging and another cut after approximately a genome-length of DNA has been packaged. This process is not relevant to our studies because, in our optical tweezers measurements, a pre-formed prohead-motor complex is observed to initiate packaging near the end of the DNA

construct and the whole genome length is never packaged. Another feature is that T4 only exhibits the lytic lifecycle pathway. [11, 12]

1.1.2 Background on Phage φ29

The length of phage φ29 genome is 19.3 kbp, which is only ~1/9 of T4, and the prohead is correspondingly smaller in internal volume. It has been observed that the motor velocity starts significantly decreasing with increasing length of DNA packaging when > 20% of genome is packaged into prohead, due in part to 'internal forces' resisting the DNA confinement. φ29, hence, is a good model for studying the dynamics of packaging in the limit when the procapsid shell is approaching the nearly full state, when the motor is observed to slow significantly and pauses in translocation and backwards slipping of the DNA out of prohead are also observed to happen more frequently.[8, 13-16] Only limited preliminary studies of φ29 packaging are reported in this thesis since most of the work focused on the T4 system.

1.1.3 Background on Phage Lambda

The lambda phage is one of the bacteriophages that can replicate by either the lytic or lysogenic lifecycle, and its packaging mechanism is more complicated than T4 and φ29 because, like T4 it has terminase 'endonuclease' (DNA cutting) function, but there are specific DNA sequences in the viral genome needed for initiation and termination of packaging. In addition, the lambda procapsid undergoes significant expansion in size during packaging. The diameter of its icosahedral procapsid is originally ~50 nm but increases to ~64 nm after expansion. The length

of lambda genome is 48,502 bp and it has complementary 12-nt single-stranded DNA (ssDNA) tails on both ends. After injection, the lambda DNA circularizes itself by connection of the complementary tails and then the nick is sealed with host bacteria cell's ligase proteins.[8, 15, 16] Lambda DNA replication is completed by utilizing the host cell's machinery thus yields a concatemer lambda genomes. There is a special 200-bp sequence (cos site) at the junction between each genome, which is composed of *cosQ*, *cosN* and *cosB*.[8, 15-17]

cosN is a nick site introduced by terminase at the time of packaging initiation or termination. The lambda motor complex produces this 12-bp staggered nicks symmetrically between genomes. *cosB* site, positioned between *cosN* and the beginning of the genome, is a site containing 3 DNA recognition sites R1, R2, and R3 where the packaging motor binds at the packaging initiation. *cosN* site, at the end of genome, is a site corresponding for the packaging termination. The motor terminates the packaging and cleaves DNA when it recognizes the *cosN* site.[17-20] In our studies we use a DNA template that has the packaging initiation site. As with the T4 measurements we only package a ~10 kbp DNA template that is much shorter than the genome length. Thus, we can study DNA translocation without complications of any significant slowing of the motor with increasing prohead filling, any effects due to capsid expansion, and termination of packaging does not occur.

1.2 Optical Tweezers Methods

1.2.1 Optical Tweezers Theory

Optical tweezers has been applied in many types of studies in biology, medicine, nanoengineering and nanochemistry, as well as in basic physics studies involving quantum optics and quantum optomechanics, etc.. The single-beam trap method we use was first proposed and demonstrated by Ashkin et al. in 1986 [19, 20] and Arthur Ashkin was awarded with the 2018 Nobel Prize in Physics for the development of the technique and early applications in biology studies including manipulation of bacteria, viruses, and cellular organelles. Fig. 1.2 illustrates the mechanism of optical tweezers. An optically transparent particle, such as a plastic particle that has a higher index of refraction than the surrounding water, is trapped by a highly focused laser beam because due to refraction the particle deflects the laser beam rays, which corresponds to changes in photon momenta. According to Newton's laws (or conservation of momentum for the system of photons and particle), the photons transfer momentum to the trapped particle and thus yield a small but significant force (called the gradient force) acting on the particle.[19, 21-24, 28] Alternatively, in classical electricity and magnetism physics picture, strictly only valid when the trapped particle is much smaller than the focused region of the laser beam, the particle acts as an electric dipole (induced by the field of the laser). The force exerted by the field on the particle can be calculated using the Lorenz force formula:

$$\vec{F} = (\vec{p} \cdot \vec{\nabla})\vec{E} + \frac{1}{c}\frac{d\vec{p}}{dt} \times \vec{B}$$

where $\vec{F}$ is applied force on particle, $\vec{p}$ is the induced dipole moment of the particle, $\vec{E}$ and $\vec{B}$ are the external electric and magnetic fields and *c* is the speed of light. The laser beam in this model has a focal spot where the electric field is very intense and induces dipole moment of a dielectric particle given by $\vec{p} = \alpha\vec{E}$, where α is the particle electric polarizability. The electric field is rapidly fluctuating, and the direction of polarization follows. The continuous wave laser leads the partial derivatives about time to be 0, i.e., $\frac{d}{dt}(\vec{E} \times \vec{B}) = 0$. The force expression becomes $\vec{F} = \alpha\vec{\nabla}E^2$ which implies the bead is pulled towards the region of highest $|\vec{E}|$ field intensity position, which is the region near the center of the laser focus. It can be understood as if there is a 3D spring (Hooke's law type force) holding the particle at the equilibrium position which is the focus of laser beam.

One of advantage of optical tweezers is it does not touch or significantly damage biological samples. To achieve this, we use near-infrared wavelength laser beam (740 – 1400 nm). The trapped beads are polystyrene particles with diameter of 2.0 – 2.5 μm, coated with biotin or protein G. The optical tweezers instrument in our lab was customized by Damian del Toro and Nicholas Keller. The 1064 nm wavelength laser beam, generated by yttrium fiber laser generator, is split into to two orthogonal linear polarizations such that we implement the dual-traps functionality with dual beam detection based on polarization. The beam separation does not affect the trapping but brings convenience and simplification to the control system. One of beams is reflected by a rotatable mirror and then focused by objective. Another one is reflected by a fixed mirror, merged with the former beam and focused by the same objective. The outgoing laser beams pass the other objective then are captured by Position Sensing Detectors (PSD). The PSD detects the angular deflections of the beam, which are proportional to the force exerted by the trap on the beads, such that we record a signal proportional to the force. After

determining the trap compliance, as discussed in Chapter 2 and Chapter 4, we can also compute the displacement of the trapped particles. When a single dsDNA molecule is attached by its ends between two trapped microspheres, the trap forces exert force to stretch the DNA. A feedback control system is implemented to precisely adjust the separation between the two traps such the force is held constant if the length of the DNA changes. Alternatively, measurements can be done in a 'fixed position' mode where the traps are held fixed and the force changes if the DNA length changes. Calibration of the instrument to be able to measure forces and displacements in real physical units is done by using DNA stretching measurements, with DNA molecules having known lengths and elastic properties, as measurement standards.[25-27] Improvements in this method developed in this thesis work are described in Chapter 4.

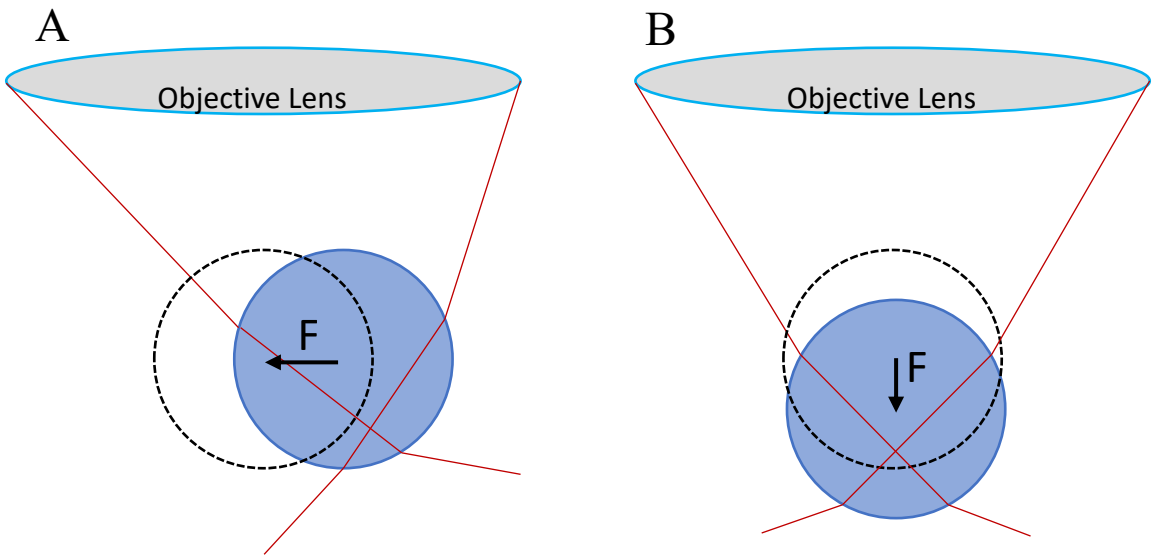

Figure 1.2 Diagram of optical tweezers mechanism. **(A)** A trapped bead (or 'microsphere' having a higher index of refraction than the surrounding water) displaced (by any external force of magnitude *F*) perpendicular to the optical axis results in a net change of the refracted photons' momentum to the right. At equilibrium the response force of magnitude *F* exerted by the trap on the bead, therefore, points to left perpendicularly to the optical axis. **(B)** A trapped bead displaced downwards parallel to the optical axis results in a net change of the refracted photons' momentum changed in the downwards direction. The force *F* exerted by the trap on bead, in this case, points upward parallel to the optical axis.

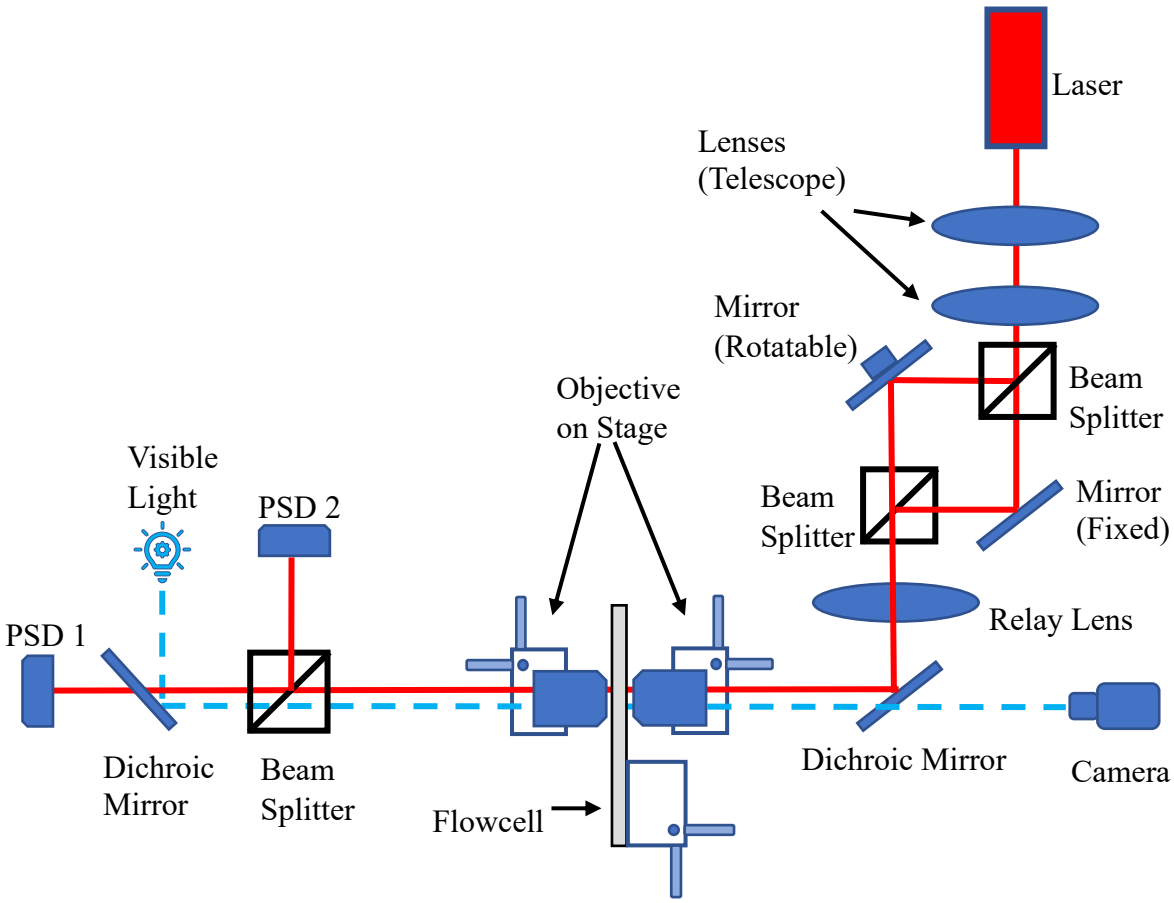

Figure 1.3 The customized optical tweezers system in our lab.

1.2.2 Experimental Setup

A fluidic micro chamber (flow cell) is placed on the focal plane, where the traps are formed, and the viral DNA packaging experiment is conducted. The flow cell has three different channels which are connected by small capillary tubes. (Fig. 1.4) There are two distinct protocols that can be used to initiate packaging: 'in situ initiation' in which single DNA packaging events are initiated in the flow cell; or use of 'pre-stalled complexes' in which packaging of many complexes is initiated in a bulk reaction and then stalled and single complexes are then manipulated and their packaging is restarted in the flow cell. The in-situ protocol is employed for phage T4 and φ29 packaging initiation in the flowcell. DNA and prohead samples are respectively incubated on streptavidin-coated and antibody-coated beads which are pumped into two side channels of the flow cell and flow into the main channel of the flow cell thru the capillary tubes. Two samples of different beads coated with DNA or phage proheads are trapped

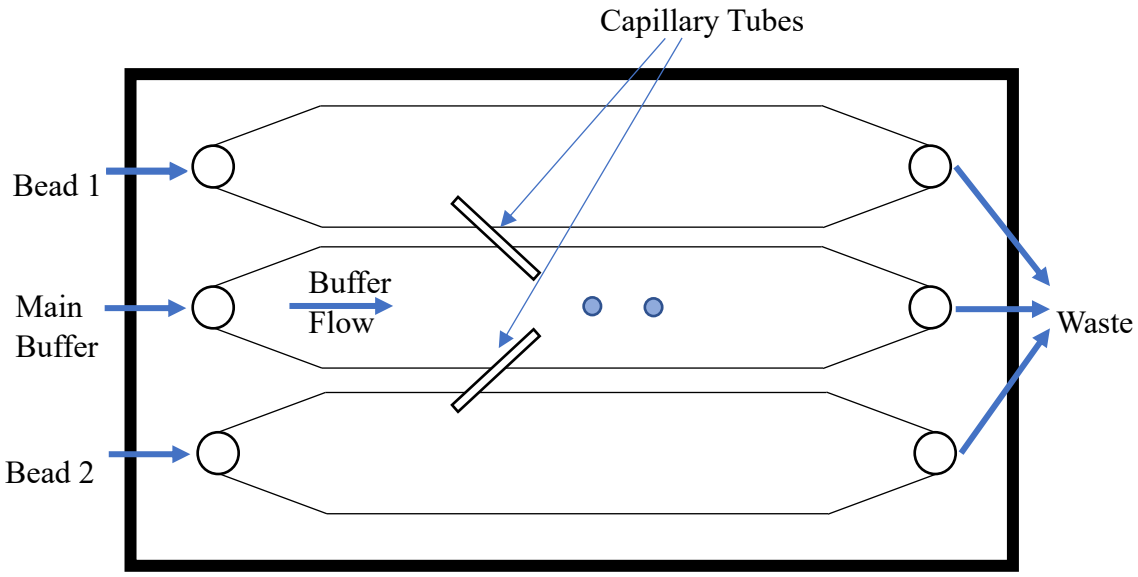

Figure 1.4 Diagram of the flowcell.

in the middle channel in the main packaging buffer which typically contains ATP or an ATP analog (depending on experiment design). We use pre-stalled complex protocol when we measure the packaging with Lambda phage of which the prohead stalled (by a non-hydrolyzable ATP analog) with partly prepackaged biotin labeled DNA is attached on antibody-coated beads and the streptavidin-coated is injected separately. These two types of beads flow into the middle channel and are then captured by optical tweezers, and then measurements are controlled automatically by computer programs which control the separation between the two traps and record the force signal. The flowing buffer in the middle, main channel of the flow cell is the same as in the bulk in vitro packaging reaction which contains ATP or an ATP analog (depending on experiment design).

The flow cell, on our optical tweezers instrument, is mounted on a 3-axis linear stage of which one of axis is parallel to the optical axis of laser beam and the other two axes positioned manually control the positions of the traps within the flow cell. While doing experiment, a researcher moves the flow cell to catch beads flowing thru top and bottom channels. Then a LabView instrument control program starts to automatically measure the viral packaging.

1.3 Brief Overview of This Dissertation

Chapter 2 describes work done in collaboration with Venigalla B. Rao Lab (who provided the T4 proheads and motor protein) and Stephen C. Harvey Lab (who designed a special DNA construct sequence for testing if packaging is sequence-dependent). In this chapter, we investigate the effect of DNA sequence on T4 packaging and found the T4 motor is insensitive to DNA sequence. The result has been published in the joural Nucleic Acids Research. Mo, Y., Keller, N.,

delToro, D., Ananthaswamy, N., Harvey, S. C., Rao, V. B., & Smith, D. E. (2020). Function of a viral genome packaging motor from bacteriophage T4 is insensitive to DNA sequence. *Nucleic Acids Research*, 48(20), 11602-11614, and the context of chapter 2 is a reprint of the text of this article.

Chapter 3 reports preliminary studies similar to those in Chapter 2 but we studied the functioning of the phage φ29 DNA packaging motor.

Chapter 4 describes improvements we developed in the methods used to calibrate the optical tweezers instrument based on measurements of single DNA molecule stretching (elasticity). We used a linear approximation of the worm-like chain model for DNA elasticity valid in a high-force measurement region to determine necessary measurement parameters for the optical tweezers system. We also used video imaging of trapped microspheres to determine displacement scale factors and trap compliances for individual trapped microspheres. My colleague graduate student Mounir Fizari assisted with these measurements and undergraduate student Kristina Koharchik assisted with video imaging of microspheres. This work was published in the journal Frontiers in Molecular Biosciences. Mo, Y., Fizari, M., Koharchik, K., & Smith, D. E. (2021). Determining Trap Compliances, Microsphere Size Variations, and Response Linearities in Single DNA Molecule Elasticity Measurements with Optical Tweezers. *Frontiers in Molecular Biosciences*, 8, 93, and the context of Chapter 4 is a reprint of the text of this article.

Chapter 5 describes preliminary studies of DNA translocation dynamics by the phage Lambda motor in conditions where the motor was slowed so that we could try to detect quantized discrete steps that the motor is likely to make in response to hydrolysis of individual ATP molecules. A novel step packaging-slip-pause dynamic behavior was also observed while Lambda packaging DNA in the packaging buffer with orthovanadate VO_4^{3-}, a putative ADP release inhibitor.

Chapter 6 presents studies I have done independently that explore novel proposed methods for signal processing of noisy signals, including detection of pauses (zero slope regions), slips (negative slope regions), and a novel method to calculate the slope of a monotonically changing section of time-series data having superposed, non-monotonic noise. Although these methods may have some advantages for the analysis of the DNA translocation data by the viral packaging motors, as discussed in Chapter 6, in the end these methods were not used for the final data analyses presented in the other chapters (in favor of using more traditional methods which proved sufficient).

Chapter 7 discusses studies of T4 packaging dynamics conducted in different buffers with phosphate (PO_4^{3-}) or its analogs to investigate whether phosphate ion rebinding slows packaging and to investigate if a putative ADP release inhibitor could slow packaging sufficiently to resolve quantized translocation steps.

References

1. Bertani, G. (1953, January). Lysogenic versus lytic cycle of phage multiplication. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 18, pp. 65-70). Cold Spring Harbor Laboratory Press.
2. Feiss, M., & Rao, V. B. (2012). The bacteriophage DNA packaging machine. *Viral molecular machines*, 489-509.
3. Rossmann, M. G., & Rao, V. B. (Eds.). (2011). *Viral molecular machines* (Vol. 726). Springer Science & Business Media.
4. Tzlil, S., Kindt, J. T., Gelbart, W. M., & Ben-Shaul, A. (2003). Forces and pressures in DNA packaging and release from viral capsids. *Biophysical Journal*, *84*(3), 1616-1627.
5. Riemer, S. C., & Bloomfield, V. A. (1978). Packaging of DNA in bacteriophage heads: some considerations on energetics. *Biopolymers: Original Research on Biomolecules*, *17*(3), 785-794.
6. Fuller, D. N., Raymer, D. M., Kottadiel, V. I., Rao, V. B., & Smith, D. E. (2007). Single phage T4 DNA packaging motors exhibit large force generation, high velocity, and dynamic variability. *Proceedings of the National Academy of Sciences*, *104*(43), 16868-16873.
7. Kanamaru, S., Kondabagil, K., Rossmann, M. G., & Rao, V. B. (2004). The functional domains of bacteriophage t4 terminase. *Journal of Biological Chemistry*, *279*(39), 40795-40801.
8. Rao, V. B., & Feiss, M. (2008). The bacteriophage DNA packaging motor. *Annual review of genetics*, *42*, 647-681.
9. Sun, S., Kondabagil, K., Gentz, P. M., Rossmann, M. G., & Rao, V. B. (2007). The structure of the ATPase that powers DNA packaging into bacteriophage T4 procapsids. *Molecular cell*, *25*(6), 943-949.
10. Sun, S., Rao, V. B., & Rossmann, M. G. (2010). Genome packaging in viruses. *Current opinion in structural biology*, *20*(1), 114-120.
11. Black, L. W., & Rao, V. B. (2012). Structure, assembly, and DNA packaging of the bacteriophage T4 head. *Advances in virus research*, *82*, 119-153.
12. Mitchell, M. S., Matsuzaki, S., Imai, S., & Rao, V. B. (2002). Sequence analysis of bacteriophage T4 DNA packaging/terminase genes 16 and 17 reveals a common ATPase center in the large subunit of viral terminases. *Nucleic acids research*, *30*(18), 4009-4021.
13. Grimes, S., & Anderson, D. (1997). The bacteriophage φ29 packaging proteins supercoil the DNA ends. Journal of molecular biology, 266(5), 901-914.
14. Zhao, W., Morais, M. C., Anderson, D. L., Jardine, P. J., & Grimes, S. (2008). Role of the CCA bulge of prohead RNA of bacteriophage ø29 in DNA packaging. *Journal of molecular biology*, *383*(3), 520-528.

15. Catalano, C. E. (2005). Viral genome packaging machines. *Viral Genome Packaging Machines: Genetics, Structure, and Mechanism*, 1-4.8.

16. Smith, D. E. (2011). Single-molecule studies of viral DNA packaging. *Current opinion in virology, 1*(2), 134-141.

17. Chemla, Y. R., & Smith, D. E. (2012). Single-molecule studies of viral DNA packaging. *Viral molecular machines*, 549-584.

18. Cue, D., & Feiss, M. (1998). Termination of packaging of the bacteriophage λ chromosome: cosQ is required for nicking the bottom strand of cosN. *Journal of molecular biology, 280*(1), 11-29.

19. Ashkin, A., Dziedzic, J. M., Bjorkholm, J. E., & Chu, S. (1986). Observation of a single-beam gradient force optical trap for dielectric particles. *Optics letters, 11*(5), 288-290.

20. Moffitt, J. R., Chemla, Y. R., Smith, S. B., & Bustamante, C. (2008). Recent advances in optical tweezers. *Annu. Rev. Biochem., 77*, 205-228.

21. Stevenson, D. J., Gunn-Moore, F. J., & Dholakia, K. (2010). Light forces the pace: optical manipulation for biophotonics. *Journal of biomedical optics, 15*(4), 041503.

22. Jonáš, A., & Zemanek, P. (2008). Light at work: The use of optical forces for particle manipulation, sorting, and analysis. *Electrophoresis, 29*(24), 4813-4851.

23. Chaumet, P. C., & Nieto-Vesperinas, M. (2000). Time-averaged total force on a dipolar sphere in an electromagnetic field. *Optics letters, 25*(15), 1065-1067.

24. Shaevitz, J. W. (2006). A practical guide to optical trapping. *University of Washington*, 138.

25. Shaevitz, J. W., Abbondanzieri, E. A., Landick, R., & Block, S. M. (2003). Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. *Nature, 426*(6967), 684-687.

26. Rickgauer, J. P., Fuller, D. N., & Smith, D. E. (2006). DNA as a metrology standard for length and force measurements with optical tweezers. *Biophysical journal, 91*(11), 4253-4257.

27. delToro, D., & Smith, D. E. (2014). Accurate measurement of force and displacement with optical tweezers using DNA molecules as metrology standards. *Applied physics letters, 104*(14), 143701.

28. Bustamante, C. J., Chemla, Y. R., Liu, S., & Wang, M. D. (2021). Optical tweezers in single-molecule biophysics. *Nature Reviews Methods Primers, 1*(1), 1-29.

Chapter 2
Function of a Viral Genome Packaging Motor from Bacteriophage T4 is Insensitive to DNA Sequence

Youbin Mo[1], Nicholas Keller[1], Damian delToro1, Neeti Ananthaswamy[2], Stephen C. Harvey[3,*], Venigalla B. Rao[2,*] and Douglas E. Smith[1,*]

[1] Department of Physics, University of California, San Diego, La Jolla, CA, 92093, USA

[2] Department of Biology, The Catholic University of America, District of Columbia, USA

[3] Department of Biochemistry and Biophysics, Univ. of Pennsylvania, Philadelphia, PA 19104, USA

[*] To whom correspondence should be addressed: Email: des@ucsd.edu

Correspondence may also be addressed to: Email: rao@cua.edu, sharvey7@icloud.com

Abstract

Many viruses employ ATP-powered motors during assembly to translocate DNA into procapsid shells. Previous reports raise the question if motor function is modulated by substrate DNA sequence: i) The phage T4 motor exhibits large translocation rate fluctuations and pauses and slips, ii) Evidence suggests that the phage phi29 motor contacts DNA bases during translocation, and iii) one theoretical model, the 'B-A scrunchworm', predicts that 'A-philic' sequences that transition more easily to A-form would alter motor function. Here, we use single-molecule optical tweezers measurements to compare translocation of phage, plasmid, and synthetic A-philic, GC rich sequences by the T4 motor. We observed no significant differences

in motor velocities, even with A-philic sequences predicted to show higher translocation rate at high applied force. We also observed no significant changes in motor pausing and only modest changes in slipping. To more generally test for sequence dependence, we conducted correlation analyses across pairs of packaging events. No significant correlations in packaging rate, pausing, or slipping versus sequence position were detected across repeated measurements with several different DNA sequences. These studies suggest that viral genome packaging is insensitive to DNA sequence and fluctuations in packaging motor velocity, pausing, and slipping are primarily stochastic temporal events.

2.1 Introduction

Most double-stranded DNA viruses, including many bacterial viruses (phages) and human/animal viruses including herpesviruses, poxviruses, and adenoviruses, follow an assembly pathway in which the viral procapsid (prohead) shell assembles first and it is subsequently filled with DNA (1,2). An ATP-powered molecular motor translocates replicated viral DNAs into the procapsids via a portal nanochannel (2,3). DNA translocation by the phage phi29, lambda, and T4 motors have been directly measured via single DNA molecule manipulation with optical tweezers (4-9). These studies revealed that viral motors are among the strongest known molecular motors, able to exert forces of at least ~60 pN (4-6). In comparison, a single skeletal muscle myosin II motor protein, also powered by ATP, exerts a maximum force of only ~2-3 pN. More broadly, viral motors have been identified as members of an ASCE (additional strand, conserved E) superfamily of ATPases that play many critical roles in cell

biology including nucleic acid unwinding, chromosome transport, and protein translocation and unfolding (10).

High force generation is a necessary property of the phage viral motors since high forces resist the tight confinement of the DNA genome, which is ultimately packed to near-crystalline density (2,11,12). Many theoretical and single molecule studies have investigated the physical factors that govern this process, including DNA bending and electrostatic self-repulsion, ionic-screening effects, entropy changes, and non-equilibrium dynamics (13-21). The viral motors are sufficiently powerful that they can nevertheless translocate DNA rapidly against the opposing forces and package the viral genome in just a few minutes (22,23). A striking example is the phage T4 motor which was measured to translocate DNA at rates as high as ~2000 bp/s and generate a power density as high as ~5000 kW/m^3 at room temperature with saturating ATP, which is twice that of a typical automobile engine (6).

The exact biochemical-structural mechanism by which viral motors function, and whether there is a universal packaging mechanism, is not fully understood although significant functional and structural information have been obtained on several model systems. Examples of recent progress, not intended to be a comprehensive review, include the following. Atomic-resolution structures of motor proteins or their ATPase domains have been obtained for bacteriophages T4, Sf6, P74-26, D6E, φ29, and herpes simplex virus 1, as well as evidence that the motor ATPases form a multi-subunit ring surrounding the DNA, and have led to the development of several models for various aspects of motor function (24-29). High-resolution optical tweezers measurements of the phage φ29 motor revealed that DNA is translocated in bursts of four 2.5 bp steps and general features of the motor's chemo-mechanical kinetic cycle have been described (7,30,31). In several systems, coordination of the multiple motor subunits

has been shown to be regulated by trans-acting ‘arginine finger’ residues (26-28). Studies on the effects of site-directed residue changes in the phage T4 and lambda motors have illustrated the roles of specific amino acids in ATP binding, catalysis of hydrolysis, motor force generation, motor pausing and slipping, and chemo-mechanical coupling (32-37).

However, fundamental questions about motor-DNA substrate interactions and how they influence various aspects of the packaging mechanism remained poorly understood. First, is the function of packaging motor influenced by DNA sequence in terms of translocation velocity and motor dynamics such as pausing and slipping? Second, do specific DNA properties such as propensity for transition from B- to A-form or local curvature affect the function? In the present work, we use the T4 motor to investigate these questions. Additional motivation for addressing these questions came from several previous experimental findings and one theoretical model that implicate potential sequence effects in motor-DNA interactions during DNA translocation. First, a striking feature of the T4 motor is that the rate of DNA translocation, even at low capsid filling where forces resisting DNA confinement are negligible, was found to be highly variable (5). The speed of individual motors was observed to vary in time, often by at least several hundreds of bp/s. The cause of these fluctuations is unknown, but since segments with varying DNA sequences were being packaged one possible hypothesis is that the local sequence of the substrate DNA affects translocation speed. DNA sequence is known to affect local structure and physical properties of DNA including intrinsic curvature, bendability, duplex stability, etc (38-40). An example relevant to our present studies is that DNA sequence has been shown to affect propensity of DNA to transition between B- and A-forms (41,42) and this propensity is an important factor in some protein-DNA interactions (43-45). Additionally, some aspects of motor function have also been shown to depend on the substrate DNA sequence. For instance, the

phage lambda motor is able to recognize a termination sequence in the DNA, which causes the motor to stop translocating and switch to an endonuclease mode (46). Other examples include the bacterial dsDNA translocating chromosome segregation motors FtsK and SpoIIIe, which can change translocation direction in response to certain sequences (47,48).

A second motivation for investigating substrate sequence dependence is the unexplained observation that all three viral motors for which translocation dynamics have been measured exhibit abrupt pauses and slips, even at low capsid filling (4-6). Since these studies measured packaging of heterogeneous DNA sequences, one possibility is that pauses or slips occur, or occur with increased probability, when the motor interacts with certain DNA sequences. An alternate possibility is that these events are independent of DNA sequence and just caused by fluctuations in the conformation of the motor protein and/or its alignment with the DNA that are stochastic in time.

A third motivation for these studies is evidence from single-molecule studies that the phage phi29 motor makes contacts with the DNA bases or sugars during translocation (49). Responses of motor translocation to a range of modified DNA substrates, including ones that contained uncharged methylphosphonate DNA strands, segments with sugars and bases removed, single-stranded gaps, or unpaired bulges, were studied as a function of applied load forces. During the dwell phase when ATP binds to multiple motor subunits, important gripping contacts are made with the phosphate backbone, but during the subsequent burst of translocation steps, nonspecific contacts are made with many parts of the DNA including the bases or sugars (49). It is therefore plausible that different DNA base sequences could cause differences in the translocation dynamics.

A fourth motivation is that one proposed model for motor function, the 'B-A scrunchworm' model hypothesized that the motor induces structural changes in the DNA substrate during translocation which would be sensitive to DNA sequence (50). This model hypothesized that the motor generates force that drives translocation by inducing repeated transitions between the B-form and A-form of DNA segments threaded into the motor channel, coordinated with motor gripping and releasing actions (Fig. 2.1). Transition of the threaded segment from the longer B-form to a shorter A-form was hypothesized to pull ~2.5 bp of DNA outside the capsid into the motor channel and transition of the segment back to B-form is then proposed to push DNA into the capsid. This model is consistent with experiments that find evidence for DNA length changes within the motor (51,52). It is also in agreement with the experimentally measured phi29 motor step size (7,30,31). This model further argues that the force generated to pull the DNA into the motor channel should depend on the ease with which particular segments of the DNA can convert between the B and A forms, characterized by the free energy difference per basepair (ΔG_{BA}) between B and A forms, which depends on DNA sequence (41). A prediction of this model is that 'A-philic' sequences that convert more easily to A-form (having lower ΔG_{BA} values) would be easier for the motor to pull in and less affected by external forces opposing translocation. Because motor velocity depends on load force, changes in ΔG_{BA} would be expected to alter DNA translocation rate if the B-A scrunchworm model is correct (6,50).

Features of the B-A transition have been reviewed previously (53,54) and we mention only a few. In solution, high concentrations of ethanol can induce B-A transitions (55) and dynamic measurements suggest that the transition timescale is ~2 μs, although evidence was found for minor relaxation components as long as 50-100 μs. This timescale is compatible with

the B-A scrunchworm model, which predicts a 2.5 bp step, because it would allow for translocation speeds up to ~2.5 bp/100 μs = 25,000 bp/s, which is higher than the measured rate. Some studies have also found evidence for intermediates between the A- and B- DNA forms that more detailed scrunchworm models may need to consider (56). One puzzling finding is that a study of single DNA stretching in ethanol unexpectedly found no detectible transitions between the ~0.34 nm/bp length expected for normal B-form DNA and the shorter length (~0.26 nm/bp) conventionally expected for A-form DNA (57). This study raises questions about the orthodox view that A-form DNA in solution has an inherently shorter length per bp than B-form. However, the B-A scrunchworm model proposes that it is interaction with the portal and/or motor protein rings that induces the threaded DNA segment to transition to a shorter A-form conformation. The plausibility of such a mechanism is supported by published structures of protein-DNA complexes which show that protein-DNA interactions can induce a shortened A-form DNA conformation (43,58).

As mentioned above, 'internal forces' resisting DNA confinement also resist translocation of DNA segments into the viral prohead and exert load force on the motor. The B-A scrunchworm model predicts that it would be more difficult for the motor to translocate A-philic segments into the capsid against internal forces because they make transition of the threaded segments from A-form to B-form more difficult. However, in the early stages of packaging that we study here, where capsid filling is low, internal forces resisting translocation are very small and are expected to have a negligible effect on translocation dynamics (6,14,15,30). Our present studies focus on predicted effects of an externally applied load force (Fig. 2.1), where A-philic segments are predicted to be translocated at a higher rate. Motor velocity slows with increasing applied force (6) and this implies that the rate limiting step in the

motor's mechanochemical cycle is the translocation step (59-61). Translocation rate is predicted to follow Arrhenius-like kinetics and depend exponentially on the height of the energy barrier between the two states (59-61). When the motor generates force to translocate against a load force, this is predicted to raise the energy barrier by an amount $F\Delta x$ and slow the translocation rate by a factor $\exp(-F\Delta x/kT)$. Since the B-A scrunchworm model proposes that motor force is proportional to ΔG_{BA}, a reduction in ΔG_{BA} is predicted to increase motor velocity.

Here we ask, by placing an A-philic, GC-rich, DNA sequence in the middle of a non-A-philic linear plasmid DNA sequence, whether viral packaging motor velocity and translocation dynamics such as pausing and slipping are altered due to structural differences caused by these sequences. In addition, we ask more generally whether any measured aspects of motor translocation dynamics are affected by DNA sequence. We present measurements of the packaging of different phage and plasmid DNA sequences as well as correlation analyses to look for sequence dependences of packaging dynamics when repeatedly packaging the same sequence versus when packaging different sequences. Our studies show no evidence, that either A-philic or other variable sequences influence motor function, aside from small differences in back slipping. Our findings therefore suggest that viral genome packaging is insensitive to DNA sequence, impose constraints on the packaging models for motor function, and provide evidence that motor velocity fluctuations, pausing, and slipping are primarily stochastic in time.

2.2 Materials and Methods

2.2.1 DNA Constructs

2.2.1.1 A-philic DNA and Control DNA Template

A 2,014 bp ‘A-philic’ dsDNA sequence, derived from a 40 bp ‘LilF’ sequence described previously (62), was synthesized and cloned into a 9,276 bp plasmid vector pPIC9K using NotI and EcoRI cloning sites (ThermoFisher Scientific, Inc., Project ID: 15ADFYGC, Construct Name: t1_lowF-seq). The construct was verified by sequencing and the sequence is provided in the following:

5′-TATCTCTCGAGAAAAGAGAGGCTGAAGCTTACGTAGAATTCGGCACCTGCGGGCTACCGCGG
TACTACCTACCTACACCACCTACACCTATACCCCCCACTACCTACCCCGTACCCCCCCCCTACCG
GGGGCCCTACGGCTCCCACACTAGGTGCTACTACACCGTACCACTACCCACCCCCCTACCACGCC
CACTACACTCCCTATACCTACCTACTACCCCCACGGTACATCGGTACGGGGGCACCCCCCTATCT
GCTACCCCCGCCCCACCTACGCGGGCGGGGGGTACCACCCACGGGGTACCCAGCCACACCCCGCTA
GCTACCCCTAGCAGGCCACCCCTACCGTAGGTACATACGTACCGCCGCCTACCTACGGCATACCT
ATACACCCCTACCCCTATGCGCTCCCCCCCCCACACTATACCCCACACCCCCACGATACCGGCCC
GCCACTGGGGTACCCCTAGCCACTATAGTACTGGGTACTACCCGCCCGCACCCGGTACCACCCAC
CCCTATAGTGGAGGTACACTACACACGTCGGCTACGGTACCCACCACGTACTCCCACGGCATACC
CACCTACCTATACCGGCCACACACCCTCGGTACCTACCGTACCCCCTATATACCACGGGCATACC
TCCCCACCATCCTACCGCACTATACGGCCCCCCCATACATACCTACTACAGTACCCATACCCGGG
ACCACCACCCCATACACCTGACTCCACCACCCCCGCTACCTATACTACCCGTATACCCTATAGCC
CACTACTACTGATATGTACCTACCACCAGCCTACGCCAGGGGGCCCCTGACGTACGCTACCACCC
CTACTACCCCCCGGTACGCCCACCCCCCGCCCCGGTATAGGGGGCTATACACCTACTAGTGCCCT
ACCTACACTACCCACCGGTGGGTACTACCGTATACACCTACTGCCACCTCGGGTATACCACTACC
ACACCCCCCGTACCTACCCAGACCCCCGGTGTACCCACACCCCTACATACTACTGGCTACGTACC
CCTGCTAGTACCCACTACACTACACCCCCACACCGGGTCGCCCTACCGGTGGCCCCCCCACCTGC
ACCTACCCCACGCCTACTACCCCTAGTAGGCCCTACCCCCTACGTGCATAGCACCCCCCCCCTAC
CCTACACCCGGGGCCCACCGTGTACCACCCACTAGGGTACCCCCTACGCTACTCCATACTGGCCC
CGTATACCCACACCACACACCATACACACCTACTACCGGGTATACGTGGACCACTACTACCACCC
CTGCGGTACTAGGCCTCCCCTACCGCCCGCCGCCGGGGGCTACTACCACTACCACCCATACTACC
CCACCCTAGGCCTAGGCTACACGTGGACACACCCCCCATCGCCCATACCACGCATGGGCTAGGGC
ACACCCCCACCCTACCTACTACCCCACTACCCCGGGTACCCCTACTACTACGCACACTAGGCCCC
ACTACCTGCACACCACACTACCTATACCATAGTAGCGGCCCCACTACGTACTACCACACCCCTAC
TACCAGCTAGGGCCACCCCTATACCCGGGGGGACATACCGGCCCCCCGGGGGGGGCCGGCCTACC
TACCCACTACGGACCACTACCTACACACCTCCCACCCCCCCCACCCTATCCACACCTACCTACGC
CGTACGGGGCCTACCGTACCGCCCCCTACCACCTACCCCTACCTCTAGCCCCCAGGGGGGGTACA
CTAGCCCTCCCCCCATACACCACGTAGGGCCCCCACGTCTGCGGCCCTACATAGTACCTAGGGGC
CCCTACCCCCTGGCCCTACCTACTACCCTCGGTACTCACCCCACTATCCCCCCCTACACCCACAG

GTACCCTACCCTACCGCTGGCCGGTACGGTACCGCCGCTACTACAGCTGTACTACTACTGGGGGC
ACCCCCACTATGCACGGGGCTGGTACTACTACGCCTACCCCACTACATACTACCACCCTCCCACG
TA-3′

A linear 11,270 bp dsDNA construct used as a substrate for packaging was prepared by PCR from this plasmid using the following primers: biotin-5'-ATGAGTGACGACTGAATCCGGTGA-3' (forward) (IDT, Inc.) and digoxygenin-5'-GGTTGTATTGATGTTGGACGAGTCGGAA-3' (reverse) (Eurofins Genomics, Inc.) using the LA Taq PCR kit (TaKaRa, Inc.). The biotin label is used to tether the DNA to streptavidin coated microspheres. The digoxygenin label was used for control experiments in which the DNA alone, in the absence of the prohead-motor complex, is tethered to anti-digoxygenin coated microspheres as described previously (63).

The 20,049 bp dsDNA 'control' non-A-philic construct with 51.8% GC content, used in control experiments, was prepared by PCR from lambda phage DNA (NEB, Inc.) using primers Biotin-5'-CTGATGAGTTCGTGTCCGTACAACTGGCGTAATC-3' (forward) (IDT, Inc.) and Digoxygenin-5'-GTGCACCATGCAACATGAATAACAGTGGGTTATC-3' (reverse) (Eurofins Genomics, Inc.) with the LA Taq PCR kit (TaKaRa, Inc.).

2.2.1.2 General Protocol of PCR Dual Labeled DNA

1. Take the following material from -20°C freezer, thawed then cooled on ice

    a. >10 µL TAKARA LA PCR Buffer II (Mg2+ free)

    b. >2 µL of ~40 mM Forward & Reverse primer

    c. >1 µL of the template DNA (50 ng/µl)

d. >16 μL 2.5mM Takara dNTP

2. Have >80 μL Purified deionized H_2O cooled on ice

3. Add water to a 0.6 mL microcentrifuge tube

4. Add 10 μL 10X Takara LA PCR Buffer II

5. Add 16 μL 2.5mM dNTP

6. Add biot-Forward primer

7. Add dig-Reverse primer

8. Add DNA template

9. Gently flip the tube to mix reagents

10. Start the thermocycler program described in Table 2.1. The lid heats to 105°C

11. Immediately take the LA Taq polymerase from -20°C freezer and add to 0.6 mL microcentrifuge tube

12. Put the polymerase back to freezer

13. Slowly stir mixture and pipet up and down ~10 times

14. Aliquot the well-mixed PCR reagent into each of the PCR tubes on ice and close the caps tightly

15. Quickly place these PCR tubes in the well of the thermocycler block

16. Close the thermocycler lid, check that the lid snugly rests on the lids of the tubes

17. Restart the thermocycler where the PCR program pauses for placing samples

18. After the program is finished, take the PCR tubes out then store them in 4°C refrigerator

2.2.1.3 PCR Solution Final Concentration

See Table 2.3.

2.2.1.4 Verify the PCR Product

It is compulsory to verify the PCR product length for T4 Packaging with A-philic content DNA because the absolute position of the A-philic content section should be exactly located.

1. Measure 0.48g of agarose and solve it into 65mL 1X TAE solution to prepare 0.8% (w/w) 1X TAE agarose gel
2. Add TE-Dye loading buffer with 2 μL of DNA samples or 1 μL of λ-HindIII ladder on DNA gel prepared in Step 1
3. Put the agarose gel into gel box (electrophoresis unit)
4. Fill gel box with 1X TAE until the gel is covered
5. Run the gel for 150 minutes at 55V
6. Check the length and concentration of PCR product on UV Spectrophotometer

2.2.2 Packaging Measurements

2.2.2.1 Optical Tweezers Measurements System

T4 phage capsids and the gp17 motor protein were prepared, complexed, and tethered to microspheres as described previously and optical tweezers measurements were conducted using the methods described previously (6,35,64,65). All the measurements were done in a solution containing 50 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$, 80 mM NaCl, 0.05 g L^{-1} BSA, and 1 mM ATP. The optical tweezers instrument was configured and calibrated as described previously (66,67). Due to variation in the sizes of individual microspheres there is uncertainty of ±135 nm (about ±400 bp) in the absolute length of DNA packaged (95% confidence interval). All measurements were recorded in 'force-clamp' mode (65), in which a feedback control system operating at 1 kHz adjusts the separation between the optical traps in 0.5 nm increments to keep the measured force constant. In the shown DNA length packaged versus time plots these data were smoothed with a 50-point moving average to reduce noise.

2.2.2.2 Preparation of T4 Antibody Coated Microspheres

It is recommended to prepare antibody T4 microspheres prior to T4 complex because it takes 5 more mins to wash the incubated antibody T4 microspheres before added into T4 complex.

1. Pipet 50 μL of protein G microspheres to a 0.6 mL microcentrifuge tube

2. Spin down microspheres in a microcentrifuge for ~1-2 min

3. Remove the supernatant

4. Wash the microspheres by resuspending them in 50 µL of 1X PBS

5. Repeat step 2-4

6. Spin down microspheres in a microcentrifuge for ~1-2 min

7. Remove the supernatant

8. Wash the microspheres by resuspending them in 5 µL of 1X PBS

9. Add 1 µL of T4 antisera and gently flip the tube to mix reagents

10. Rotate the tube on tube rotator at room temperature (~25°C) for 45 min

11. Take the tube from rotator and add 45 µL 1X PBS

12. Spin down microspheres in a microcentrifuge for ~1-2 min

13. Remove the supernatant

14. Wash the microspheres by resuspending them in 50 µl of 1x T4 buffer

15. Repeat step 12-14

16. Spin down microspheres in a microcentrifuge for ~1-2 min

17. Remove the supernatant

18. Wash the microspheres by resuspending them in 10 µL of 1X T4 buffer

19. The microspheres should be stored at 4°C and are typically usable for at most 1 week

2.2.2.3 Preparation of T4 Complexes

The final volume of the reaction is ~12 µL.

1. Add 5 μL 2.5X T4 buffer to a 0.6 ml tube

2. Add 1 μL 20 mM γ-S-ATP

3. Add 3 μL 120 bp DNA (~200 ng/μL)

4. Add 1.5 μL gp17 (~40 μM)

5. Add 3 μL T4 heads (~1.1e10/μL)

6. Gently flip the tube to mix reagents

7. Rotate the tube on tube rotator at room temperature (~25°C) for 45 min

2.2.2.4 Preparation of T4 Complex Microspheres

1. Add 3 μL of T4 coated antibody microspheres to T4 complex solution

2. Mix the sample by gently flipping the tube

3. Rotate the tube on tube rotator at room temperature (~25°C) for 45 min

4. Add all T4 complex solution to syringe solution

2.2.2.5 Preparation of T4 Complex Microsphere Syringe Solution

1. Pipet 500 μL solution of 1X T4 buffer to a 2 mL microcentrifuge tube

2. Add 10 μL T4 complex microspheres

3. Gently flip the tube to mix microspheres

4. Suck solution into 1 mL syringe

2.2.2.6 Preparation of Streptavidin Microspheres with Biotinylated DNA

1. Add 10 μL of Streptavidin microspheres (0.5 % w/v) to a 0.6 mL microcentrifuge tube
2. Spin down microspheres in a microcentrifuge for ~1-2 min
3. Remove the supernatant
4. Wash the microspheres by resuspending them in 10 μL of 1X PBS
5. Add 30-50 ng of dual labeled A-philic content DNA
6. Add 0.25 μL of 100 mg/mL BSA
7. Gently flip the tube to mix microspheres
8. Rotate the tube on a tube rotator at room temperature (~25°C) for 30 min
9. The microspheres should be stored at 4°C and are typically usable for at most 3 weeks

2.2.2.7 Preparation of DNA Microsphere Syringe Solution

1. Pipet 500 μL solution of 1X T4 buffer to a 2 mL microcentrifuge tube
2. Add 3 μL DNA microspheres
3. Gently flip the tube to mix microspheres
4. Suck solution into 1 mL syringe

2.2.2.8 Preparation of T4 Packaging Buffer

The T4 packaging buffer is diluted from 10X T4 buffer then added 50 μL of 100mM ATP. This T4 packaging buffer is prepared at room temperature and could store in 4°C refrigerator for reused in next day experiment.

2.2.2.9 Microsphere Capture Tweezers Operation

1. Flowing enough purified water (> 2 mL) into the flow chamber to wash out the detergents;
2. Flowing enough purified water (> 2 mL) into each of the microsphere chambers to wash out the detergents and microspheres;
3. Bump T4 packaging buffer into the flow chamber such that the entire chamber is filled with packaging buffer;
4. Add enough microsphere solution from up and down microsphere chambers until the microspheres are visibly flowing out of the capillary tube;
5. Adjust the flowrate by using a tube-crimping ‘valve’ on the polyethylene tubing until the microspheres are easily captured by optical traps but the flowrate is fast enough to pump ATP into environment.

2.2.3 Motor Velocity and Pausing & Slipping Analyses

DNA packaging rate versus time was calculated, as in previous studies (64), by linear fitting of the length versus time data in a 0.5 s sliding time window slid in 5 ms increments.

Negative control data in which fixed length DNA molecules, tethered without the head-motor complex, were recorded to determine the effect of noise/drift on the measurements were analyzed in the same manner. The term 'motor velocity' is used to refer to DNA translocation rate during active translocation, omitting periods where pauses or slips occurred. As in previous studies, sections of data in which active translocation occurred were identified based on a velocity threshold criterion considering the effect of noise/drift measured in control experiments with fixed-length tethered DNA molecules (64). Windows in which rate was >55 bp/s were scored as active translocation (97.8% confidence) and sections with rate <-55 bp/s were scored as slipping (97.8% confidence). Average velocities (for packaging specified DNA segments with specified applied forces) were calculated by averaging velocities in all time windows scored as having active translocation over all events. Uncertainties in average motor velocity were estimated using the bootstrap method (68).

2.2.3.1 Bacteriophage Packaging Rate and Motor Velocity Calculations

Computing Velocity by Sliding Temporal Window

A 0.5 second width window slide from the start packaging track to the end which sliced the data into 500 points per segment (with the 1000 Hz sampling rate). The speed of a trapped bead was the slope of the linear fitting of time and bead displacement that computed by Matlab *fit* function. The bead speed computed by sliding time window represented the measured bacteriophage packaging rate during the moment of t to $t+0.5$ second in the position from x_{start} to

x_{end}. In this A-philic project, the packaging rate of A-philic section only counted the segments of which the middle point was in the Section II.

Computing Velocity by Sliding Spatial Window

The Sliding Spatial Window was similar to the temporal method except the window slides along translocation axis. A spatial sliding window cropped packaging track into segments of 500bp (~170 nm). The bead translocation speed for each segment was computed by the same linear fitting method as temporal window method but it represented the speed of the middle point of section. In this A-philic project, the packaging rate of A-philic section only counted the segments of which the middle point was in the Section II.

2.2.3.2 Noise Measurement and Packaging/Pausing/Slipping Criteria

A segment of static bead position data (trapped bead speed = 0) was collected by stretching a double-labelled A-philic DNA at 5pN. Optical traps position and simultaneous noise were recorded for 10 minutes. This static data was cropped into small segments and used for calculating the bead fluctuation as described in 2.2.3.1. The average speed of bead was -0.0536 bp/s and the standard deviation was 27.4 bp/s.

A threshold-based algorithm of judging which class a segment was classified was listed below. The threshold (criteria) was acquired from static bead fluctuation calculation as discussed above. A segment with linear fitting slope less than 54 bp/s but more than -54 bp/s, i.e, (-2σ, $+2\sigma$), was classified as pause class. The segment with slope more than 54bp/s was classified as packaging segment and the segment with slope less than -54bp/s was treated as slipping segment.

2.2.3.3 Alignment and A-Philic Section Identification

There was a measurement error brought by T4 packaging initiation with using the protocol of T4 packaging experiment and the sensitivity of the absolute position for each oligo. The measurement errors came from 3 aspects:

First, there was diameter error on polystyrene particles that not all beads were in the same size. Meanwhile, the beads were not uniform spheres so that the average diameter measured by bead provider was untrustworthy. In practice, we caught the beads with similar diameter in observation. The criteria to determine if the trapped beads were acceptable were judged by experimenter's empirical determination. For increase the preciseness, an object detection program was implemented to decide whether the caught beads were in acceptable range.

Second, the initial point of packaging contributed another kind of error. T4 packages DNA in ~700bp/second. The experiment initialization for each complex took 1-3 seconds. It meant there were at most 1500bp alignment error if we aligned the package traces at the beginning. Notice polystyrene particles could be bumped together after packaging was done, we aligned the packaging traces at the end point, i.e., the point when 2 beads touched each other to reduce zeroing error.

Third, the beads in the optical traps underwent power supply baseline drifting and Brownian motion. The baseline drifting was from temperature affects that the devices room was sealed to reduce air fluctuation which caused heat cumulates in the room. To reduce heat drifting, the optical system built in a covered lower thermal expansion coefficient box which

helped insulate the optical components (as well as reduce air fluctuations and to keep out ambient light). The Brownian noise was reduced by smoothing packaging traces.

End Point Alignment for T4

The T4 packaging experiment was done with force-claim method in which the DNA was stretched by constant force. T4 phage continuously packaged the DNA until the whole DNA was swallowed into the prohead. At that moment, 2 beads touched each another and extra DNA on the Streptavidin coated bead would interact with the Protein G coated bead. This DNA and Protein G bead entanglement caused multiple tethering signal when the control program separates 2 microspheres. Those ambiguous signal induced mis-classification during data analysis. To avoid this problem, the data collection for each packaging activity might be stopped before 2 beads got close and left enough space. On another hand, one would like to collect more data in one T4 phage packaging event, i.e., let the phage packaging to the end of DNA. To tradeoff these 2 requirements, the force-claim was set to stop when the moveable bead was 1500bp far away to the fixed one. In data processing, packaging track of T4 with A-philic DNA were aligned at the same ending point which was 9830bp.

2.2.3.4 Normalized Motor Velocity

T4 bacteriophage packaging rate varied from one T4 complex to another ranging from 300 to 1800 bp/s. It brought high external variance while finding the average rate over different complexes. Hence, a normalization pre-procedure was needed to reduce external statistics error. This normalization method computed the average of motor velocity of all sections as the

normalization factor. Then all of the speeds got from 2.2.3.1 divided this normalization factor so that the external variance vanished.

2.2.4 Packaging Rate Fluctuation Analyses

Fourier transforms of DNA length packaged versus time were calculated for each event using the FFT function in Matlab (R2019, Mathworks, Inc.). The average FFT was calculated by averaging computed FFT amplitudes in frequency bins over all events. The same analysis was done for the negative control datasets (measurements on fixed length DNA tethers) to characterize the measurement noise. Signal-to-noise ratio (SNR) was calculated by dividing the average FFT amplitudes calculated for signal to those calculated for the control datasets.

2.2.5 Correlation Analyses for Packaging Rate

Packaging rate versus position along DNA template was calculated by linear fitting of the DNA length versus time data in a sliding 500 bp window, slid in steps of 5 bp. Correlations between rates measured in one event versus another event were analyzed by calculating Pearson correlation coefficients. The following procedure was implemented to account for the effect that the roughly ±400 bp uncertainty in absolute position would have on the detection of correlations. The correlation coefficient for each pair of events was recalculated after shifting one of the datasets by ±100, ±200, ±300, and ±400 bp and the maximum value was determined for each pair. If two events are statistically correlated this shifting procedure would detect the larger correlation that would occur when pairs are more closely aligned in absolute DNA position.

These maximum correlation coeffients were determined for all pairs of datasets and averaged to define the ‘correlation score’. By definition, Pearson correlation coefficients have values between -1 and 1 and would average to zero for a large ensemble of pairs of completely uncorrelated datasets. Because we calculate the maximum correlation coefficients for different position shifts, we obtain non-zero positive average values for the correlation scores. To determine whether these values indicate statistically significant correlations we compare correlation scores when correlating pairs of events recorded with the same sequence versus when cross-correlating pairs of events recorded with different sequences. Datasets recorded for the linear ~11 kbp plasmid sequence and the ~20 kbp phage sequence, both linear molecules, were used. When cross-correlating, data recorded with the first ~10 kbp of the plasmid sequence were correlated against data recorded with either the first 10 kbp or second 10 kbp of the phage sequence.

2.2.6 Correlation Analyses for Pausing and Slipping

Analyses for correlations between pauses and slips occurring in pairs of different events were conducted using a similar approach as that described above used to analyze correlations in packaging rate. Specifically, we first computed pausing and slipping frequencies versus position in a sliding 500 bp window for each event. Then correlation scores were computed for pairs of these records in the same manner as described above.

2.3 Results

2.3.1 DNA Construct Design to Analyze Sequence Dependence of Packaging Motor Function

We used optical tweezers to measure motor-driven packaging of single DNA molecules into single phage T4 heads using the techniques we developed previously (6,35,65). In brief, a head-motor complex is attached to one microsphere held in one steerable optical trap and a DNA molecule is attached by one end to a second microsphere held in a second optical trap. To initiate packaging, the two microspheres are brought into near contact in a buffer solution containing ATP, to allow the motor to grip the free end of the DNA and begin translocating it into the head. To test for translocation, we move the two microspheres apart while measuring the force on the second microsphere. If a DNA is tethered the force rises as it is pulled taut between the two microspheres. We then turn on a 'force-clamp' feedback control system which adjusts the separation between the two traps to maintain a specified constant applied force. If the motor is translocating the DNA, a rapid decrease in the separation between the two traps is observed as the two microspheres are pulled closer together. In this manner, we track the length of the DNA packaged versus time under a constant applied load force (external force opposing translocation).

To test whether DNA sequence would affect motor function, and to test the B-A scrunchworm model specifically, we designed a linear DNA construct containing a ~2 kbp synthetic A-philic sequence inserted into the middle of a ~9 kbp 'normal' (non-A-philic) plasmid DNA sequence, illustrated schematically in Fig. 2.2A. This approach allowed us to measure the

effects of sequence within a single packaging event, especially when the motor transitions from the flanking plasmid segments into the A-philic segment, or vice-versa.

Our synthetic sequence was designed to be A-philic based on principles determined by experiments by Minchenkova et al. (55), Ivanov et al. (69), and Tolstorukov et al. (41) which measured B-A transitions for 24 different 9-14 bp sequences to determine how ΔG_{BA} depends on sequence. These studies established that both high GC content and the presence of certain dimers and trimers cause sequences to be more A-philic. The results were shown to be well fit by an empirical 'T-32' model with experimentally-determined parameters (41). The B-A scrunchworm model, which predicts that a ~10 bp segment in the motor channel is induced to transition to A-form, specifically proposed that these ΔG_{BA} values for short DNA segments are relevant to motor function (50). The A-philic sequence we designed has ~64% GC content whereas the flanking plasmid DNA sequence has only ~46% (Fig. 2.2B). In addition, 60% of the included dimers are ones classified by Tolstorukov et al. as A-philic (versus 53% of those in the flanking plasmid DNA) and 33% of the included trimers are ones classified as highly A-philic (versus only 14% in the flanking plasmid DNA). The ΔG_{BA} values calculated with the T-32 model for this sequence are significantly lower than for the flanking plasmid DNA sequences (Fig. 2.2C), implying that transition to A-form is more easily induced. In the B-A scrunchworm model, translocation of A-philic DNA segments is predicted to be less-inhibited by an applied load force.

2.3.2 A-philic DNA Sequences do not Alter Motor Velocity

Single DNA molecule packaging measurements with the linear plasmid DNA sequence were made at saturating ATP concentration (1 mM) using both low applied force (5 pN), where

the motor translocates DNA at nearly maximum speed, and high applied force (30 pN), where the motor is slowed by ~60% (6). 30 pN is the estimated maximum internal force the motor experiences during the end stages of packaging the full-length viral genome (14,30,70,71). We collected N=134 packaging events at 5 pN and N=45 events at 30pN. It is considerably more difficult to obtain measurements with the higher force because tethered complexes often detach from the microspheres before the measurement is completed, likely due to dissociation of the antibody-capsid bonds (63). Examples of measured length of DNA packaged versus time records are shown in Figs. 3A and 3B. From these datasets one can see that there is variability in the motor velocity, as we have reported previously. However, there are no obvious changes when the motor transitions from the flanking plasmid section (blue) into A-philic (red) section, or vice versa.

We analysed the full ensemble of packaging events and found no significant differences in the average motor velocity when packaging the flanking plasmid versus A-philic segments at both low (5 pN) and high (30 pN) applied forces (Fig. 2.4A and Table 2.1). As in our previous work (6), we define 'motor velocity' as the rate of DNA translocation during active packaging, i.e. excluding pauses and slips. Additional 'negative control' measurements were conducted using a ~20 kbp phage DNA construct. Like the non-A-philic sections of the flanking plasmid sequences, this sequence has much lower average GC content (51.8%) and a much higher average ΔG_{BA} (0.6487 kcal/mol) than the synthetic A-philic sequence. We again found no significant differences in average motor velocity (Fig. 2.4A and Table 2.1).

Uncertainties in the determination of the average velocities were ~4% for the 5 pN measurements and ~10% for the 30 pN measurements. These uncertainties are mostly due to the fact that, as we reported previously, different individual T4 packaging events exhibit different

average motor velocities (6). Thus, an alternative and more sensitive test is to examine the ratios of motor velocities measured during each packaging event when a single motor packages through A-philic versus flanking plasmid sections of the same DNA molecule. We calculated the ratios of motor velocity when packaging the A-philic segment to that when packaging the flanking plasmid segments for each event and then calculated the average ratio over all events. For both the 5 pN and 30 pN data we find that the average ratio is close to unity (Fig. 2.4B and Table 2.1), again providing evidence that there is no significant effect of the A-philic sequence on motor velocity. As expected, the uncertainties in the ratios are much lower, ~2.5% for the 5 pN data and ~5% for the 30 pN data, and thus establish stricter bounds on the null effect of sequence. As a control, we also calculated ratios in the same manner for motor velocities measured when packaging segments of the control phage DNA (which has no A-philic section) starting/ending at the same positions that delineate the flanking plasmid versus A-philic sections in the DNA sequence. As expected, the average ratios determined by this analysis are close to unity (Fig. 2.4B and Table 2.1).

2.3.3 Effects of A-philic DNA Sequences on Motor Pausing and Slipping

As reported in previous studies, T4 and other phage packaging motors exhibit occasional pauses where the translocation stops transiently, and slips, where the motor transiently loses grip on the DNA resulting in rapid release of packaged DNA under the applied force (5,64). To investigate whether the substantially different A-philic sequence affects pausing or slipping, we analysed the data to determine the percent time the motor paused or slipped during packaging at

both 5 pN and 30 pN applied forces (Table 1). No significant differences were observed in pausing between the A-philic, flanking plasmid, and control sequences with either applied force. However, modest differences were detected in percent time slipping (Table 1). The largest difference in percent time slipping was 9.38% ± 1.14% for the A-philic DNA versus 15.85% ± 2.3% with the control phage DNA sequence with the higher 30 pN force. Although modest, this difference is statistically significant (P=0.007) suggesting that the strength of the motor's grip on DNA can be affected by DNA sequence. However, it is important to note that this difference does not cause a significant change in the overall rate of DNA translocation.

2.3.4 Tests for General Sequence Dependence of Motor Function

The data presented above provide evidence that an A-philic DNA sequence with high GC-content does not significantly affect motor function. We also sought to investigate if *any* DNA sequence differences, aside from A-philic propensity or GC content, could influence motor function. As reported previously, the T4 packaging rate fluctuates considerably not only between different packaging events but also during each packaging event (6). Examples of the latter are shown in Fig. 2.5A. To demonstrate that these fluctuations are not simply due to measurement noise/drift, we conducted control force-clamp measurements with fixed-length DNA molecules tethered between microspheres (in the absence of head-motor complexes and thus the absence of DNA translocation) to characterize the instrumental noise/drift. These measurements show that fluctuations in measured translocation rate caused by noise/drift are much smaller than those observed during DNA packaging (Fig. 2.5A).

To characterize the timescales of the translocation rate fluctuations we computed Fourier transforms (FFTs) averaged over all datasets, and then signal-to-noise ratio (SNR) versus frequency by dividing the average FFT amplitudes for the packaging data by the average FFT amplitudes calculated for the control noise/drift measurements (Fig. 2.5B). This analysis shows that the inherent packaging rate fluctuations span frequencies from at least ~0.05 to 1 Hz. The cause of these fluctuations is unknown, but one possible hypothesis is that they could be caused by variations in the substrate DNA sequence. To test this hypothesis, we first calculated packaging rate versus position along DNA for each packaging event (Fig. 2.6A) and then looked for correlations in the fluctuations across different events when packaging the same sequence versus different sequences. We used datasets of ~10 kbp recorded with the linear plasmid-derived DNA construct and two different segments of the longer phage DNA.

An important consideration in this analysis is accounting for the effect of uncertainty in the measurement of absolute lengths of DNA packaged, which is about ±135 nm or ±400 bp. This uncertainty is caused by variations in sizes of individual trapped microspheres. To account for the effect of this uncertainty in the analysis, we defined a 'correlation score' that considers the effect of relative position shifts between pairs of data sets (see Methods). An important control in this analysis is comparing correlation scores across pairs of events when packaging the same sequence with those across pairs of events when packaging different sequences. Together, these analyses covered a large variety of sequence space since three different ~10 kbp-long sequences were tested. No significant differences were found in the correlation scores (Fig. 2.6B), and thus no evidence that the packaging rate fluctuations are attributable to DNA sequence. We further investigated the occurrence of pauses or slips using a similar correlation score analysis (see Methods) and again found no significant differences when correlating pairs of events when

packaging the same DNA sequence versus pairs of events when packaging different sequences (Fig. 2.6C and 2.6D). Thus, we found no evidence that the pausing or slipping positions are influenced by DNA sequence.

2.4 Discussion

One of the key questions in viral genome packaging is how DNA sequence and/or structure influence packaging motor-DNA substrate interactions and dynamics of translocation. Previous experimental data implicate potential sequence effects, and one theoretical model hypothesized that structural transitions between B and A forms might be important for force generation and DNA movement. We addressed these issues through single-molecule optical tweezers measurements of DNA translocation by the phage T4 motor.

Our studies show that there are no differences in average motor velocities of the T4 motor, within the experimental uncertainties, when the motor is packaging A-philic and high GC DNA sequences versus the non-A-philic plasmid and phage sequences. This finding contrasts with the B-A scrunchworm model which predicts that the A-philic sequence would be more rapidly packaged against an externally applied load force. The model predicts that motor force is proportional to ΔG_{AB} (50). Since motor velocity decreases with increasing applied force, the predicted effect of the A-philic sequence on motor velocity is calculated by multiplying forces by the ratio of ΔG_{AB} calculated for the flanking plasmid sequence to that for the A-philic sequence (6,50). As indicated by the dashed lines in Fig. 2.4, a negligible difference is predicted for the 5 pN force, consistent with our findings, but a significant ~20% higher velocity is predicted for the A-philic sequence with a 30 pN force. In contrast, our measurements with 30 pN force find an

average ratio of velocities for A-philic versus flanking plasmid sections of 0.9739 ± 0.0478, which indicates no significant change is caused by the A-philic, high GC content sequence to within an experimental uncertainty of ~5%. This finding provides evidence against the B-A scrunchworm model and further shows that GC content does not strongly influence motor function.

Our studies examined translocation dynamics at low prohead filling, where internal forces resisting packaging are negligible. The B-A scrunchworm model predicts that packaging of A-philic DNA would be less-inhibited by an external force but more-inhibited by 'internal forces' resisting DNA packaging that occur during the latter stages of capsid filling. We are not presently able examine this regime because it is technically challenging to manipulate DNA of the full ~170 kbp T4 genome length with optical tweezers. Based on studies of phage phi29 and theoretical models (14,30,71) internal forces are estimated to rise to ~20-30 pN near the end of packaging. In this case the B-A scrunchworm model would predict a lower translocation rate for the A-philic DNA in the late stages of packaging. However, since our present results provide evidence against the B-A scrunchworm model, we do not believe this is a likely outcome. A related consideration is that sequence could influence physical properties such as DNA curvature and bendability and thus potentially influence internal forces by influencing the conformation and dynamics of the packaged DNA (38,72,73). But again, we would expect such effects to only potentially influence motor dynamics in the high capsid filling regime.

After the B-A scrunchworm model was proposed (50), computational studies on DNA within the phi29 connector channel (albeit lacking the packaging ATPase) supported the argument that DNA might be driven to a scrunched conformation (62,74). However, this conformation was even shorter than the standard A-form and the effect was found to be

independent of DNA sequence (62), which is consistent with our experimental results. An alternative class of models based on structures of several viral motor proteins proposes that DNA translocation is primarily driven by lever-like conformational changes in the motor protein ATPase (24-26,28). However, it remains possible that a scrunchworm-type mechanism in which expansion and contraction of the threaded DNA segment could play a role in translocation but without involving an A-form DNA structure. Additional and more recent simulations of DNA threaded into the portal rings of phages phi29, T4, and P22 showed that contraction or lengthening of the threaded DNA segment in the channel can occur due to the electrostatic potentials generated by the portal rings (75). This led to a proposed 'electrostatic scrunchworm' model. It proposes that ATPase-driven conformational changes in the proteins of the motor/portal complex lead to cyclical changes in the electrostatic potential of the portal channel causing DNA scrunching-unscrunching motions. These are coupled to a protein-DNA grip and release cycle to rectify the DNA motion (75).

Recent structures of the phage P23-45 portal channel (76,77) may have implications for scrunchworm-type models. The channel at its narrowest point was observed to have a different conformation ~8 Å wider in empty procapsids than in expanded capsids. It was suggested that the portal may have this more 'open' conformation during DNA packaging at low procapsid filling and transition to the tighter-fitting conformation at high filling to restrict DNA slipping. It is then possible the portal protein might not interact strongly enough with the DNA to induce structural transitions. However, these studies describe static portal structures in the absence of the motor ATPase and DNA. Interactions between the motor and portal could affect their conformations and interactions with DNA during packaging, or the ATPase could induce DNA transitions instead of the portal.

One recently published study and two preprints provide complementary information consistent with our findings. First, it was shown that 25-35 bp DNA:RNA duplexes can be packaged by the T4 motor (78). Since such duplexes are expected to have A-form structure and not undergo B-to-A transitions, this argues against the B-A scrunchworm model. Second, a recent preprint reports that the phage phi29 motor can also translocate both DNA:RNA duplexes as well as RNA:RNA duplexes and the motor's step size changes to match their shorter helical pitch (79). Third, a recent preprint reports a cryo-electron microscopy structure of stalled phi29 packaging complex in which the five motor subunits are arranged in a helical 'lock-washer' structure with symmetry complementary to the DNA substrate (80). It is proposed that the phi29 motor may function by cycling between this helical structure and a planar one, a mechanism which attributes translocation to conformational changes of the motor complex rather than of the DNA. Interestingly, it was noted that the threaded DNA in this structure is 'stretched or partially unwound in some places, compressed in others, and has a prominent kink'. Whether these features represent dynamic changes in the DNA conformation that could play a role in the motor mechanism remains an open question.

Our measurements revealed modest effects of sequence on motor slipping. A lower percent time slipping was measured with the A-philic, high GC sequence than with the control phage DNA sequence. However, the amount of slipping observed with the flanking sections of the non-A-philic plasmid sequence was similar to that observed with the A-philic sequence, indicating that it is not A-philic property or high GC content that causes the difference. Our findings are consistent with experimental (49) and computational (62) evidence from the phi29 system suggesting that the motor and/or portal proteins make contact with the DNA bases during translocation steps. The detailed nature of these contacts could influence the strength of the

motor's grip on DNA in a manner that depends on the identity of the bases. It has been shown that binding of ATP induces the motor to transition into a conformation where it tightly grips DNA and that slipping can occur due to either transient ATP dissociation or force-induced rupture of the motor's grip (36,61,64). It is conceivable that either effect could be influenced by the interacting DNA sequence, the former potentially via allosteric effects (81). On the other hand, the small difference in percent time slipping does not result in a significant difference in overall translocation rate. Moreover, slips do not occur at the same positions in every packaging event, and the observed change in percent time slipping with sequence at high force is less than the average percent time slipping. These findings suggest that while slipping propensity can be affected by sequence it is primarily a stochastic temporal process.

No significant differences in motor pausing were observed with the A-philic, GC rich sequence. The cause of motor pausing is not completely clear, but it has been observed with all the three well-characterized motors from T4, lambda, and phi29 phages. In the case of T4, pausing was attributed to misaligned DNA in the motor channel when the ATP binding site was unoccupied (82). Alternatively, studies of lambda motor mutants exhibiting altered pausing suggested that that pausing can occur due to binding of ATP in a misaligned orientation that leads to temporary blockage of hydrolysis (36). In addition, studies of the phi29 motor revealed that an increase in pausing at high prohead filling (>75% of genome length packaged) is attributable to non-equilibrium dynamics of the tightly packed DNA via fluctuating internal forces and/or allosteric regulation of motor function (20,70). This latter effect is not relevant to the present measurements with T4 because we are measuring at low prohead filling. Our finding of no evidence for sequence dependence suggests that pausing is mainly a stochastic temporal process.

Our additional analyses with three different ~10 kbp DNA sequences found no evidence that packaging rate fluctuations, pausing, or slipping were correlated with position along the DNA templates. A limitation of this analysis is that there are limits on the time resolution of measurements of translocation rate because they require calculating the derivative of the DNA length versus time data which is affected by noise. For the analyses in Fig. 2.6, we calculated packaging rate and frequency of slipping and pausing in a sliding 500 bp length window, so these analyses do not address whether shorter sequences could influence packaging dynamics unless such sequences occurred with significantly different probabilities in different 500 bp windows. However, the results in Fig. 2.5B show that large fluctuations in the packaging rate that we sought to explain occur at frequencies from 0.05 to 1 Hz. Since the average DNA translocation rate is ~700 bp/s, these rate fluctuations occur on DNA length scales ranging from 700 to 14,000 bp, which implies that a 500 bp sliding window is appropriate for the analysis and we can conclude that these fluctuations are not caused by sequence dependence. We also conducted additional correlation analyses using a smaller 200 bp window size and again found no evidence for sequence-dependent packaging dynamics (Supplemental Data Fig. 2.9).

Since pauses and slips are abrupt events which likely occur at specific positions along DNA it seems likely they could be influenced by DNA sequences shorter than 200 bp. During translocation the motor likely contacts a small section of the threaded DNA comprising perhaps just ~1-10 basepairs (49). The probability to slip or pause may vary depending on the identities of those basepairs. Our measurements with the engineered high GC DNA segment provide clear evidence that G-C and C-G basepairs don't cause major differences versus A-T or T-A basepairs. As mentioned above, differences caused by sequences <200 bp could be detected if the sequences of interest occur with significantly different frequency within different 200-500 bp

sections of the DNA template. Any sequences of interest, of any length, could be tested using the methods presented here by using engineered DNA templates containing a section enriched in these sequences.

To investigate whether the synthetic A-philic sequence we designed exhibits other physical properties distinct from those of the flanking plasmid sequences we used the 'plot.it' and 'bend.it' software packages (40). Differences in several properties are predicted (Fig. 2.10 and Table 2.2). The most notable is that average roll angle between basepairs (72) is predicted to be ~5-fold higher for the A-philic sequence. Roll angles can be positive or negative, but for the A-philic segment they are predicted to be predominantly positive. Average twist angle (72) is predicted to be ~0.6° lower for the A-philic section, which is a significant difference since the standard error in the mean (SEM) is 0.034°. Intrinsic curvature (38) is predicted to be 43% lower (SEM = 6.9%) and average bendability (73) 7% higher (SEM = 0.72%). Average free energy (ΔG) of melting (39), a measure of duplex stability, is predicted to be 8% higher (SEM = 1.3%). Since our measurements found no significant differences in DNA translocation dynamics when packaging the A-philic versus flanking plasmid segments, our results provide evidence that these DNA property changes do not significantly affect motor function.

In summary, our analyses of a viral genome packaging motor show that variable DNA sequences do not significantly affect the function of the motor. No influences on motor translocation rate or pausing were detected. Only modest differences in slipping were detected, which suggest that the motor's grip on DNA can vary with sequence, but this does not significantly affect the overall DNA packaging rate. The finding that an engineered A-philic DNA sequence has no significant effect on motor velocity with a high applied load force provides evidence against the B-A scrunchworm model (50). These results do not provide

evidence for or against the electrostatic scrunchworm model (75). Overall, our results suggest that motor velocity fluctuations, pausing, and slipping are primarily stochastic temporal events. These insights impose constraints on the plausible packaging models for motor function, particularly those requiring structural changes in the DNA substrate for DNA movement. Furthermore, insensitivity to sequence relieves any sequence dependent genome packaging constraints on virus evolution. Otherwise, it could create evolutionary bottlenecks for regulatory sequences that control viral life cycle such as transcription, replication, recombination, and repair.

As additional data become available from molecular genetic analyses of the motor proteins, and from high-resolution structural studies of the packaging motor complexes, more detailed models for motor force generation in the phage DNA packaging system will be developed. The results reported here, along with those from other single-molecule experiments (9,31,64,70,82-84), will be critical for evaluating those models, and for developing a complete structural, kinetic and thermodynamic understanding of how these motors work.

Acknowledgement and Funding

We thank Mariam Ordyan and Mounir Fizari for advice and assistance with optical tweezers measurements.

This work was supported by the US National Institutes of Health [grant numbers R01GM118817 to D.E.S., AI081726 to V.B.R.] and National Science Foundation [grant number MCB-0923873 to V.B.R]. Funding for open Funding for open access charge: US National Institutes of Health [grant numbers R01GM118817, AI081726].

Chapter 2, in full, is a reprint of the material as it appears in the journal Nucleic Acids Research. Mo, Y., Keller, N., delToro, D., Ananthaswamy, N., Harvey, S. C., Rao, V. B., & Smith, D. E. (2020). Function of a viral genome packaging motor from bacteriophage T4 is insensitive to DNA sequence. *Nucleic Acids Research*, 48(20), 11602-11614. The dissertation author was the primary investigator and author of this paper.

Table 2.1 Measured parameters characterizing motor function. Each parameter is an average over all events and uncertainties are reported as standard errors. 'Normal' refers to measurements with the non-A-philic flanking sections I and III of the linear plasmid DNA sequence, 'A-philic' to section II, and 'Control' to the phage DNA sequence.

DNA Segment and applied force	Average Velocity (bp/sec)	Velocity Ratio	% Time Pausing	% Time Slipping	# of Events
A-philic, 5pN	731.85 ± 30.51	1.02 ± 0.02564	1.63% ± 0.57%	2.36% ± 0.59%	134
Normal, 5pN	736.9 ± 30.88	1 (by definition)	2.88% ± 0.66%	1.02% ± 0.46%	134
Control, 5pN	734.6 ± 30.7	0.9870 ± 0.0364	1.60% ± 0.17%	2.97% ± 0.2%	139
A-philic, 30pN	293.09 ± 28.85	0.9739 ± 0.0478	13.10% ±1.41%	9.38% ± 1.14%	45
Normal, 30pN	303.35 ± 29.17	1 (by definition)	12.29% ± 1.18%	11.43% ± 1.11%	45
Control, 30pN	313.7 ± 29	1.0349 ± 0.0644	11.59% ± 1.32%	15.85% ± 2.3%	51

Table 2.2 Mean predicted properties for the plasmid DNA construct containing the synthetic A-philic segment. Means and standard errors in the means (SEM) for the properties plotted in Fig. 2.10. 'Normal' refers to values for the flanking plasmid segments. 'Difference' refers to the percent difference between the means for the A-philic and Normal sections.

Parameter	A-philic Mean	A-philic SEM	Normal Mean	Normal SEM	Difference
(A) Roll	1.47	0.048	0.23	0.035	539%
(B) Twist	33.03	0.034	33.65	0.015	-2%
(C) Bendability	5.38	0.039	5.04	0.030	7%
(D) ΔG melt	-2.05	0.027	-1.89	0.010	8%
(E) Curvature	2.18	0.150	3.81	0.124	-43%

Table 2.3 PCR Solution Mixing Order and Concentration.

Total	20 µL	100 µL	Final Concentration
H_2O	14.283 µL	71.667 µL	
10X Buffer	2 µL	10 µL	1X
60.6mM Forward Primer	0.165 µL	0.825 µL	0.5 µM
66mM Reverse Primer	0.1516 µL	0.758 µL	0.5 µM
DNA template	0.10 µL	0.25 µL	50 ng
2.5mM dNTP	3.2 µL	16 µL	0.2 mM
Taq Polymerase	0.1 µL	0.5 µL	0.025U/µL

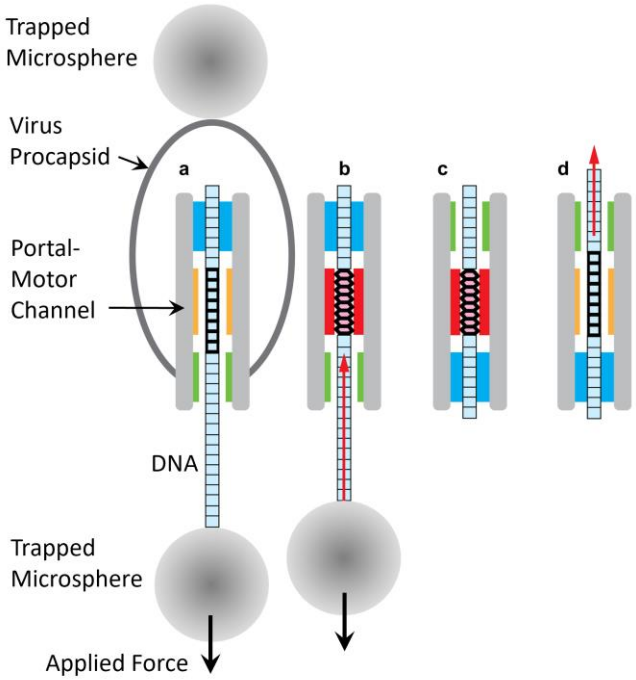

Figure 2.1 Schematic illustration of the single DNA molecule packaging measurement and B-A scrunchworm model (modified from S. Harvey, J. Struct. Biol. 189 (2015); note that the elements are not drawn to scale). (a) The viral procapsid-motor complex is attached to one trapped microsphere and the unpackaged end of the DNA is attached to a second trapped microsphere. Translocation by the motor pulls the microspheres together and a resisting force is applied as indicated. The hypothetical mechanism proposed by the B-A scrunchworm model is also illustrated: (b) The portal/motor induces a threaded section of DNA (pink) to transition to a shortened A-form structure while an upper section is gripped, which results in ~2.5 bp of DNA being pulled into the channel (red arrow). (c) The upper section of DNA is released, and a lower section of DNA is gripped. (d) The grip is released, and the threaded DNA is induced to transition to the longer B-form structure, resulting in translocation of ~2.5 bp of DNA into the procapsid (red arrow). The portal/motor then cycles back to state (a). This model may be contrasted with other types of models in which the DNA is assumed to have a static structure and conformational changes in the motor protein complex are proposed to drive DNA translocation steps.

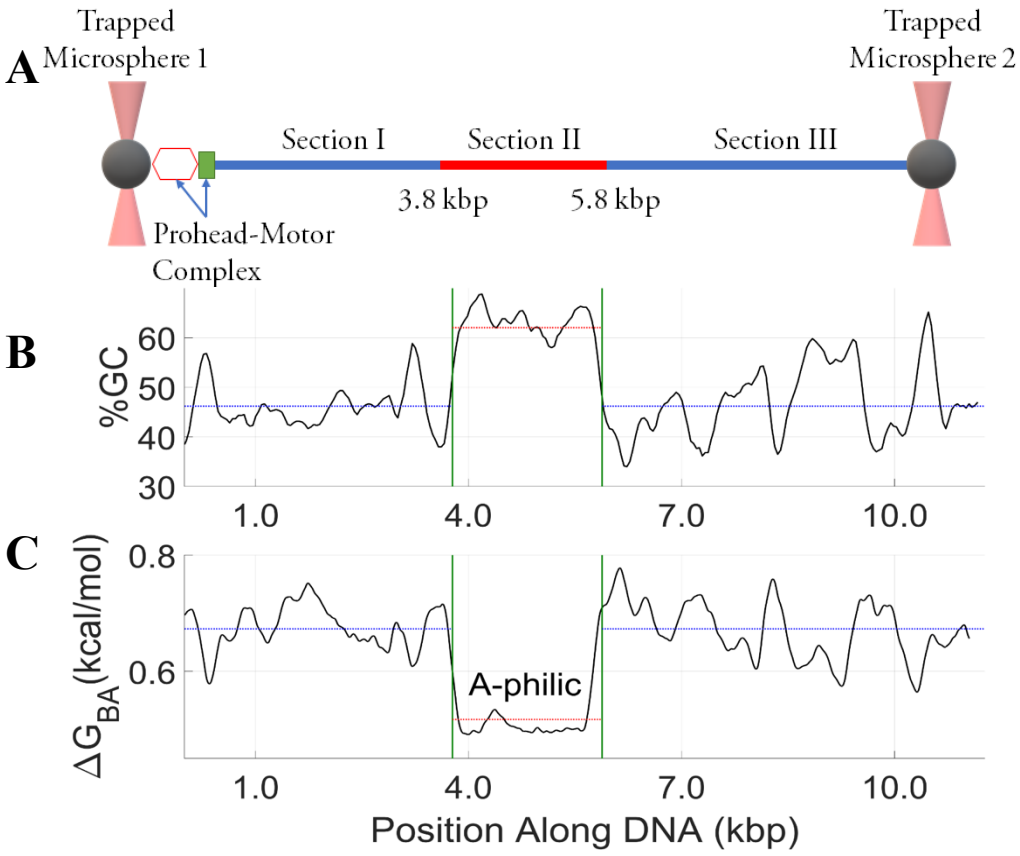

Figure 2.2 Schematic illustration of the DNA construct design and packaging measurement. (A) T4 prohead-motor complexes are attached to one microsphere trapped with optical tweezers (left) and DNA is attached to a second trapped microsphere (right). A linear plasmid DNA sequence was designed with a synthetic A-philic, high GC content sequence (section II) between two 'normal' non-A-philic flanking plasmid DNA sequences (sections I and III). DNA translocation by the motor begins at left and proceeds from section I to II to III. (B) Average % GC versus position along the DNA calculated in a 200 bp sliding window. (C) Predicted free energy difference per basepair (ΔG_{BA}) between B-form and A-form DNA structures versus position calculated in a 200 bp sliding window. In panels (B) and (C) the vertical green lines indicate the beginning and end of section II, the dashed blue lines indicate average % GC and ΔG_{BA} for sections I & III, and the dashed red line indicates average % GC and ΔG_{BA} in section II.

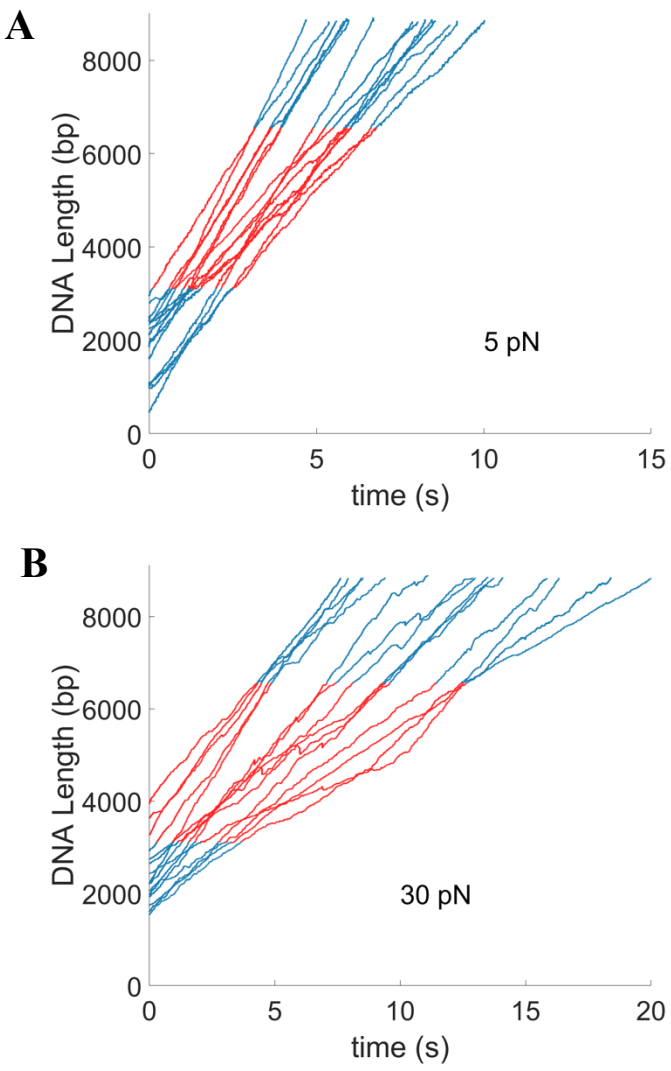

Figure 2.3 Measurements of length of DNA translocated versus time with the linear plasmid DNA sequence. Parts of the dataset containing the non-A-philic flanking plasmid sequences (sections I and III) are indicated in blue and parts containing the synthetic A-philic, high % GC sequence (section II) are indicated in red. (A) Measurements with 5 pN applied force (examples from N=134 recorded events) (B) Measurements with 30 pN applied force (examples from N=45 recorded events).

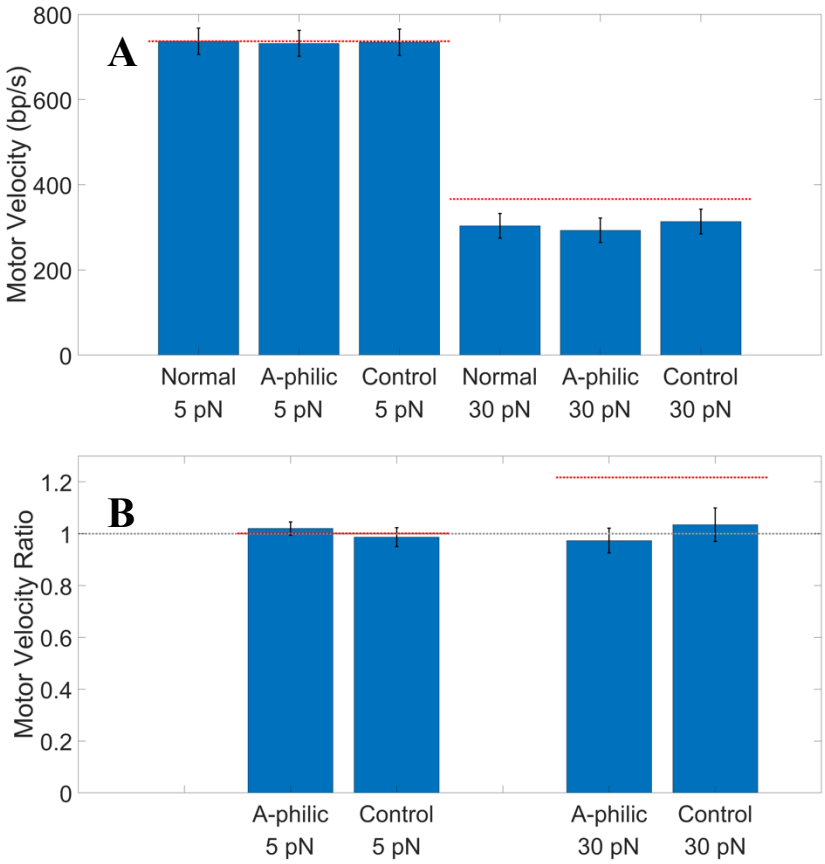

Figure 2.4 (A) Average motor velocities. 'Normal' refers to measurements with the non-A-philic flanking sections I and III of the linear plasmid DNA sequence, 'A-philic' to section II, and 'Control' to the phage DNA sequence. Averages were determined from 134 recorded events at 5 pN and 45 recorded events at 30 pN for the linear plasmid sequence and 80 events at 5 pN and 50 events at 30 pN for the control DNA. Error bars indicate standard errors in the means. The dashed red lines indicate the velocity for the A-philic sequence predicted by the B-A scrunchworm model and force-velocity relationship. **(B)** Average velocity ratios. Bars labelled 'A-philic' refer to the average ratio of the velocity when packaging the A-philic segment to that when packaging the normal flanking plasmid segments calculated for each event. Bars labelled 'control' refer to velocity ratios calculated in the same manner for events recorded with the phage DNA. Error bars indicate standard errors. The dashed grey line indicates a ratio of '1', expected if there is no sequence dependence. The dashed red lines indicate ratios predicted by the B-A scrunchworm model and measured force-velocity relationship.

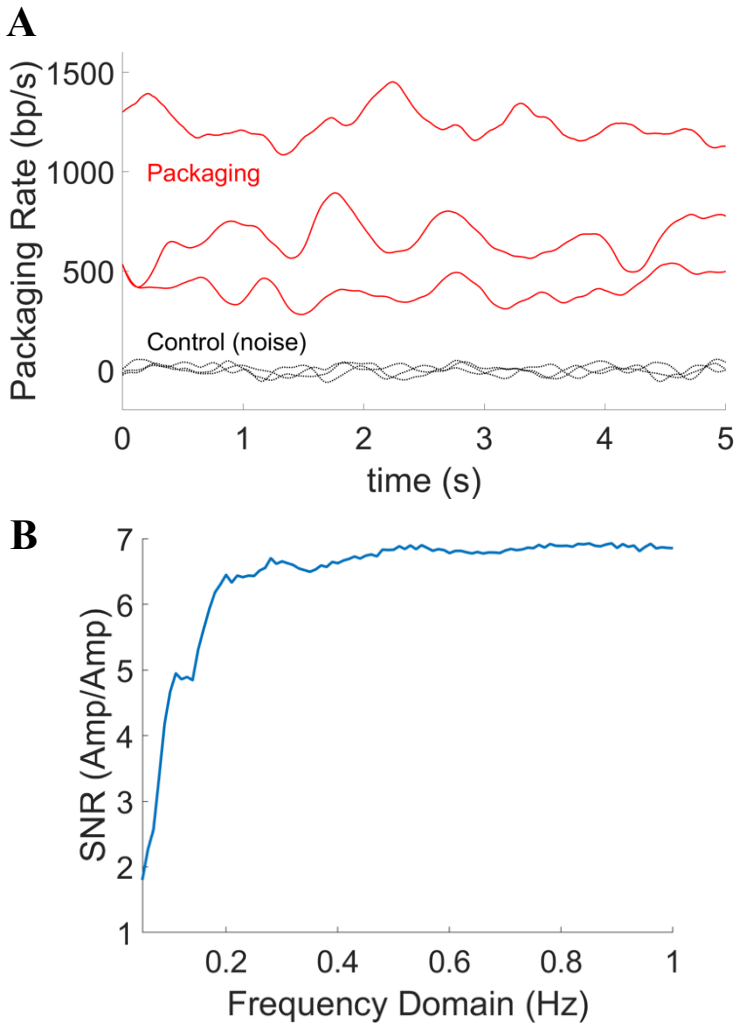

Figure 2.5 (A) Examples of measured packaging rate versus time with the linear plasmid DNA sequence with 5 pN applied force (red lines). Shown for comparison are rate fluctuations caused by Brownian and instrumental noise measured in control experiments with fixed tethered DNA molecules (black lines). **(B)** Signal-to-noise ratio versus frequency determined by calculating the average Fourier transform amplitudes for packaging rate measurements ('signal') (from N=134 packaging events) and dividing them by the average amplitudes calculated from control ('noise') rate measurements.

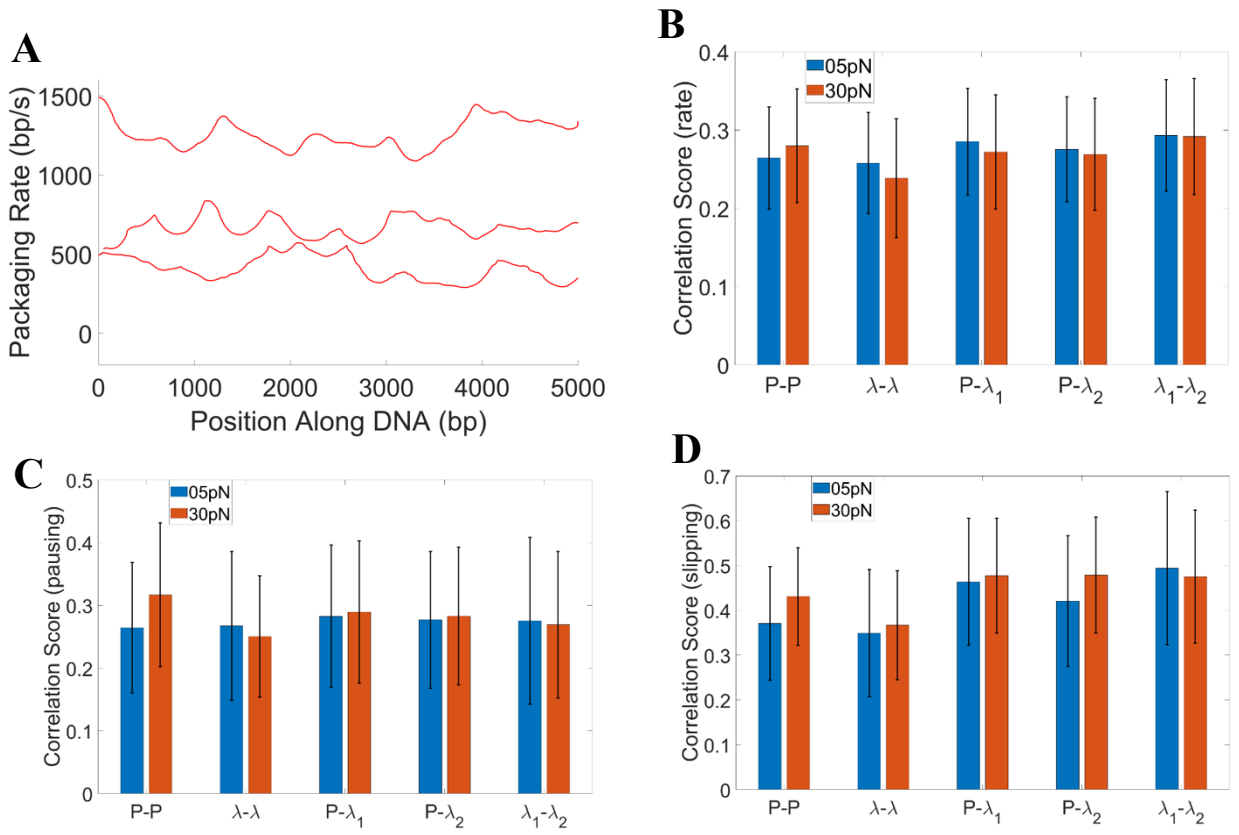

Figure 2.6 Correlation analyses. (A) Examples of measured packaging rate versus position along DNA measured with the linear plasmid DNA sequence. **(B)** Correlations analyzed between rate versus position along DNA measured in one event and that measured in other events when packaging either the same sequence or different sequences. 'Correlation score' is defined in the methods section. 'P' refers to the linear plasmid DNA sequence, λ to the control phage DNA, λ_1 to one 10 kbp section of the control DNA, and λ_2 to another 10 kbp section of the control DNA (see Methods). 'P-P' indicates correlation score for pairs of events where P was packaged. 'λ-λ' refers pairs of events where λ was packaged. 'P-λ_1' and 'P-λ_2' refer to correlations between P datasets and λ_1 or λ_2 datasets. 'λ_1-λ_2' refers to correlations between λ_1 and λ_2 datasets. Results were determined from 134 recorded events at 5 pN and 45 recorded events at 30 pN for the linear plasmid sequence and 80 events at 5 pN and 50 events at 30 pN for the control DNA. Error bars for all plots indicate one standard deviation. **(C)** Correlations analyzed between pausing versus position measured in one event and that measured in other events, as in (**B**). Correlation score for pausing is defined in methods. **(D)** Correlations analyzed between slipping versus position measured in one event and that measured in other events, as in (**B**). Correlation score for slipping is defined in methods.

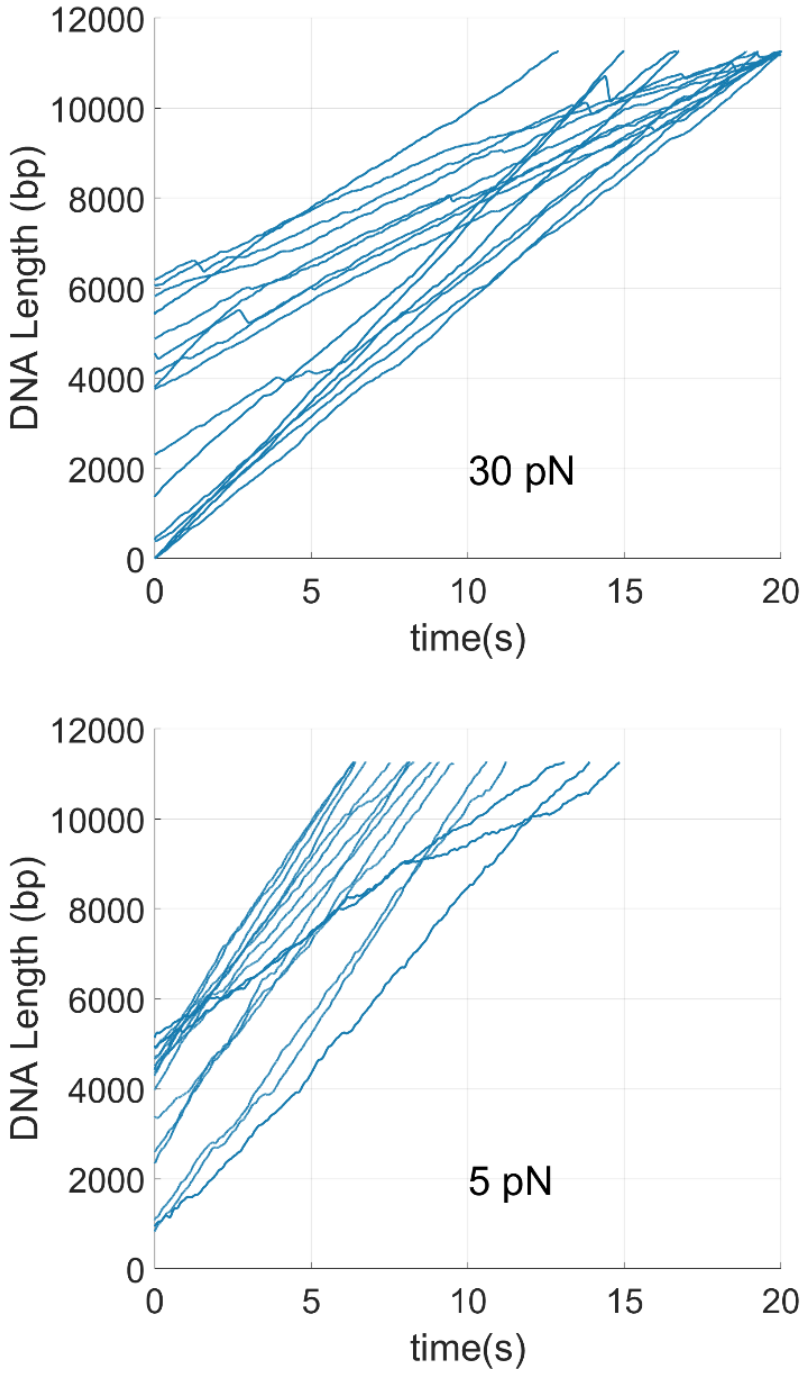

Figure 2.7 Measurements of length of DNA translocated versus time with the control phage DNA construct. Left plot are measurements with 5 pN applied force and right plot are measurements with 30 pN applied force.

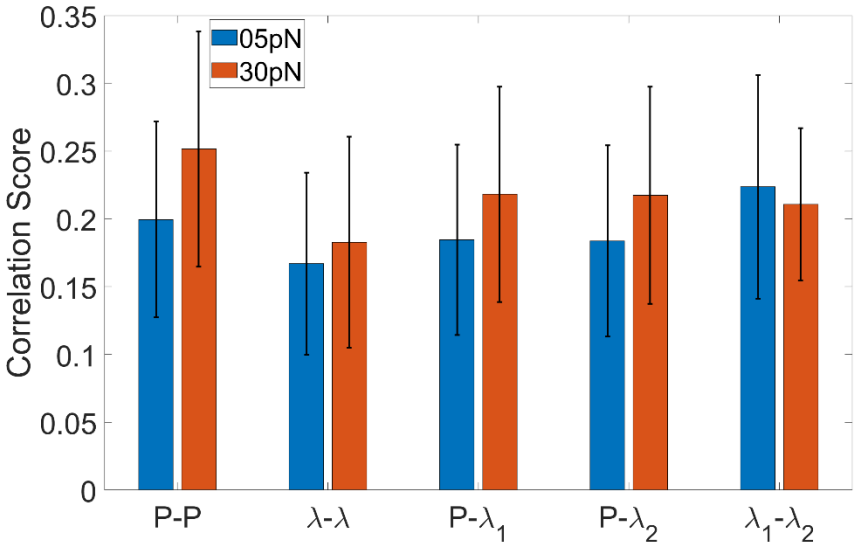

Figure 2.8 Additional packaging rate correlation analyses. These analyses were performed in the same manner as described for Fig. 2.5B, except that a 200 bp sliding window was used instead of a 500 bp window (see methods).

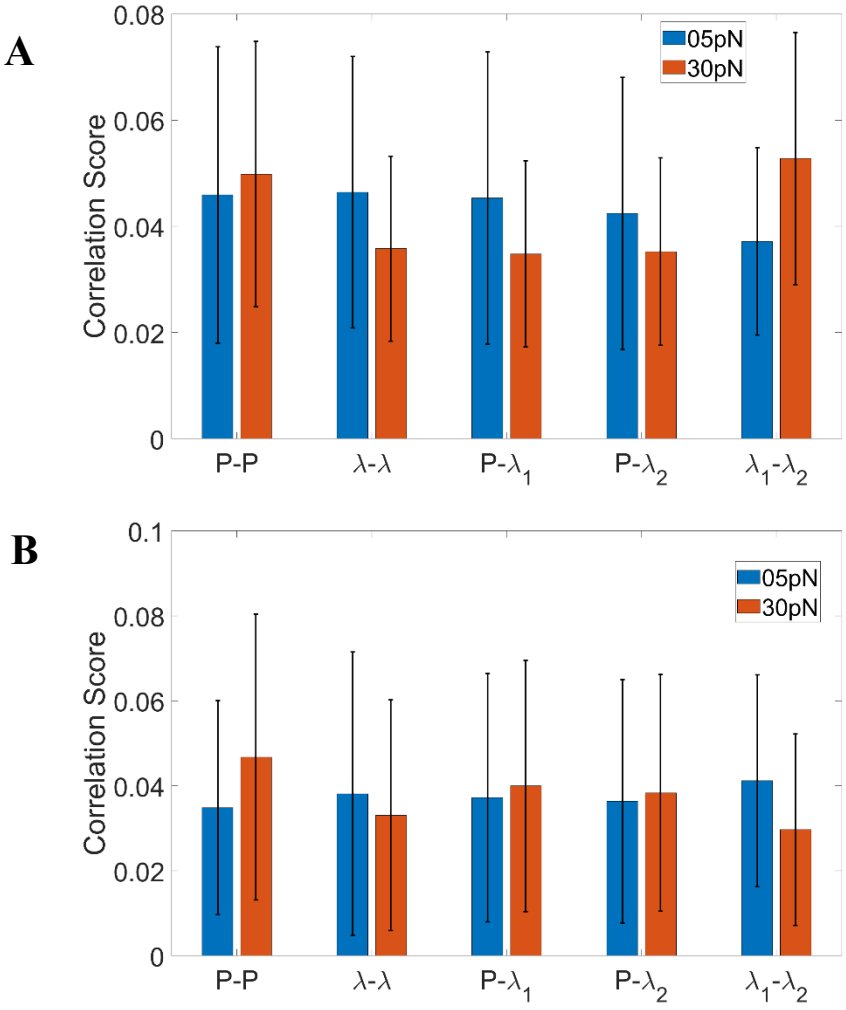

Figure 2.9 Additional pausing and slipping correlation analyses. These analyses, for **(A)** pausing and **(B)** slipping, were performed in the same manner as described for Fig. 2.5C and 2.5D, except that a 200 bp sliding window was used instead of a 500 bp window (see methods).

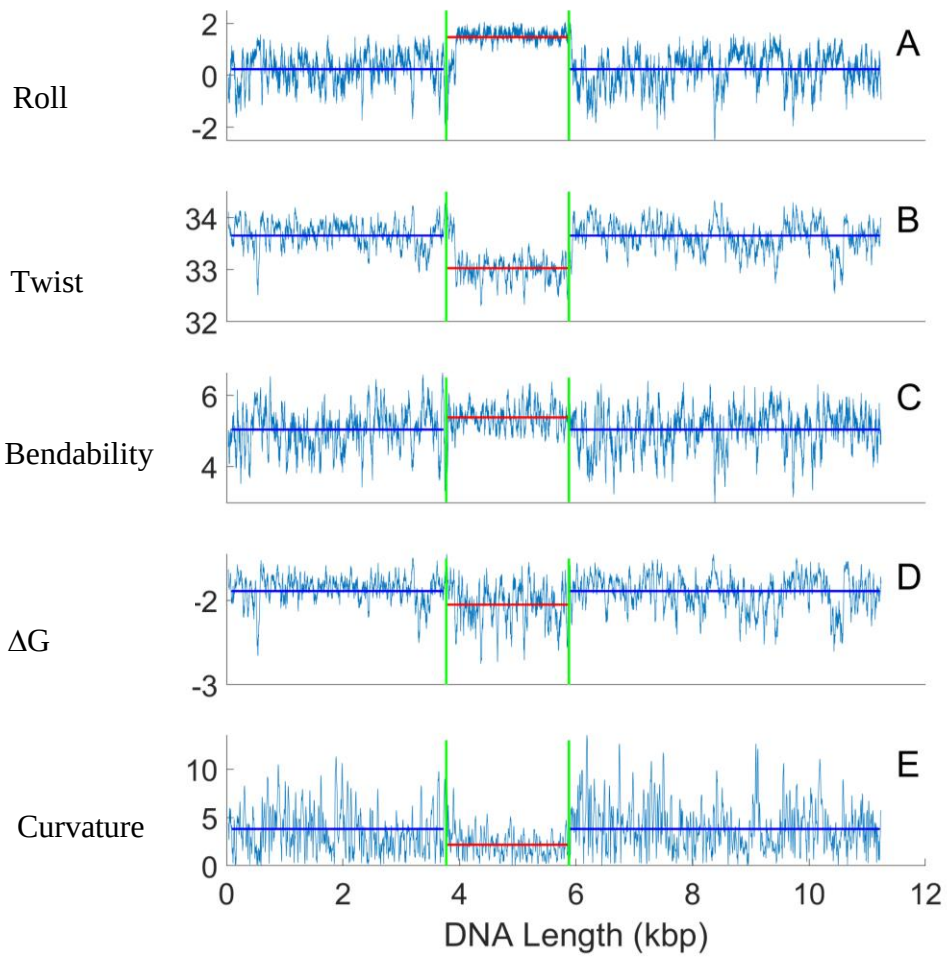

Figure 2.10 Predicted sequence-dependent properties for the plasmid DNA construct containing the synthetic A-philic segment. Predictions calculated using the 'plot.it' and 'bend.it' software packages using default settings (30 bp window size). All plots use the same horizontal axis at the bottom. Vertical green lines indicate the beginning and end of the synthetic A-philic segment. Horizontal red lines indicate the average property values for the A-philic section and horizontal blue lines indicate average property values for the flanking plasmid segments. (**A**) Roll angle (degrees; from conformational energy calculations). (**B**) Twist angle (degrees per helical turn; from conformational energy calculations). (**C**) Bendability (degrees; from DNAse I digestion experiments). (**D**) Free energy (ΔG) of melting (kcal/mol; from calorimetric studies). (**E**) Intrinsic curvature (degrees; consensus scale from DNAse I and nucleosome positioning data).

References

1. Rao, V.B. and Feiss, M. (2015) Mechanisms of DNA packaging by large double-stranded DNA viruses. *Ann Rev Virology*, **2**, 351-378.
2. Casjens, S.R. (2011) The DNA-packaging nanomotor of tailed bacteriophages. *Nature Reviews Microbiology*, **9**, 647-657.
3. Feiss, M. and Rao, V.B. (2012) In Rao, V. and Rossmann, M. G. (eds.), *Viral Molecular Machines*. Springer, New York, NY, pp. 489-509.
4. Smith, D.E., Tans, S.J., Smith, S.B., Grimes, S., Anderson, D.L. and Bustamante, C. (2001) The bacteriophage phi29 portal motor can package DNA against a large internal force. *Nature*, **413**, 748-752.
5. Fuller, D.N., Raymer, D.M., Rickgauer, J.P., Robertson, R.M., Catalano, C.E., Anderson, D.L., Grimes, S. and Smith, D.E. (2007) Measurements of single DNA molecule packaging dynamics in bacteriophage lambda reveal high forces, high motor processivity, and capsid transformations. *J Mol Biol*, **373**, 1113-1122.
6. Fuller, D.N., Raymer, D.M., Kottadiel, V.I., Rao, V.B. and Smith, D.E. (2007) Single phage T4 DNA packaging motors exhibit large force generation, high velocity, and dynamic variability. *Proc Nat Acad Sci USA*, **104**, 16868-16873.
7. Moffitt, J.R., Chemla, Y.R., Aathavan, K., Grimes, S., Jardine, P.J., Anderson, D.L. and Bustamante, C. (2009) Intersubunit coordination in a homomeric ring ATPase. *Nature*, **457**, 446-450.
8. Smith, D.E. (2011) Single-molecule studies of viral DNA packaging. *Current Opinion in Virology*, **1**, 134.
9. Liu, S., Tafoya, S. and Bustamante, C. (2017) Deciphering the Molecular Mechanism of the Bacteriophage φ 29 DNA Packaging Motor. *Optical Tweezers: Methods and Protocols*, 343-355.
10. Lyubimov, A.Y., Strycharska, M. and Berger, J.M. (2011) The nuts and bolts of ring-translocase structure and mechanism. *Curr Opin Struct Biol*, **21**, 240-248.
11. Gelbart, W.M. and Knobler, C.M. (2008) The Physics of Phages. *Physics Today*, **61**, 42-47.
12. Speir, J.A. and Johnson, J.E. (2012) Nucleic acid packaging in viruses. *Curr Opin Cell Biol*, **22**, 65-71.
13. Arsuaga, J., Tan, R.K.Z., Vazquez, M., Sumners, D.W. and Harvey, S.C. (2002) Investigation of viral DNA packaging using molecular mechanics models. *Biophysical chemistry*, **101**, 475-484.
14. Tzlil, S., Kindt, J.T., Gelbart, W.M. and Ben-Shaul, A. (2003) Forces and pressures in DNA packaging and release from viral capsids. *Biophys J*, **84**, 1616-1627.

15. Purohit, P.K., Inamdar, M.M., Grayson, P.D., Squires, T.M., Kondev, J. and Phillips, R. (2005) Forces during bacteriophage DNA packaging and ejection. *Biophys J*, **88**, 851-866.

16. Ali, I., Marenduzzo, D. and Yeomans, J.M. (2006) Polymer packaging and ejection in viral capsids: Shape matters. *Physical Review Letters*, **96**, 208102-208102.

17. Forrey, C. and Muthukumar, M. (2006) Langevin dynamics simulations of genome packing in bacteriophage. *Biophys J*, **91**, 25-41.

18. Fuller, D.N., Rickgauer, J.P., Jardine, P.J., Grimes, S., Anderson, D.L. and Smith, D.E. (2007) Ionic effects on viral DNA packaging and portal motor function in bacteriophage phi 29. *Proc Nat Acad Sci USA*, **104**, 11245-11250.

19. Petrov, A.S. and Harvey, S.C. (2008) Packaging double-helical DNA into viral capsids: Structures, forces, and energetics. *Biophysical journal*, **95**, 497-502.

20. Berndsen, Z.T., Keller, N., Grimes, S., Jardine, P.J. and Smith, D.E. (2014) Nonequilibrium dynamics and ultraslow relaxation of confined DNA during viral packaging. *Proc Nat Acad Sci USA*, **111**, 8345–8350.

21. Keller, N., Grimes, S., Jardine, P.J. and Smith, D.E. (2014) Repulsive DNA-DNA interactions accelerate viral DNA packaging in phage phi29. *Phys Rev Lett*, **112**, 248101.

22. Black, L.W. (2015) Old, new, and widely true: The bacteriophage T4 DNA packaging mechanism. *Virology*, **479**, 650-656.

23. Grimes, S., Jardine, P.J. and Anderson, D. (2002) Bacteriophage phi 29 DNA packaging. *Advances in Virus Research*, **58**, 255-294.

24. Sun, S., Kondabagil, K., Draper, B., Alam, T.I., Bowman, V.D., Zhang, Z., Hegde, S., Fokine, A., Rossmann, M.G. and Rao, V.B. (2008) The Structure of the Phage T4 DNA Packaging Motor Suggests a Mechanism Dependent on Electrostatic Forces. *Cell*, **135**, 1251-1262.

25. Zhao, H., Christensen, T.E., Kamau, Y.N. and Tang, L. (2013) Structures of the phage Sf6 large terminase provide new insights into DNA translocation and cleavage. *Proc Nat Acad Sci USA*, **110**, 8075-8080.

26. Hilbert, B.J., Hayes, J.A., Stone, N.P., Duffy, C.M., Sankaran, B. and Kelch, B.A. (2015) Structure and mechanism of the ATPase that powers viral genome packaging. *Proc Natl Acad Sci U S A*, **112**, E3792-3799.

27. Xu, R.-G., Jenkins, H.T., Antson, A.A. and Greive, S.J. (2017) Structure of the large terminase from a hyperthermophilic virus reveals a unique mechanism for oligomerization and ATP hydrolysis. *Nucleic Acids Res*, **45**, 13029-13042.

28. Mao, H., Saha, M., Reyes-Aldrete, E., Sherman, M.B., Woodson, M., Atz, R., Grimes, S., Jardine, P.J. and Morais, M.C. (2016) Structural and Molecular Basis for Coordination in a Viral DNA Packaging Motor. *Cell Rep*, **14**, 2017-2029.

29. Yang, Y., Yang, P., Wang, N., Chen, Z., Su, D., Zhou, Z.H., Rao, Z. and Wang, X. (2020) Architecture of the herpesvirus genome-packaging complex and implications for DNA translocation. *Protein & Cell*, 1-13.

30. Liu, S., Chistol, G., Hetherington, C.L., Tafoya, S., Aathavan, K., Schnitzbauer, J., Grimes, S., Jardine, P.J. and Bustamante, C. (2014) A Viral Packaging Motor Varies Its DNA Rotation and Step Size to Preserve Subunit Coordination as the Capsid Fills. *Cell*, **157**, 702-713.

31. Chistol, G., Liu, S., Hetherington, C.L., Moffitt, J.R., Grimes, S., Jardine, P.J. and Bustamante, C. (2012) High degree of coordination and division of labor among subunits in a homomeric ring ATPase. *Cell*, **151**, 1017-1028.

32. Rao, V.B. and Mitchell, M.S. (2001) The N-terminal ATPase site in the large terminase protein gp17 is critically required for DNA packaging in bacteriophage T4. *Journal of Molecular Biology*, **314**, 401-411.

33. Mitchell, M.S. and Rao, V.B. (2006) Functional analysis of the bacteriophage T4 DNA-packaging ATPase motor. *J Biol Chem*, **281**, 518-527.

34. Tsay, J.M., Sippy, J., DelToro, D., Andrews, B.T., Draper, B., Rao, V., Catalano, C.E., Feiss, M. and Smith, D.E. (2010) Mutations altering a structurally conserved loop-helix-loop region of a viral packaging motor change DNA translocation velocity and processivity. *J Biol Chem*, **285**, 24282-24289.

35. Migliori, A.D., Keller, N., Alam, T.I., Mahalingam, M., Rao, V.B., Arya, G. and Smith, D.E. (2014) Evidence for an electrostatic mechanism of force generation by the bacteriophage T4 DNA packaging motor. *Nature Communications*, **5**, 4173.

36. delToro, D., Ortiz, D., Ordyan, M., Sippy, J., Oh, C.S., Keller, N., Feiss, M., Catalano, C.E. and Smith, D.E. (2016) Walker-A Motif Acts to Coordinate ATP Hydrolysis with Motor Output in Viral DNA Packaging. *J Mol Biol*, **428**, 2709-2729.

37. Ortiz, D., delToro, D., Ordyan, M., Pajak, J., Sippy, J., Catala, A., Oh, C.-S., Vu, A., Arya, G. and Feiss, M. (2018) Evidence that a catalytic glutamate and an 'Arginine Toggle' act in concert to mediate ATP hydrolysis and mechanochemical coupling in a viral DNA packaging motor. *Nucleic Acids Research*, **47**, 1404-1415.

38. Goodsell, D.S. and Dickerson, R.E. (1994) Bending and curvature calculations in B-DNA. *Nucleic Acids Research*, **22**, 5497.

39. Breslauer, K.J., Frank, R., Blöcker, H. and Marky, L.A. (1986) Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences*, **83**, 3746-3750.

40. Vlahovicˇek, K., Kajan, L. and Pongor, S. (2003) DNA analysis servers: plot. it, bend. it, model. it and IS. *Nucleic Acids Research*, **31**, 3686-3687.

41. Tolstorukov, M., Ivanov, V., Malenkov, G., Jernigan, R. and Zhurkin, V. (2001) Sequence-dependent B↔ A transition in DNA evaluated with dimeric and trimeric scales. *Biophys J*, **81**, 3409-3421.

42. Ivanov, V. and Minchenkova, L. (1994) The A-form of DNA: in search of the biological role. *Molekuliarnaia biologiia*, **28**, 1258-1271.

43. Lu, X.-J., Shakked, Z. and Olson, W.K. (2000) A-form conformational motifs in ligand-bound DNA structures. *J Mol Biol*, **300**, 819-840.

44. Olson, W.K. and Zhurkin, V.B. (2000) Modeling DNA deformations. *Current Opinion in Structural Biology*, **10**, 286-297.

45. Ivanov, V.I., Minchenkova, L.E., Chernov, B.K., McPhie, P., Ryu, S., Garges, S., Barber, A.M., Zhurkin, V.B. and Adhya, S. (1995) CRP-DNA Complexes: Inducing the A-likeForm in the Binding Sites with an Extended Central Spacer. *J Mol Biol*, **245**, 228-240.

46. Cue, D. and Feiss, M. (2001) Bacteriophage lambda DNA packaging: DNA site requirements for termination and processivity. *Journal of Molecular Biology*, **311**, 233-240.

47. Pease, P.J., Levy, O., Cost, G.J., Gore, J., Ptacin, J.L., Sherratt, D., Bustamante, C. and Cozzarelli, N.R. (2005) Sequence-directed DNA translocation by purified FtsK. *Science*, **307**, 586-590.

48. Ptacin, J.L., Nollmann, M., Becker, E.C., Cozzarelli, N.R., Pogliano, K. and Bustamante, C. (2008) Sequence-directed DNA export guides chromosome translocation during sporulation in Bacillus subtilis. *Nature Struc & Mol Biol*, **15**, 485-493.

49. Aathavan, K., Politzer, A.T., Kaplan, A., Moffitt, J.R., Chemla, Y.R., Grimes, S., Jardine, P.J., Anderson, D.L. and Bustamante, C. (2009) Substrate interactions and promiscuity in a viral DNA packaging motor. *Nature*, **461**, 669-673.

50. Harvey, S.C. (2015) The scrunchworm hypothesis: Transitions between A-DNA and B-DNA provide the driving force for genome packaging in double-stranded DNA bacteriophages. *J Struc Biol*, **189**, 1-8.

51. Ray, K., Sabanayagam, C.R., Lakowicz, J.R. and Black, L.W. (2010) DNA crunching by a viral packaging motor: Compression of a procapsid-portal stalled Y-DNA substrate. *Virology*, **398**, 224-232.

52. Dixit, A.B., Ray, K. and Black, L.W. (2012) Compression of the DNA substrate by a viral packaging motor is supported by removal of intercalating dye during translocation. *Proceedings of the National Academy of Sciences*, **109**, 20419-20424.

53. Kulkarni, M. and Mukherjee, A. (2017) Understanding B-DNA to A-DNA transition in the right-handed DNA helix: Perspective from a local to global transition. *Progress in Biophysics and Molecular Biology*, **128**, 63-73.

54. Ivanov, V.I. and Krylov, D.Y. (1992), *Methods in Enzymology*. Elsevier, Vol. 211, pp. 111-127.

55. Minchenkova, L.E., Schyolkina, A.K., Chernov, B.K. and Ivanov, V.I. (1986) CC/GG Contacts Facilitate the B to A Transition of DNA in Solution. *Journal of Biomolecular Structure and Dynamics*, **4**, 463-476.

56. Knee, K.M., Dixit, S.B., Aitken, C.E., Ponomarev, S., Beveridge, D. and Mukerji, I. (2008) Spectroscopic and molecular dynamics evidence for a sequential mechanism for the A-to-B transition in DNA. *Biophysical Journal*, **95**, 257-272.

57. Hormeno, S., Moreno-Herrero, F., Ibarra, B., Carrascosa, J.L., Valpuesta, J.M. and Arias-Gonzalez, J.R. (2011) Condensation prevails over BA transition in the structure of DNA at low humidity. *Biophysical Journal*, **100**, 2006-2015.

58. DiMaio, F., Yu, X., Rensen, E., Krupovic, M., Prangishvili, D. and Egelman, E.H. (2015) A virus that infects a hyperthermophile encapsidates A-form DNA. *Science*, **348**, 914-917.

59. Wang, M.D., Schnitzer, M.J., Yin, H., Landick, R., Gelles, J. and Block, S.M. (1998) Force and velocity measured for single molecules of RNA polymerase. *Science*, **282**, 902-907.

60. Keller, D. and Bustamante, C. (2000) The mechanochemistry of molecular motors. *Biophysical journal*, **78**, 541-556.

61. Chemla, Y.R., Aathavan, K., Michaelis, J., Grimes, S., Jardine, P.J., Anderson, D.L. and Bustamante, C. (2005) Mechanism of force generation of a viral DNA packaging motor. *Cell*, **122**, 683-692.

62. Waters, J.T., Kim, H.D., Gumbart, J.C., Lu, X.-J. and Harvey, S.C. (2016) DNA Scrunching in the Packaging of Viral Genomes. *The Journal of Physical Chemistry B*, **120**, 6200-6207.

63. Fuller, D.N., Gemmen, G.J., Rickgauer, J.P., Dupont, A., Millin, R., Recouvreux, P. and Smith, D.E. (2006) A general method for manipulating DNA sequences from any organism with optical tweezers. *Nucleic Acids Res*, **34**, e15.

64. Ordyan, M., Alam, I., Mahalingam, M., Rao, V.B. and Smith, D.E. (2018) Nucleotide-dependent DNA gripping and an end-clamp mechanism regulate the bacteriophage T4 viral packaging motor. *Nature Communications*, **9**, 1-9.

65. Keller, N., delToro, D. and Smith, D.E. (2018) In Lavelle, C. (ed.), *Molecular Motors Methods and Protocols*. Humana Press, New York, NY.

66. Rickgauer, J.P., Fuller, D.N. and Smith, D.E. (2006) DNA as a metrology standard for length and force measurements with optical tweezers. *Biophys J*, **91**, 4253-4257.

67. delToro, D. and Smith, D.E. (2014) Accurate measurement of force and displacement with optical tweezers using DNA molecules as metrology standards. *Appl Phys Lett*, **104**, 143701.

68. Efron, B. and Tibshirani, R.J. (1994) *An introduction to the bootstrap*. CRC press.

69. Ivanov, V., Minchenkova, L., Minyat, E. and Schyolkina, A. (1983), *Cold Spring Harbor Symposia on Quantitative Biology*. Cold Spring Harbor Laboratory Press, Vol. 47, pp. 243-250.

70. Berndsen, Z.T., Keller, N. and Smith, D.E. (2015) Continuous Allosteric Regulation of a Viral Packaging Motor by a Sensor that Detects the Density and Conformation of Packaged DNA. *Biophys J*, **108**, 315-324.

71. Keller, N., Berndsen, Z.T., Jardine, P.J. and Smith, D.E. (2017) Experimental comparison of forces resisting viral DNA packaging and driving DNA ejection. *Phys Rev E*, **95**, 052408.

72. De Santis, P., Palleschi, A., Savino, M. and Scipioni, A. (1990) Validity of the nearest-neighbor approximation in the evaluation of the electrophoretic manifestations of DNA curvature. *Biochemistry*, **29**, 9269-9273.

73. Brukner, I., Sanchez, R., Suck, D. and Pongor, S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *The EMBO journal*, **14**, 1812-1818.

74. Kumar, R. and Grubmüller, H. (2016) Phi29 Connector-DNA Interactions Govern DNA Crunching and Rotation, Supporting the Check-Valve Model. *Biophysical journal*, **110**, 455-469.

75. Sharp, K.A., Lu, X.-J., Cingolani, G. and Harvey, S.C. (2019) DNA conformational changes play a force-generating role during bacteriophage genome packaging. *Biophysical journal*, **116**, 2172-2180.

76. Bayfield, O.W., Steven, A.C. and Antson, A.A. (2020) Cryo-EM structure in situ reveals a molecular switch that safeguards virus against genome loss. *Elife*, **9**, e55517.

77. Bayfield, O.W., Klimuk, E., Winkler, D.C., Hesketh, E.L., Chechik, M., Cheng, N., Dykeman, E.C., Minakhin, L., Ranson, N.A. and Severinov, K. (2019) Cryo-EM structure and in vitro DNA packaging of a thermophilic virus with supersized T= 7 capsids. *Proceedings of the National Academy of Sciences*, **116**, 3556-3561.

78. Black, L.W., Yan, B. and Ray, K. (2020) The T4 TerL Prohead Packaging Motor Does Not Drive DNA Translocation by a Proposed Dehydration Mechanism. *Viruses*, **12**, 522.

79. Castillo, J.P., Tong, A., Tafoya, S., Jardine, P.J. and Bustamante, C. (2020) A DNA translocase operates by cycling between planar and lock-washer structures. *Preprint posted at https://www.biorxiv.org/content/10.1101/2020.05.22.101154v1.abstract*.

80. Woodson, M., Pajak, J., Zhao, W., Zhang, W., Arya, G., White, M.A., Jardine, P.J. and Morais, M.C. (2020) A viral genome packaging motor transitions between cyclic and helical symmetry to translocate dsDNA. *Preprint posted at https://www.biorxiv.org/content/10.1101/2020.05.23.112524v2*.

81. Besprozvannaya, M., Pivorunas, V.L., Feldman, Z. and Burton, B.M. (2013) SpoIIIE protein achieves directional DNA translocation through allosteric regulation of ATPase activity by an accessory domain. *Journal of Biological Chemistry*, **288**, 28962-28974.

82. Kottadiel, V.I., Rao, V.B. and Chemla, Y.R. (2012) The dynamic pause-unpackaging state, an off-translocation recovery state of a DNA packaging motor from bacteriophage T4. *Proceedings of the National Academy of Sciences*, **109**, 20000-20005.

83. Chemla, Y.R. and Smith, D.E. (2012) In Rao, V. and Rossmann, M. G. (eds.), *Viral Molecular Machines*. Springer, Boston, MA, Vol. 726, pp. 549-584.

84. Ortiz, D., Ordyan, M., Pajak, J., Sippy, J., Catala, A., Oh, C.-S., Vu, A., Arya, G., Smith, D.E. and Catalano, C.E. (2019) Functional Dissection of a Viral DNA Packaging Machine's Walker B Motif. *Journal of molecular biology*, **431**, 4455-4474.

Chapter 3
A-philic DNA Sequence Does Not Significantly Affect Phage φ29 Motor Function

Abstract

Phage φ29 is one of the systems we use to study ATP-powered viral DNA packaging motors. In the studies reported in this chapter, we investigate whether 'A-philic' DNA sequences affect the function of the φ29 motor, with the major motivation being to test the 'B-A scrunchworm' model which predicts that motor velocity with high applied external load force would be increased with A-philic DNA. As reported in Chapter 2 of this thesis, our measurement of bacteriophage T4 packaging A-philic, GC rich DNA sequences showed no significant correlation between DNA sequence and either motor velocities or pauses/slips. Here we reconducted the measurement with phage φ29 to investigate whether this finding is universal for different virus systems. As mentioned in the introduction, the φ29 system exhibits some notable differences from T4, such as the motor containing RNA elements and having much slower DNA translocation rate than T4, so it is of interest to see whether it has the same response to A-philic DNA. Here we report preliminary results which suggest that, like we found for T4, there is no significant change in packaging dynamics. Limitations of these studies are that the uncertainties in the measurements are higher than for the T4 studies because we do not have as large datasets, and since φ29 has a shorter genome these measurements are in a regime where motor velocity is affected by prohead filling which also causes increased uncertainty. Thus, we

caution that the results in this chapter must be considered preliminary and more experiments are needed to fully confirm them.

3.1 Introduction

In many viruses DNA packaging into the container protein shell is a vital step in the process of viral assembly. Viral proheads (~40-100 μm in diameter) are filled with a much longer DNA molecule (~7-70 μm long) packed to near-crystalline density.[1-10] A powerful nanoscale motor drives in and rearranges the DNA molecule from a loose random coil in solution into a tightly spooled packing conformation.[7, 11, 12] The motor transduces ATP energy into mechanical work against large resisting forces arising from entropy loss, DNA bending and repulsion between DNA segments.[1, 7, 10, 13, 14] Single molecule experiments with optical tweezers show evidence that bacteriophages' motors (of phages φ29, T4, and Lambda) can generate > 60 pN force to prevent the DNA ejection.[4, 15] It has become a prevailing topic to investigate the model of phage package to understand fundamental biological mechanisms.

Even though various static structures of motor proteins or their ATPase domains of some bacteriophage, including T4, Sf6, P74-26, D6E, φ29, and herpes simplex virus 1, have been acquired by X-ray and cryo-EM studies, viral packaging dynamics (initiation, translocation and termination) and details of the dynamic motor mechanism have not been fully understood.[16-23] Furthermore, the question of whether there is a universal packaging mechanism for all different phages is unclear. As described in Chapter 2, there are two quite distinct types of models for the packaging motor mechanism. The first, most commonly proposed type, is one in

which translocation of DNA is proposed to be driven by a conformational change in the motor proteins coupled with the ATP hydrolysis cycle. The motor is proposed to grip DNA, undergo a conformational change that moves it by a lever-like action, then release DNA, then undergo a conformational change that resets the protein conformation back to the original one. Such models have been proposed based on the structural data cited above, although these remain largely speculative. As described in more detail in Chapter 2, an intriguing second type of model called the 'scrunchworm' model has been proposed by Harvey in 2015. In this model it is proposed that it is induced conformational change in segment of the DNA that is threaded through the motor channel that drives translocation of the DNA. The specific model proposed by Harvey in 2015 proposed that the motor induces conformational changes in the DNA between the B-form (0.34 nm/bp) and the shorter A-form (0.26 nm/bp), coordinated with DNA gripping and releasing actions, and the ATP hydrolysis cycle, that causes the DNA to be translocated. [18-20, 24-31]

Our studies reported in Chapter 2 of this thesis found no evidence to prove that DNA sequence, and in particular A-philic DNA sequences designed to test the B-A scrunchworm model, correlates to either motor velocity or its pause and slip frequency in the case of the phage T4 motor.[32] These findings argue against the model. However, it is unclear whether this conclusion is universally true for different viral motors. Also, it is notable that the T4 motor differs in various ways, as reviewed above, and also exhibits large fluctuations of up to several hundred bp/s in the motor velocity that the φ29 motor does not exhibit. Here we conduct the viral packaging experiment with the A-philic high-GC content dsDNA construct described in Chapter 2, with phage φ29 — a smaller phage with slower ~150 bp/s average DNA packaging rate. In the case of φ29 it is important to keep in mind that, due to the smaller size of the virus and genome, a much larger fraction of the genome is packaged during our measurements and this results in a

decrease in the packaging rate with increasing length of DNA packaging which must be kept in mind when analyzing the data.

3.2 Methods and Results

3.2.1 DNA Constructs

3.2.1.1 A-philic DNA and Control DNA Template

A-philic DNA and control DNA template are prepared using the same method described in Section 2.2.1.1.

3.2.2 φ29 Materials

3.2.2.1 DNA

1. gp3-DNA was obtained from Dr. Paul Jardine (Univ. of Minnesota); preparation procedure described in ref. [33-35].

2. Lambda Phage DNA (500 μg/ml, NEB Inc.)

3.2.2.2 Nucleotides

1. Biotinylated (Biotin) forward 25.3 kbp primer:

Biotin-5’-CTGATGAGTTCGTGTCCGTACAACTGGCGTAATC-3'

2. Digoxigenin (Dig) reverse 25.3 kbp primer:

Dig-5’-ATCCGATCTGCGTTACCGAATGGATGGATG-3'

3. 10 kbp forward primer:

Biotin-5’-CTGATGAGTTCGTGTCCGTACAACTGGCGTAATC-3’

4. 10 kbp reverse primer:

Dig-5’-CGTGTTCCTGACGGTGTTGCTGAATACAGCGTAT-3’

6. 10 mM dNTP mix

3.2.2.3 Enzymes for DNA Manipulation

1. TaKaRa Long Range LA Taq PCR Kit

2. Thermo Fisher Hot Start Phusion PCR Kit

3.2.2.4 DNA Manipulation Protocol

1. ‘Control’ DNA and A-philic DNA construct is prepared the same way as described in Chapter 2.

3.2.2.5 Buffer Solutions Used with φ29 Components and Complexes

1. 10X TMS buffer: 1M NaCl, 500 mM Tris HCl pH 7.5, 100 mM $MgCl_2$

2. Packaging buffer: 0.5X TMS, 0.5 mM ATP, 0.1 mg/mL BSA

3.2.2.6 Proteins

1. φ29 gp16 motor protein, diluted to ~0.1 μg/μL in 0.5X TMS was obtained from Dr. Paul Jardine (Univ. of Minnesota); preparation procedure is described in Ref. 36.

2. φ29 proheads (with the associated pRNA), diluted to ~1 μg/μL in 0.5X TMS were obtained from Dr. Paul Jardine (Univ. of Minnesota); preparation procedure is described in Ref. 36.

3. Antibodies against φ29 proheads were obtained from Dr. Paul Jardine (Univ. of Minnesota); prepared from Rabbit antisera by a commercial company and purified with a protein A column.

4. Superase In RNAse inhibitor (ThermoFisher Scientific, catalog #AM2696)

For all procedures that follow a P-2 pipettor was used when pipetting liquid in any volume below 2 μL, a P-20 pipettor was used when pipetting liquid in any volume > 2 μL and ≤ 20 μL, a P-200 pipettor was used when pipetting liquid in any volume > 20 μL and ≤ 200 μL, a P-1000 pipettor was used when pipetting liquid in any volume > 200 μL and ≤ 1000 μL.

3.2.3 φ29 DNA Substrate for in vitro Packaging

3.2.3.1 Dual Labeled 25 kbp DNA

1. Follow the procedure described in Section 2.2. Use Lambda phage DNA as the template DNA and use the 5' end labeled biotinlyated 25 kbp forward primer and 5' end labeled digoxygenin 25 kbp reverse primer.

2. Verify the PCR yield and measure the DNA concentration with the UV spectrophotometer.

3.2.4 φ29 in vitro Optical Tweezers Packaging Protocol

The assembly of stalled empty φ29 prohead-motor complexes required purified proheads and purified gp16 protein. After the motor assembled and bound to the prohead, the prohead-motor complex was stalled or stabilized by adding γ-S-ATP. Proheads and gp16 should be thawed, diluted to ~1 μg/μL and ~0.1 μg/μL in 0.5X TMS respectively, and kept on ice before beginning the reaction. Use a siliconized 0.6 mL microcentrifuge tube. The biotinylated 25 kbp DNA needed to be prepared prior to beginning the procedure. For this procedure the RCF of the micro-centrifuge, which was used to spin down the microspheres, was ~2000g.

3.2.4.1 Assembly of Empty φ29 Prohead-motor Complexes

1. Pipet 4 μL of water to a 0.6 ml tube

2. Add 1 μL 10x TMS

3. Add 1 μL BSA (~2.5 mg/mL)

4. Mix the sample by gently flicking the tube

5. Add 2 μL proheads (~1 μg/μL)

6. Add 2 μL gp16 (~0.1 μg/ μL)

7. Mix the sample by gently flicking the tube

8. Incubate sample at room temperature for 2 min

9. Add 2 μL γ-S-ATP (3.5 mM)

10. Mix the sample by gently flicking the tube

11. Place tube on the tube rotator and keep it rotating at room temperature for 60 min

12. Store the complexes at 4 °C. They are typically usable for at most 24 hours.

3.2.4.2 Anti-φ29 Antibody Coated Microspheres

The anti-φ29 antibodies should be thawed and kept on ice before beginning this procedure.

1. Add 50 μL of protein G microspheres to a 0.6 mL tube

2. Pellet the microspheres by spinning them in a microcentrifuge for ~1-2 min

3. Remove the supernatant

4. Wash the microspheres by resuspending them in 50 μL of 1X PBS

5. Repeat steps 2-4

6. Pellet the microspheres by spinning them in a microcentrifuge for ~1-2 min

7. Remove the supernatant

8. Wash the microspheres by resuspending them in 5 μL of 1X PBS

9. Add 1 μL of antibodies (antisera) and mix the sample by gently flicking the tube

10. Place tube on the tube rotator and keep it rotating at room temperature for 45 min

11. Add 45 μL 1X PBS

12. Pellet the microspheres by spinning them in a microcentrifuge for ~1-2 min

13. Remove the supernatant

14. Wash the microspheres by resuspending them in 50 μL of 0.5X TMS

15. Repeat step 12-14

16. Pellet the microspheres by spinning them in a microcentrifuge for ~1-2 min

17. Remove the supernatant

18. Wash the microspheres by resuspending them in 5 μL of 0.5X TMS

19. Store the microspheres 4 °C. Typically they are usable for ~1 week.

3.2.4.3 Binding of Empty φ29 Prohead-Motor Complexes to Antibody Coated Microspheres

1. Add 5 μL of 0.5X TMS to a 0.6 mL microcentrifuge tube

2. Add 2 μL antibody microspheres

3. Mix the sample by gently flicking the tube

4. Add 4 μL complexes

5. Mix the sample by gently flicking the tube

6. Place tube on tube rotator and keep it rotating at room temperature for 60 min.

7. Keep the sample on the rotator after this incubation to avoid the microspheres from settling to the bottom of the tube. Typically, the samples are usable for at most ~6 hours before loss of activity is observed.

3.2.4.4 Complex Microsphere Syringe

This step should be done immediately prior to beginning the optical tweezers experiment.

1. Add 500 μL of 0.5X TMS to a 2 mL tube

2. Add 4 μL of 20 mM γ-S-ATP

3. Add 4 μL complex microspheres

4. Mix the sample by gently flicking the tube

5. Suck solution into 1 mL syringe

3.2.4.5 Binding of Biotinylated DNA to Streptavidin Microspheres

1. Add 10 μL of Streptavidin microspheres (0.5 % w/v) to a 0.6 mL microcentrifuge tube

2. Pellet the microspheres by spinning them in a microcentrifuge for ~1-2 min

3. Remove the supernatant

4. Wash the microspheres by resuspending them in 10 μL of 1X PBS

5. Add 30-50 ng of dual labeled 25 kbp DNA

6. Add 0.2 μL of 100 mg/mL BSA

7. Mix the sample by gently flicking the tube

8. Place tube on tube rotator and keep it rotating at room temperature for 20 min

9. Store in 4 °C. Typically these microspheres are usable for ~4 weeks before loss of activity is observed.

3.2.4.6 DNA Microsphere Syringe

1. Add 500 μL solution of 0.5X TMS to a 2 mL microcentrifuge tube

2. Add 2 μL DNA microspheres

3. Mix the sample by gently flicking the tube

4. Suck solution into 1 mL syringe

3.2.4.7 φ29 Packaging Buffer

This buffer should be prepared fresh each day that experiments are performed and at room temperature. See materials above for the contents of this buffer.

3.2.5 Phi29 Packaging Rate Calculation

3.2.5.1 Pausing and Slipping Criteria

φ29 packaging rate (DNA translocation rate) and motor velocity (rate not including times when pauses or slips occurred) were calculated in the same method as described in 2.2.3. The packaging/pause/slipping classification criteria differed to the T4 study in Chapter 2 because the average packaging rate of phage φ29 was ~4 times slower than T4 and the packaging rate fluctuations were smaller. For these studies, to determine sections of translocation vs. pausing vs. slipping we used velocity (slope of the length vs. time data) threshold criteria. Due to instrument and Brownian noise even data recorded with a non-translocating tether (DNA tethered between two beads with no translocating motor complex) exhibits non-zero fluctuating velocity. This was measured in control experiments and here we use the measured standard deviation in these control velocity measurements for the threshold criteria. Data segments with packaging rate < 27.4 bp/s, i.e, (-1σ, +1σ), were classified as pauses. Segment with slope more than 27.4 bp/s was classified as packaging sections and segments with slope less than -54 bp/s (two standard deviations; since slipping is typically much faster than packaging and thus easier to distinguish) were classified as slips.

3.2.5.2 φ29 Packaging Dynamics

For these preliminary studies we conducted ensembles of measurements of φ29 packaging events using the ~10 kbp, 25 kbp, and 'A-philic' DNA substrates. As mentioned above, since the φ29 genome is only ~19.3 kbp, an appreciable fraction of the genome length is

packaged such that slowing of the motor with increasing length of DNA packaged may occur during the measurements. This occurs because forces resisting packaging increase with increasing length packaged and these forces exert a load on the motor which causes it to slow. For phage φ29, the packaging rate achieved maximum (~150-170 bp/s) when the prohead filling (% of genome length packaged) was 0% and it decreased to ~20 bp/s when the filling approached 100%. An example of a real-time recording of DNA length packaged vs. time for φ29 was shown in Fig. 3.1A. Phage φ29 packaged DNA in a constant packaging rate when its filling < 50%, i.e., translocated DNA < 9.7 kbp, but its packaging rate slowed down when prohead filling is > 50%.

Fig 3.1B plots the average determined rate vs. filling level for a small ensemble of events, in the standard in vitro packaging experimental conditions and under 5 pN applied external stretching force. Records of N=27 packaging events were cropped into 21 slices, corresponding to filling levels ranging from 0% to 100% respectively. Motor velocities in each slice were calculated by the sliding window method described in Chapter 2. The height of bars represented the average motor velocities over all velocities, which the motor velocities were the packaging rate > 27 bp/s, in the same filling levels. And the error bars shown for each data point represent the standard error in the means for each filling slice.

Use of the A-philic DNA Construct

We used the same DNA construct as in the T4 studies discussed in chapter 2, a ~10 kbp plasmid with a ~2 kbp A-philic sequence inserted. The A-philic DNA section was inserted in between 3.8 kbp and 5.8 kbp positions, or 20% to 30% prohead filling with respect to the φ29 genome length. The motor velocities in this interval were treated as the data for studying any

potential effect of the A-philic section on the φ29 packaging dynamics. Note that because φ29 packages at a ~4x slower rate than T4 there was significantly less uncertainty in the absolute position during the measurement than in the T4 measurements described in Chapter 2, since the initial DNA tether length could be used as a reference for the starting length, so it was possible to identify with certainty a larger section of the data corresponding to the A-philic sequence than in the T4 measurements.

Motor Velocities Calculation

It was rare to observe pausing or slipping while measuring φ29 with A-philic plasmid DNA (they were less frequent even that the pretty low pausing or slipping measured with phage T4 packaging). Thus the packaging rate was approximately equal to motor velocity (rate not including pausing and slipping) and could in practice simply be calculated by

$v = \Delta y/\Delta t$, where Δy was the change in displacement of trapped beads and Δt was the change in time. Motor velocities during translocation of the A-philic section, therefore, was obtained by the displacement measured during the A-philic section (20%-30% of filling) divided by duration the of packaging that section. Motor velocities of 'Normal' section were calculated as the average over the 10%-20% and 30%-40% filling sections because the trend of the motor velocity with filling revealed by the most accurate measurements with the 25 kbp construct in Fig. 3.1 indicated it is approximately linearly decreasing when the prohead filling was < 40%, so the average motor velocities of these two sections should equal the middle A-philic section. Note also that this procedure for comparing would also be valid even if the velocity were constant (not decreasing with filling) as it unexpectedly appears to be in the measurements of the A-philic construct and 10 kb control DNA packaging, so this would not introduce any bias in the

comparison. ‘Control’ motor velocities were computed by the same method but for the 20%-30% filling section of the 10 kbp phage DNA sequence. We measured the φ29 motor velocities with both 5 pN applied force (low force) and 30 pN applied force (high force). Error bars report the standard error in the means of the corresponding measurements.

3.3 Results and Discussion

In the preliminary work, we conducted small ensembles of measurements of φ29 packaging events with three different DNA sequences. These are the same DNAs also used in the T4 measurements in Chapter 2: the construct containing the A-philic sequence, the 10 kbp Phage Lambda DNA sequence, and 25 kbp Phage Lambda DNA sequence, under the same experimental conditions. As in the T4 studies the repeated experiments have been done with both low external applied force (5 pN) and high external force (30 pN). Examples of individual events measured with the A-philic construct, with the A-philic and non-A-philic sections of the data marked, are shown in Fig. 3.2. In these plots one does not see any clearly obvious distinct difference in behavior when packaging the A-philic section. Calculated average velocities versus prohead filling level (% of genome length packaged) for all three DNA constructs are shown in Fig. 3.3. An unexpected property of these datasets are that the motor velocities measured with the 10 kbp Phage Lambda DNA and the A-philic DNA were ~20-30 bp/s higher than in the measurement of the 25 kbp samples, and the small, gradual decrease in average velocity with filling observed in the 25 kbp dataset is not clearly seen in the 10 kbp and A-philic datasets. On the other hand, the measurements are mostly consistent to within the experimental errors. We suspect that the apparent slightly higher velocities measured in the 25 kbp data could be due to

these measurements having been done approximately one year earlier than the others and the instrument was re-aligned and re-calibrated before the other measurements were done, which could lead to small systematic errors.

In any case, the data in Fig. 3.3, and summarized in Table 3.1, shows that there is no statistically significant difference in average motor velocity when packaging the A-philic DNA sequence compared to the flanking sections of DNA in the construct or when packaging the other sequences. In a more precise analysis, to account for the expected trend of motor velocity decreasing with increasing filling, we calculated the average motor velocity when packaging the A-philic section (20-30% filling) to the average motor velocity when packaging the "normal" (non-A-philic sequence) flanking sections of the plasmid just before (10-20% filling) and just after (30-40% filling) the A-philic section. These calculations also find no significant differences in average motor velocity (Fig. 3.4A). In this figure the same procedure was also used to analyze the data recorded with the 10 kbp phage DNA and no difference in average motor velocity was detected either, which serves as a negative control experiment.

As in the T4 studies discussed in Chapter 2, as a more precise check that can reduce the influence of variations in packaging rates measured in individual events, we also calculated the ratio of the velocity measured during each event for the A-philic section to the velocity measured in the flanking normal (non-A-philic) sections during the same event. The average ratio for all events was then calculated and found to be close to unity, to within experimental error, again showing that the motor velocity is not significantly affected by the A-philic sequence. This preliminary data thus serves as an additional check of the "B-A scrunchworm" model as discussed in Chapter 2. That model predicts, as shown by the horizontal red lines in Fig. 3.4, that the motor velocity would be slightly higher when packaging A-philic DNA. The fact that no such

change is observed provides additional evidence against this model. We caution again that this φ29 data presented in this chapter should be considered preliminary because not a large number of events were recorded and it would be desirable to repeat these measurements to confirm reproducibility.

Table 3.1 Measured parameters characterizing motor function. Each parameter is an average over all events and uncertainties are reported as standard errors in the means.

DNA Substrate	Motor Velocity (bp/sec)	Motor Velocity Ratio	# of Complexes
A-philic @5pN	164.8 ± 3.5	0.95 ± 0.02	32
Normal @5pN	173.5 ± 3.0		32
Control @5pN	146.7 ± 3.8	0.98 ± 0.03	14
A-philic @30pN	95.0 ± 6.3	0.93 ± 0.06	15
Normal @30pN	101.8 ± 6.1		15
Control @30pN	112.3 ± 11.0	0.92 ± 0.10	16

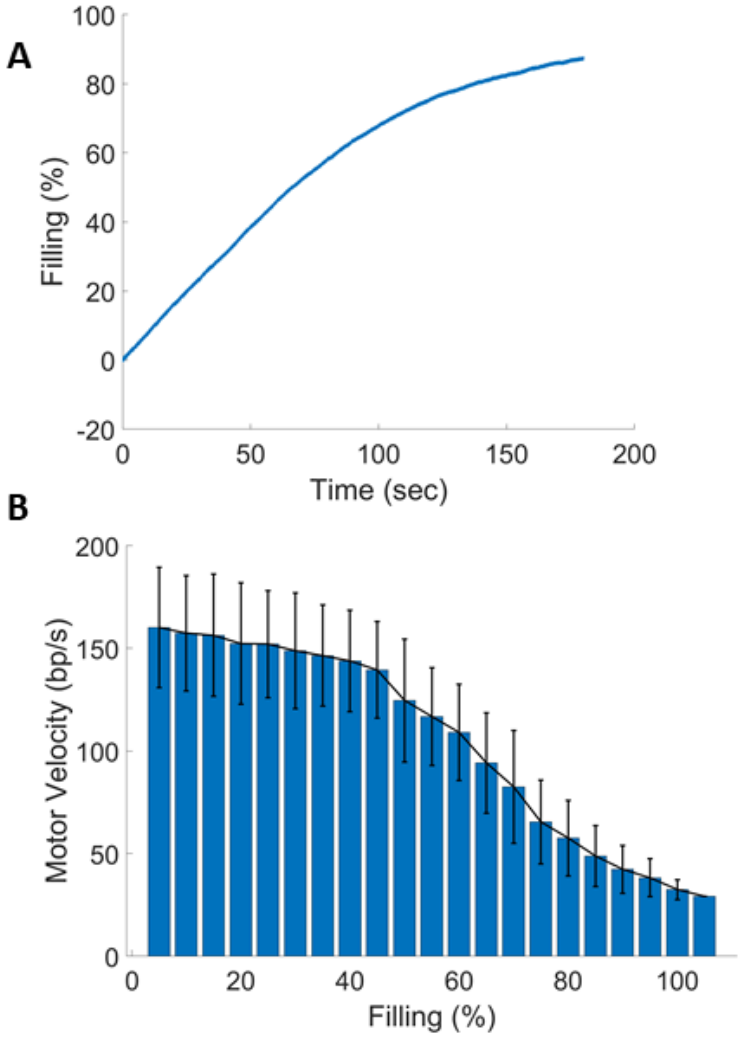

Figure 3.1 (A) Measurement of filling of DNA translocated versus time with the DNA sequence. Phi29 packages dsDNA that the motor velocity slows down while filling level increases. Phi29 Packaging trace is collected within regular Phi29 in vitro experiment as described in Methods. Y-axis is in unit of percentage which refers to ratio of translocated DNA length vs. the length of Phi29 phage genome 19.3 kbp. **(B)** Average Phi29 motor velocities at different filling levels. Averages were determined from N=27 recorded events at 5 pN. Filling refers to translocated DNA length ratio as (**A**). Motor velocities are computed in the same method as in Chapter 2. Error bars indicate standard errors in the means.

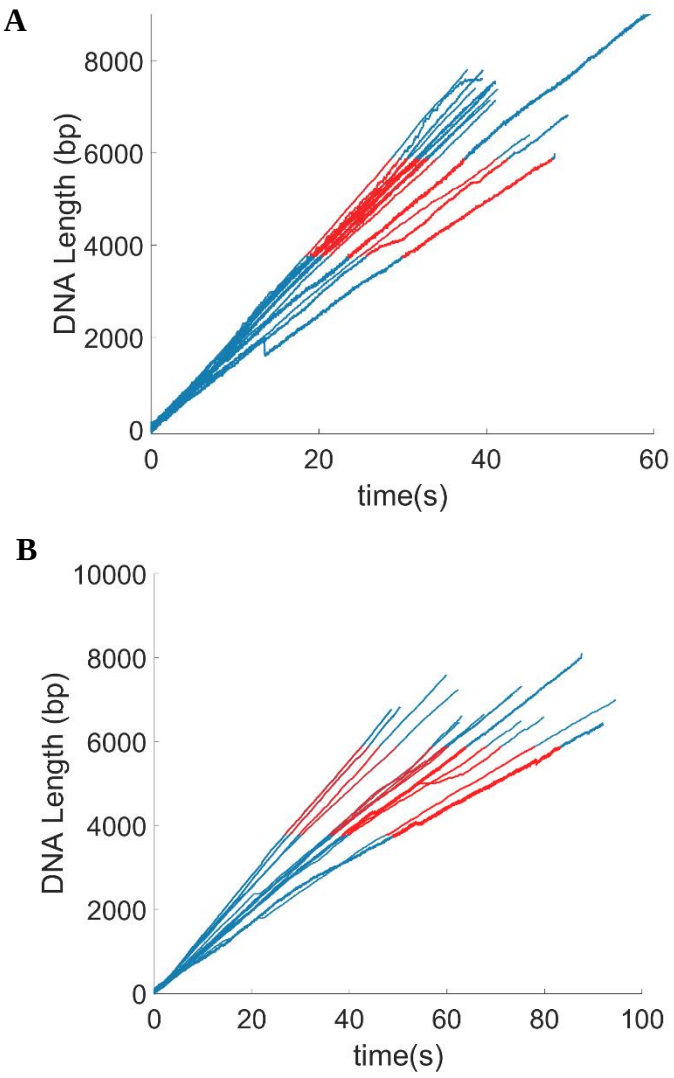

Figure 3.2 Length of DNA packaged versus time. (A) Examples of measurements with the plasmid DNA construct illustrated in Fig. 3.1, with a 5 pN applied force. Sections of packaging containing the normal plasmid sequence (sections I and III) are indicated in blue and sections containing the synthetic A-philic, high % GC sequence (section II) are indicated in red. Each line is a measurement on a single complex and 15 representative events, out of a larger total number of events, are shown. **(B)** Same type of measurements as in (**A**) but with a higher, 30 pN, applied force.

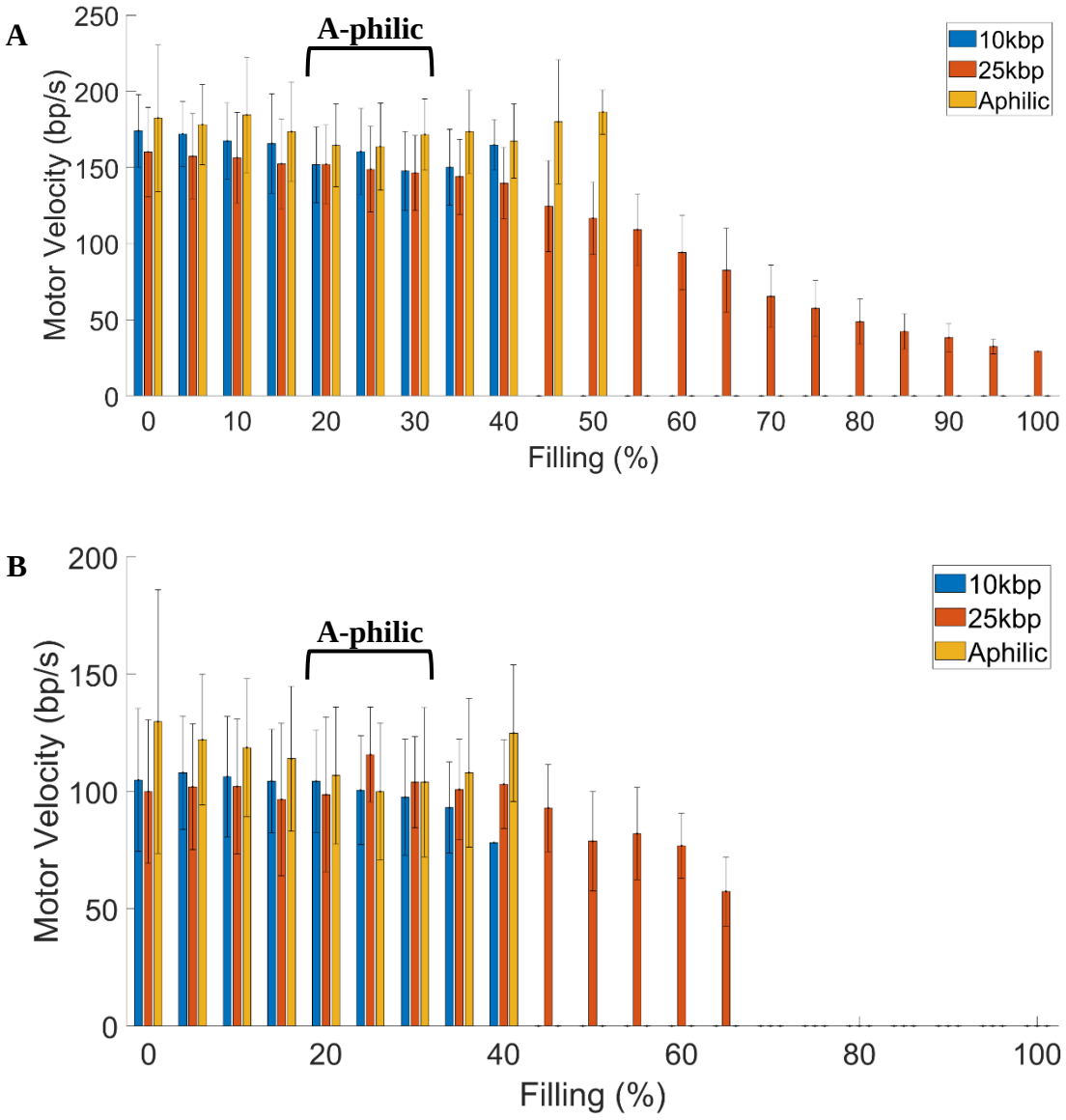

Figure 3.3 Average velocities of Phi29 translocated different DNA sequences. '10kbp' refers to 10 kbp lambda control DNA. '25kbp' refers to 25 kbp lambda DNA. 'Aphilic' refers to the manipulated A-philic DNA described in Chapter 2.

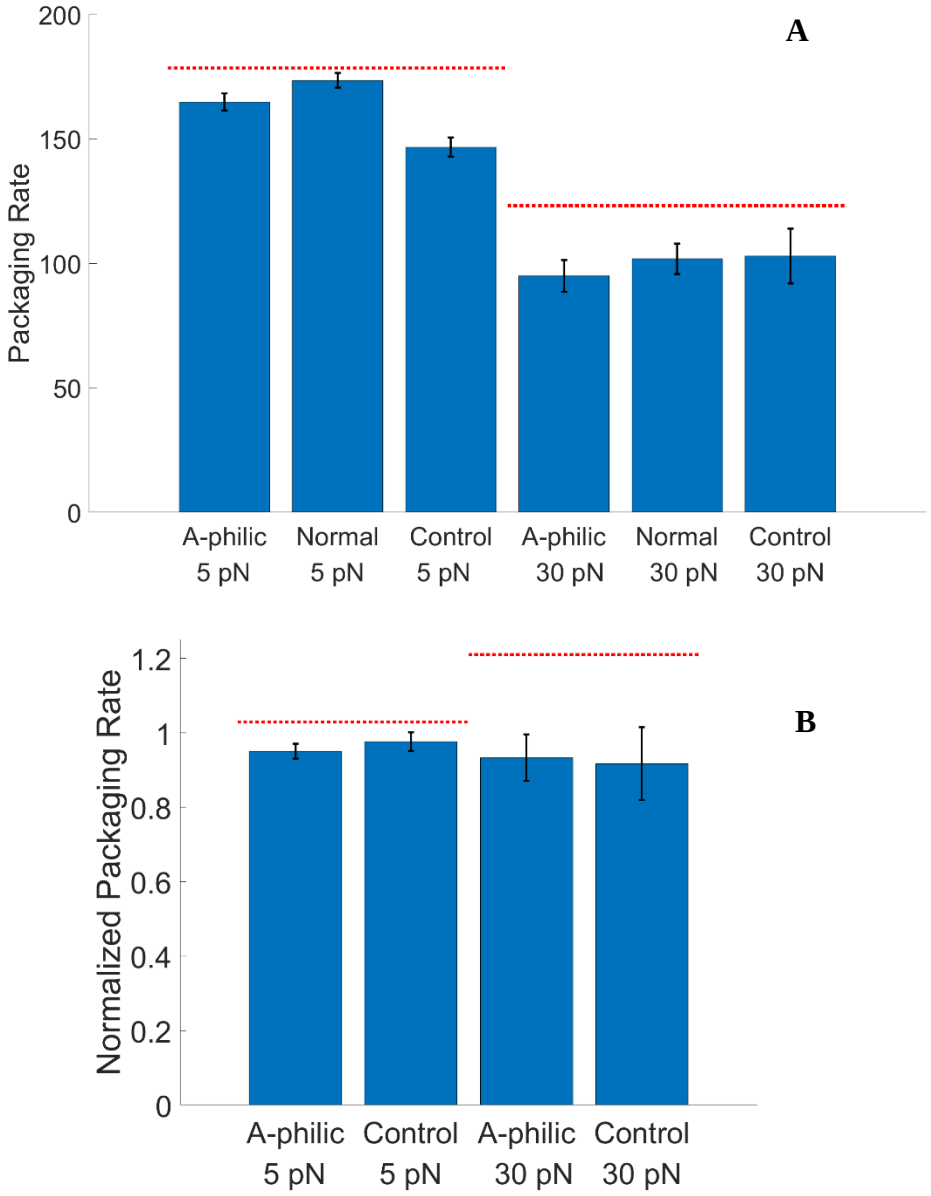

Figure 3.4 Motor velocity ratio analyses. The ratio of the average velocity when packaging the A-philic segment to that when packaging the normal segment was calculated for each packaging event and then the average ratio for all events was calculated (bars labeled "A-philic" with either 5 pN or 30 pN applied loads). Ratios were calculated in the same manner for velocities during packaging of segments at the same positions within the control DNA (bars labeled "Control"). Errors indicated for each bar are standard errors in the means. The dashed grey line indicates a ratio of "1", expected if there is no sequence dependence. The dashed red lines indicate the predicted ratios for the A-philic sequence based on the B/A transition model and measured force-velocity relationship.

References

1. Rao, V.B. and Feiss, M. (2015) Mechanisms of DNA packaging by large double-stranded DNA viruses. *Ann Rev Virology*, 2, 351-378.
2. Casjens, S.R. (2011) The DNA-packaging nanomotor of tailed bacteriophages. *Nature Reviews Microbiology*, 9, 647-657.
3. Feiss, M. and Rao, V.B. (2012) In Rao, V. and Rossmann, M. G. (eds.), *Viral Molecular Machines.* Springer, New York, NY, pp. 489-509.
4. Riemer, S. C., & Bloomfield, V. A. (1978). Packaging of DNA in bacteriophage heads: some considerations on energetics. *Biopolymers: Original Research on Biomolecules*, 17(3), 785-794.
5. Odijk, T. (1998). Hexagonally packed DNA within bacteriophage T7 stabilized by curvature stress. *Biophysical journal*, 75(3), 1223-1227.
6. Kindt, J., Tzlil, S., Ben-Shaul, A., & Gelbart, W. M. (2001). DNA packaging and ejection forces in bacteriophage. *Proceedings of the National Academy of Sciences*, 98(24), 13671-13674.
7. Tzlil, S., Kindt, J. T., Gelbart, W. M., & Ben-Shaul, A. (2003). Forces and pressures in DNA packaging and release from viral capsids. *Biophysical journal*, 84(3), 1616-1627.
8. Purohit, P. K., Kondev, J., & Phillips, R. (2003). Mechanics of DNA packaging in viruses. *Proceedings of the National Academy of Sciences*, 100(6), 3173-3178.
9. Harvey, S. C., Petrov, A. S., Devkota, B., & Boz, M. B. (2009). Viral assembly: a molecular modeling perspective. *Physical Chemistry Chemical Physics*, 11(45), 10553-10564.
10. Smith, D. E. (2011). Single-molecule studies of viral DNA packaging. *Current opinion in virology*, 1(2), 134-141.
11. Arsuaga, J., Tan, R.K.Z., Vazquez, M., Sumners, D.W. and Harvey, S.C. (2002) Investigation of viral DNA packaging using molecular mechanics models. *Biophysical chemistry*, 101, 475-484.
12. Serwer, P. (1986). Arrangement of double-stranded DNA packaged in bacteriophage capsids: An alternative model. *Journal of molecular biology*, 190(3), 509-512.
13. Gelbart, W.M. and Knobler, C.M. (2008) The Physics of Phages. *Physics Today*, 61, 42-47.
14. Speir, J.A. and Johnson, J.E. (2012) Nucleic acid packaging in viruses. *Curr Opin Cell Biol*, 22, 65-71.
15. Fuller, D.N., Raymer, D.M., Kottadiel, V.I., Rao, V.B. and Smith, D.E. (2007) Single phage T4 DNA packaging motors exhibit large force generation, high velocity, and dynamic variability. *Proc Nat Acad Sci USA*, 104, 16868-16873.

16. Sun, S., Kondabagil, K., Draper, B., Alam, T.I., Bowman, V.D., Zhang, Z., Hegde, S., Fokine, A., Rossmann, M.G. and Rao, V.B. (2008) The Structure of the Phage T4 DNA Packaging Motor Suggests a Mechanism Dependent on Electrostatic Forces. *Cell*, 135, 1251-1262.

17. Zhao, H., Christensen, T.E., Kamau, Y.N. and Tang, L. (2013) Structures of the phage Sf6 large terminase provide new insights into DNA translocation and cleavage. *Proc Nat Acad Sci USA*, 110, 8075-8080.

18. Hilbert, B.J., Hayes, J.A., Stone, N.P., Duffy, C.M., Sankaran, B. and Kelch, B.A. (2015) Structure and mechanism of the ATPase that powers viral genome packaging. *Proc Natl Acad Sci USA*, 112, E3792-3799.

19. Xu, R.-G., Jenkins, H.T., Antson, A.A. and Greive, S.J. (2017) Structure of the large terminase from a hyperthermophilic virus reveals a unique mechanism for oligomerization and ATP hydrolysis. *Nucleic Acids Res*, 45, 13029-13042.

20. Mao, H., Saha, M., Reyes-Aldrete, E., Sherman, M.B., Woodson, M., Atz, R., Grimes, S., Jardine, P.J. and Morais, M.C. (2016) Structural and Molecular Basis for Coordination in a Viral DNA Packaging Motor. *Cell Rep*, 14, 2017-2029.

21. Yang, Y., Yang, P., Wang, N., Chen, Z., Su, D., Zhou, Z.H., Rao, Z. and Wang, X. (2020) Architecture of the herpesvirus genome-packaging complex and implications for DNA translocation. *Protein & Cell*, 1-13.

22. Bayfield, O. W., Klimuk, E., Winkler, D. C., Hesketh, E. L., Chechik, M., Cheng, N., Dykeman, E.C., Minakhin, L., Ranson, N.A., Severinov, K., & Antson, A. A. (2019). Cryo-EM structure and in vitro DNA packaging of a thermophilic virus with supersized T= 7 capsids. *Proceedings of the National Academy of Sciences*, 116(9), 3556-3561.

23. Bayfield, O. W., Steven, A. C., & Antson, A. A. (2020). Cryo-EM structure in situ reveals a molecular switch that safeguards virus against genome loss. *Elife*, 9, e55517.

24. Rao, V.B. and Mitchell, M.S. (2001) The N-terminal ATPase site in the large terminase protein gp17 is critically required for DNA packaging in bacteriophage T4. *Journal of Molecular Biology*, 314, 401-411.

25. Mitchell, M.S. and Rao, V.B. (2006) Functional analysis of the bacteriophage T4 DNA-packaging ATPase motor. *J Biol Chem*, 281, 518-527.

26. Tsay, J.M., Sippy, J., DelToro, D., Andrews, B.T., Draper, B., Rao, V., Catalano, C.E., Feiss, M. and Smith, D.E. (2010) Mutations altering a structurally conserved loop-helix-loop region of a viral packaging motor change DNA translocation velocity and processivity. *J Biol Chem*, 285, 24282-24289.

27. Migliori, A.D., Keller, N., Alam, T.I., Mahalingam, M., Rao, V.B., Arya, G. and Smith, D.E. (2014) Evidence for an electrostatic mechanism of force generation by the bacteriophage T4 DNA packaging motor. *Nature Communications*, 5, 4173.

28. delToro, D., Ortiz, D., Ordyan, M., Sippy, J., Oh, C.S., Keller, N., Feiss, M., Catalano, C.E. and Smith, D.E. (2016) Walker-A Motif Acts to Coordinate ATP Hydrolysis with Motor Output in Viral DNA Packaging. *J Mol Biol*, 428, 2709-2729.

29. Harvey, S.C. (2015) The scrunchworm hypothesis: Transitions between A-DNA and B-DNA provide the driving force for genome packaging in double-stranded DNA bacteriophages. *J Struc Biol*, 189, 1-8.

30. Ray, K., Sabanayagam, C.R., Lakowicz, J.R. and Black, L.W. (2010) DNA crunching by a viral packaging motor: Compression of a procapsid-portal stalled Y-DNA substrate. *Virology*, 398, 224-232.

31. Dixit, A.B., Ray, K. and Black, L.W. (2012) Compression of the DNA substrate by a viral packaging motor is supported by removal of intercalating dye during translocation. *Proceedings of the National Academy of Sciences*, 109, 20419-20424.

32. Mo, Y., Keller, N., delToro, D., Ananthaswamy, N., Harvey, S. C., Rao, V. B., & Smith, D. E. (2020). Function of a viral genome packaging motor from bacteriophage T4 is insensitive to DNA sequence. *Nucleic Acids Research*, 48(20), 11602-11614.

33. Grimes, S., & Anderson, D. (1989). In vitro packaging of bacteriophage φ29 DNA restriction fragments and the role of the terminal protein gp3. *Journal of molecular biology*, *209*(1), 91-100.

34. Zhao, W., Morais, M. C., Anderson, D. L., Jardine, P. J., & Grimes, S. (2008). Role of the CCA bulge of prohead RNA of bacteriophage ø29 in DNA packaging. *Journal of molecular biology*, *383*(3), 520-528.

35. Grimes, S., & Anderson, D. (1997). The bacteriophage φ29 packaging proteins supercoil the DNA ends. *Journal of molecular biology*, *266*(5), 901-914.

36. Berndsen, Z. T., Keller, N., Grimes, S., Jardine, P. J., & Smith, D. E. (2014). Nonequilibrium dynamics and ultraslow relaxation of confined DNA during viral packaging. *Proceedings of the National Academy of Sciences*, *111*(23), 8345-8350.

Chapter 4
Determining Trap Compliances, Microsphere Size Variations, and Response Linearities in Single DNA Molecule Elasticity Measurements with Optical Tweezers

Youbin Mo[1], Mounir Fizari[1], Kristina Koharchik[1], Douglas E. Smith[1,*]

[1]Department of Physics, University of California San Diego, La Jolla, CA 92093, USA

*** Correspondence:**
Douglas E. Smith
des@ucsd.edu

Keywords: Optical Trap, Laser Tweezers, Single-Molecule, DNA Elasticity, Calibration, Microsphere Size, Force, Trap Stiffness

Abstract

We previously introduced the use of DNA molecules as metrology standards for biophysical force and displacement measurements with optical tweezers. Force and length scale factors can be determined from measurements of DNA stretching. Trap compliance can be determined by fitting the data to a nonlinear DNA elasticity model, however, noise/drift/offsets in the measurement can affect the reliability of this determination. Here we demonstrate a more robust method that uses a linear approximation for DNA elasticity applied to high force range (25-45 pN) data. We show that this method can be used to assess how small variations in microsphere sizes affect DNA length measurements and demonstrate methods for correcting for these errors. We further show that these measurements can be used to check assumed linearities

of system responses. Finally, we demonstrate methods using microsphere imaging and DNA stretching to check the compliance and positioning of individual traps.

4.1 Introduction

Optical tweezers have many applications in biophysics [1-8], with one powerful technique being single biomolecule manipulation [9-18]. Here we describe several useful methods for system calibration/characterization based on single DNA manipulation.

In our instrument two laser beams create two traps and the position of one is adjusted [19-22]. The force acting on a trapped microsphere is determined by measuring laser deflection [23]. Single DNA molecules are tethered between two microspheres and we apply stretching force [22]. In many biophysical studies one wants to measure force and changes in the molecular length or extension due to translocation by a molecular motor or interactions with other biomolecules/ligands that induce conformational changes. In our system, several parameters need to be determined: force and trap displacement scale factors, relative trap positions, and trap compliances [24]. Methods for calibrating displacement include microsphere tracking via calibrated imaging systems and displacement with calibrated positioning stages, while methods for calibrating force and compliance include analyses of Brownian fluctuations, applied fluid drag forces, and trapping beam momentum changes [3, 23, 25-33]. We introduced an alternative approach using DNA molecules as metrology standards [22, 24]. This is not intended to be a more accurate method than others, but rather a complementary one that has several useful attributes: (1) all calibration factors can be determined simultaneously via a single type of measurement; (2) it is relatively easy to implement especially if one is already working with

DNA; (3) it does not require a calibrated imaging system, characterization of the optical system, precise control of sample stage position, or application of fluid drag forces; (4) it is an independent calibration method that can be used to check other methods; and (5) calibration extends to high forces (45 pN) and the linearity of system responses can be assessed. Advantages of DNA as a metrology standard are that its elasticity is well characterized [9, 34, 35], its length can be precisely controlled in increments of 1 base-pair (0.34 nm), and particular DNAs can be exactly replicated in any lab.

We determine force scale factor based on the DNA overstretch transition that occurs at ~64 pN in the conditions used [22, 35, 36]. Displacement scale is determined by measuring two DNA molecules having different lengths [22, 24]. Series compliance of the traps and relative trap positions can be determined by fitting of a nonlinear DNA elastic force law model to measurements, however this may not be reliable due to measurement noise and offsets, particularly in the low-force regime [22, 24]. We describe here a more reliable method using high-force range data (25-45 pN) where a linearized DNA elasticity model is accurate. This method can be used to check assumed linearities of system responses and errors in DNA length measurements caused by variations in the microsphere sizes. We describe methods to correct these errors. We also describe how combined microsphere imaging can be used to check positioning and compliances of individual traps.

The concept of using DNA measurements for calibration could also be applied, with minor modifications, to other types of optical tweezers setups, magnetic tweezers, and AFM/microneedle instruments that use force-cantilevers [37], in any case where single DNA stretching can be measured. In single optical trap and cantilever systems, DNA can be attached at one end to the sample chamber surface and stretched via a piezo-actuated stage [37].

4.2 Methods, Results, and Discussion

4.2.1 Linear Approximation for DNA Elasticity in High Force Range

The elasticity of DNA is well described by the extensible worm-like chain (WLC) model that predicts:

$$x/L = 1 - \sqrt{\frac{kT}{4FP}} + \frac{F}{S} \qquad (1)$$

where F is the stretching force, x is the DNA end-to-end extension, L is the unstretched DNA contour length (0.34 nm per basepair), k the Boltzmann constant, T the temperature (kT=4.14 pN·nm at room temperature), S the DNA stretch modulus, and P the DNA persistence length [34, 38, 39]. In the conditions we use, 10 mM Tris-HCl, pH 7.5, 150 mM NaCl, S=1275 pN and P=45 nm according to published studies by Wenner et al. that consider ionic dependence [9, 35, 40].

A difficulty is that parameter determination via nonlinear fitting of data to this model is sensitive to experimental noise/drift/offsets, particularly at low force [22]. However, a linear approximation of the square-root term in Eq. (1) is valid in a restricted high-force range. We use a maximum of 45 pN to stay well below the onset of non-linear behavior caused by the DNA overstretch transition [35] and reduce the probability DNA detachment [41]. From 25-45 pN the square-root term can be approximated with the function y=(3.78E-4)F + 0.0391 (Fig. 4.1A; F in pN). The average error is ~0.3% and maximum error ~2% (Fig. 4.1B). This results in a linearized Eq (1):

$$\frac{x}{L} = A + BF \quad (2)$$

Where A=0.961 and B=1.16E-3 for the conditions we use.

4.2.2 Method to Determine Trap Compliances by DNA Stretching

A single DNA molecule is stretched by increasing the separation, d, between the two traps as illustrated in Fig. 4.1C. This is done by steering one of the beams using a mirror tilted by a piezoelectric actuator. Further details and a schematic diagram of the system are given in the Supplementary Materials [21]. The voltage applied to control the mirror actuator is referred to as V_{mirror} and the value of V_{mirror} when the two traps overlap is referred to as $V_{overlap}$. The system is intended to have a linear response so that $d = \beta(V_{mirror} - V_{overlap})$. Once the calibration parameters β and $V_{overlap}$ are known the separation of the two traps in nanometers can be determined. We showed previously that the displacement scale factor β can be accurately determined by measuring two DNA molecules of known lengths (see Supplementary Materials and Ref. [22]). For our system β=980 nm/volt. Here we will describe methods for determining $V_{overlap}$ and checking the assumed linearity of this relationship.

Force $F=\alpha V_{PSD}$ is determined by measuring laser deflections with a position-sensing detector (PSD), where V_{PSD} is the detector signal and α the force scale factor. We showed previously that α can be accurately determined by measurements of the DNA overstretch transition (see Supplementary Materials and Refs. [22, 42]). For our system α=38.3 pN/volt. Below we will discuss a method for checking the linearity of this relationship. Note that in a

dual-trap system F can be measured with either trap. DNA is stretched under tension between the two microspheres, so the magnitude of the force acting on each is the same.

In the Hookean regime a trapped microsphere subject to force *F* is displaced from its equilibrium position by $\Delta x = \gamma' F$, where γ' is the trap compliance [5]. The two traps may have different compliances $\gamma 1$ and $\gamma 2$, but determination of the series compliance $\gamma = \gamma 1 + \gamma 2$ is usually the only parameter needed for our applications. The sum of the displacements of the two microspheres when a force is applied is $\Delta x = \Delta x_1 + \Delta x_2 = \gamma 1 F + \gamma 2 F = \gamma F$. We will discuss a method for checking the assumed linearity of this relationship.

When a single DNA molecule is stretched between the two trapped microspheres, as illustrated in Fig. 4.1C, the separation between the centers of the two traps is

$$d = x + (\Delta x_1 + \Delta x_2) + (r_1 + r_2) \qquad (3)$$

where x is the end-to-end extension of the DNA, Δx_1 and Δx_2 are the displacements of the microspheres from the trap centers, and r_1 and r_2 are the radii of the microspheres.

When d = $\beta(V_{mirror} - V_{overlap})$ and $\frac{x}{L} = A + BF$ are substituted into Eq (3) we obtain

$$\beta(V_{mirror} - V_{overlap}) = (LB + \gamma) F + (LA + r_1 + r_2) \qquad (4)$$

This equation has a linear form in which βV_{mirror} is experimentally controlled and F is measured. Thus, linear fits can be used to determine the slope η and thereby the series compliance $\gamma = \eta - LB$, since *L* and *B* are known.

4.2.3 Experimental Results for Trap Compliance

We prepared 10.7 kilobasepair (kbp) DNA molecules with one end biotin labeled for attachment to streptavidin-coated polystyrene microspheres (~2.1 μm diameter; Spherotech) and the other end labeled with digoxygenin for attachment to anti-digoxygenin coated microspheres (~2.3 μm diameter; Spherotech); these are commonly-used, non-covalent attachments and details are given in the Supplementary Materials and prior publications [21, 22, 24, 41]. We recorded N=99 stretching measurements with F=25 to 45 pN. V_{mirror} was varied and F was measured. Examples shown in Fig. 4.2A confirm the dependence is linear as expected. Each dataset was fit to Eq. (4) to determine series compliance $\gamma=\eta-LB$, yielding an average value γ=12.8 nm/pN (standard deviation=0.56 nm/pN). Additional measurements were done with a 25.3 kbp DNA and yielded a consistent value γ=12.5 nm/pN (standard deviation=1.8, N=180).

Sources of error in determining γ include (details are given in Supplementary Materials): (1) 0.2% uncertainty due to uncertainty in the value of P reported by Wenner et al. [35]; (2) 3.8% due to uncertainty in S [35]; (3) 1.5% due to uncertainty in force calibration (α) [22, 35]; (4) 0.95% due to uncertainty in trap position calibration (β); (5) 0.96% error due to the linearized DNA elasticity approximation (Eq. 2); (6) 0.3% due to noise in the force measurement. Here (1-5) are systematic errors. They are not independent, but together contribute a maximum uncertainty of ~6%. Although (6) considers a source of random (measurement) error, the measured standard deviation reported above is higher (e.g., 4.3% for the 10.7 kbp DNA). While this is acceptably small uncertainty for our applications it suggests there are additional random error sources. Our intuition is these include effects of factors such as the variations in

microsphere size and variations in DNA elasticity due to, for example, occasional nicks (single-stranded breaks) occurring in individual molecules.

4.2.4 Characterization of the Effect of Microsphere Size Variations

The y-intercept of Equation (4) is $\epsilon = LA + r_1 + r_2$. L and A are constants, but the microsphere radii r_1 and r_2 vary by small amounts. This is undesirable since it causes errors in DNA length measurements. The ϵ values determined by the linear fits can be used to characterize these variations. A histogram of ϵ values, with the mean subtracted, is shown in Fig. 4.2B. The standard deviation is 144 nm. A similar standard deviation of 138 nm was determined from the measurements with the 25.3 kbp DNA. These values are consistent with the standard deviations in the microsphere radii of ~140 nm reported by the manufacturer. Below we will discuss methods for correcting for these errors.

A related question is whether the differences in microsphere size cause detectible differences in trap compliances. In regime we operate, where microsphere diameter is larger than the trapping laser focal spot, a larger microsphere could cause larger trap compliance [43]. To investigate if this effect is significant, we checked to see if there was any correlation between ϵ and γ values in the ensemble of individual measurements. We did not find a significant correlation (correlation coefficient = -0.2). This is likely because the standard deviation in the microsphere size is only ~7%, so any effect is too small to measure.

4.2.5 Determination of an Average Displacement Offset Factor

A simplified method can be used when neglecting the small variations in microsphere sizes is acceptable. Instead of considering $V_{overlap}$, we define $V_{contact}$ to be the value of the mirror control voltage V_{mirror} when the two microspheres come into contact. In principle this could be detected by measuring the force signal when the two microspheres touch, but we find this can be inaccurate due to optical cross-talk/interference between the beams and offsets in trap positions [22]. Instead, one can determine this parameter from the DNA measurements. The value $\beta V_{contact} = \beta V_{overlap} + (r_1 + r_2)$, so Eq (4) an be rewritten as

$$\beta V_{mirror} = (LB + \gamma)\,\mathrm{F} + (LA + \beta V_{contact}) \qquad (5)$$

Measurements of $\epsilon = LA + \beta V_{contact}$ allow one to determine values of $V_{contact}$, since L, A, and β are known. Individual measurements depend on the individual microsphere sizes but from an ensemble of measurements an average value $\bar{V}_{contact}$ is determined. The imposed DNA extension is then given by

$$x = \beta(V_{mirror} - \bar{V}_{contact}) - \gamma\mathrm{F}, \qquad (6)$$

which can be controlled since all the variables are known or measured.

4.2.6 Correction for Microsphere Size Variations

Variation in the trap separation determination caused by variations in microsphere sizes causes error in DNA length measurements. Above we determined a standard deviation of 144 nm in $r_1 + r_2$. Here we describe methods to correct for this error. While in many types of

biophysical studies only relative changes in DNA length may be of interest, in some cases it is desirable to determine absolute DNA length [44].

As an example, we discuss studies of motor-driven viral DNA packaging. We attach a viral procapsid-motor complex to one microsphere and a DNA molecule is translocated into the procapsid by the motor [44, 45]. We attach the other end of the DNA to a second trapped microsphere, such that the motor pulls the two microspheres together as the tethered DNA is translocated. The length of DNA packaged into the procapsid is equal to the full DNA substrate minus the unpackaged DNA between the two microspheres. Since the operation of the motor and rate of DNA translocation is affected by the length of DNA packaged in the procapsid [46], we want to measure the absolute DNA length.

One of the viruses we study, bacteriophage phi29, has a 19.3 kbp genome (~6600 nm), so uncertainty of 144 nm would cause uncertainty of ±2% in the determination of the fraction of the genome packaged. However, since the motor has a relatively slow translocation rate (maximum of ~180 bp/s in the conditions we use), a simple method to reduce this error is to use the measured starting tether length as a reference. Packaging is initiated by moving the microsphere carrying DNA near a second trapped microsphere carrying procapsid-motor complexes. The time delay between initiation of packaging and start of data recording is ~0.2 to 0.8 s. With a packaging rate of ~180 bp/s, ~36-144 bp is packaged during this time, corresponding to ~12 to 50 nm, yielding an uncertainty of ~38 nm. Since this is less than that of ~144 nm caused by microsphere size variations, use of the measured starting DNA length as a reference in each measurement should be useful for improving accuracy.

To demonstrate this, we analyzed a dataset of N=60 phage phi29 packaging events, where the DNA translocation rate was ~180 bp/s, and found a standard deviation in measured

starting lengths of ~175 nm. This is roughly consistent with the expected uncertainty of ~144 nm due to microsphere size variations plus ~38 nm due to variations in initial length of DNA packaged, which implies that subtracting the initial starting length would reduce the uncertainty in measurement of absolute length of DNA from ~175 nm to ~38 nm. A limitation of this technique is that it would not be beneficial if the DNA translocation rate was so fast that the error caused by the uncertainty in the time delay between initiation and data recording was larger than the error caused by the variation in microsphere sizes.

A second method is based on defining a minimum separation where the two trapped microspheres nearly touch as a reference. This method can be used in cases where DNA translocation proceeds for long enough to bring the microspheres into near contact, or if they can be moved together after the measurement. To test this, we conducted measurements in which we brought two microspheres together into near contact. They are observed using a video imaging system described in the next section. Each microsphere appears as a bright spot surrounded by a dark circular ring. We defined the minimum separation reference as the point where the two dark rings were first observed to overlap as the separation was decreased. We then recorded the value of the control voltage, V'_{mirror}, and repeated this measurement with N=32 different pairs of microspheres. The standard deviation in the inferred relative positions was 136 nm, which is close to the estimated uncertainty of 144 nm in $r_1 + r_2$ determined from the DNA stretching measurements discussed above. This indicates that if $\beta V'_{mirror}$ values is recorded for each pair of microspheres these can be used to correct length measurements to reduce error caused by microsphere size variations.

4.2.7 Checking System Response Linearities

Equation (4) assumes several instrument response relationships are linear. We expect these to be valid based on the system design, but describe here how they can be checked by analysis of the DNA stretching data.

Trap separation is assumed to obey d = $\beta(V_{mirror} - V_{overlap})$, where V_{mirror} is the control voltage. This is expected since the trapping beam is steered by a feedback-controlled piezo-actuated mirror. However, the DNA stretching measurements provide a check. Suppose there was a nonlinear response in which the actual relationship deviated from the assumed one by a quadratic term, so that $d_{actual} = d_{assumed} + \delta_1(\Delta d)^2$, where δ_1 is a constant and Δd is the separation change. The actual DNA extension would be greater than assumed based on the linear relationship, but this would cause the force at each value of the assumed extension to be higher than predicted by the DNA force-extension relationship. The F vs. d plot is predicted to be linear over 25-45 pN but the error term would cause curvature. That our data does not show significant curvature (Fig. 4.2A) suggests there is no significant error of this type. To quantify the effect of such an error, we subtract the error term $\delta_1(\Delta d)^2$ from the plotted extension values and keep the measured force values unchanged. As an example, assume the error increases from zero at *F*=25 pN to 15% of Δd at *F*=45 pN. This results in simulated data plot where curvature is resolvable within the experimental noise (Fig. 4.3A). That curvature of this magnitude is not observed in the actual data implies that an error of this magnitude does not occur. A limitation is that this method only probes a narrow range of trap separations. Below we discuss a method using microsphere imaging to test a much wider range of trap separations.

Force measurement is assumed to obey $F = \alpha V_{PSD}$, where V_{PSD} is the detector signal and α the force scale factor. This is expected based on the system design, but it is conceivable that a nonlinear error could occur, for example due to optical misalignments. We perform a similar analysis as above to consider an error $F_{actual} = F_{assumed} + \delta_2(\Delta F)^2$, where δ_2 is a constant and ΔF is the change in force. As shown in Fig. 4.3A we again find that if this error increased from zero at 25 pN to 15% of ΔF at 45 pN this would cause detectible curvature in the F vs. d plot, but this is not observed in our recorded data.

Displacement of the microsphere from the trap is assumed to be Hookean, $\Delta x = \gamma F$, where γ is the compliance, but there could be deviations. Some studies find that beyond a low-force Hookean regime compliance decreases slightly, before ultimately increasing again at the highest forces when the bead begins to escape the trap [29, 47]. If this occurred and was neglected, it would cause the measured forces in the force vs. DNA extension ($x = d - \Delta x$) plot to be higher than predicted. We find that a quadratic error term of 15% maximum magnitude in Δx would cause detectible curvature, but curvature of this magnitude is not observed (Fig. 4.3B).

4.2.8 Checking Trap Positioning Linearity Over a Wide Range

The DNA force-extension measurements described above provided a test of the validity of the trap separation control linearity $d = \beta(V_{mirror} - V_{overlap})$, but only over a limited range of separations. To check that the relationship is valid over the full range we used video imaging and tracking of the microsphere centroid.

The field of view was illuminated by a ~10 mW blue LED imaged into the back focal plane of the downstream microscope objective. The image formed by the upstream microscope objective was recorded by video camera (WAT-902H2, Watec, Inc.) and digitized by a video capture card (NI-PCI-1405, National Instruments, Inc.). The centroid of the microsphere was tracked by locating the dark circular ring in the microsphere image with the 'imfindcircles' function in the Image Processing Toolbox in Matlab, which employs a circular Hough transform algorithm. In this analysis, pixels in the video image were converted to nm using on the known value of β. In this manner we could confirm that trap position is linearly proportional to mirror-tilt control voltage over the full range of ~13 μm (Fig. 4.4A).

4.2.9 Determination of Individual Trap Compliances

Fitting DNA stretching data to Equation (4) described above provides a convenient way to determine the series compliance of the traps, which is sufficient for most of our studies since one can determine the change in DNA extension. However, some studies may require knowledge of individual trap compliances [20]. This can be done by combining microsphere imaging with DNA stretching to apply controlled forces. It can be challenging to determine the compliance of a movable trap because a small displacement due to force needs to be discerned from a much larger imposed movement of the trap. The method described below provides a solution to this issue.

When high forces are applied to stretch the DNA after some time the molecule detaches due to force-induced dissociation of the digoxygenin-anti-digoxygenin linkage [41]. An example is shown in Fig. 4.4B. The DNA was stretched by increasing the trap separation in small steps

but it suddenly detached at F~40 pN, causing the force to drop to zero. Such events can be used to determine the trap compliance because the separation between the traps remains constant while the microsphere suddenly moves a distance $\Delta x_1 = \gamma_1 \Delta F$, where γ_1 is the trap compliance and ΔF is the force drop. Because detachments occur randomly at different forces ranging from ~5 to 50 pN, the predicted relationship $\Delta x_1 = \gamma_1 \Delta F$ can be tested over a wide force range to confirm a Hookean response and determine γ_1. Examples of such measurements are shown in Fig. 4.4C and are consistent with an assumed linear $\Delta x_1 = \gamma_1 \Delta F$ relationship. The value of γ_1 we obtain by a linear fit to these data is 10.9 ± 1.1 nm/pN. The significant scatter in the data points is attributable to the low resolution of our imaging system, which was originally only designed to allow the user to check for the presence of a microsphere. Since methods have been developed to measure microsphere movements with nanometer-level resolution [48], the method has potential to be improved significantly.

4.3 Conclusions

We demonstrated methods by which trap compliance can be determined in a robust matter via measurements of DNA stretching in the regime where a linear elasticity approximation is valid. This method is especially useful if one is already working with DNA, but more generally provides an independent method to confirm other calibration methods in the literature [3, 23, 25-33]. In comparison to many other methods, it does not require a calibrated imaging/optical system, fluid flow, or precise sample stage control, and four calibration parameters (force scale, trap displacement scale, relative trap position, and compliance) can be simultaneously determined. Calibration extends to high forces (45 pN) and the linearity of

system responses can be tested. The method is also useful for characterizing and correcting for the effect of variations in microsphere sizes on extension measurements. Finally, we show that combined microsphere imaging can be further used to check individual trap positions and compliances.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

YM and MF developed methods, conducted measurements, analyzed data, and edited the text. KK conducted measurements and developed methods. DES supervised the project, developed methods, analyzed data, and wrote the text.

Acknowledgements

We thank Damian delToro and Nick Keller for optical tweezer design and construction and protocols, Paul Jardine for supplying phage phi29 components, and Brandon Rawson for discussions.

This research was supported by NSF grant MCB-1716219 and NIH/NIGMS grant R01GM118817 to DES, and by an MBTG Fellowship award to MF (NIH Grant T32 GM008326).

Chapter 4, in full, is a reprint of the material as it appears in the journal Frontiers in Molecular Biosciences. Mo, Y., Fizari, M., Koharchik, K., & Smith, D. E. (2021). Determining Trap Compliances, Microsphere Size Variations, and Response Linearities in Single DNA Molecule Elasticity Measurements with Optical Tweezers. *Frontiers in Molecular Biosciences*, 8, 93. The dissertation author was the primary investigator of this paper.

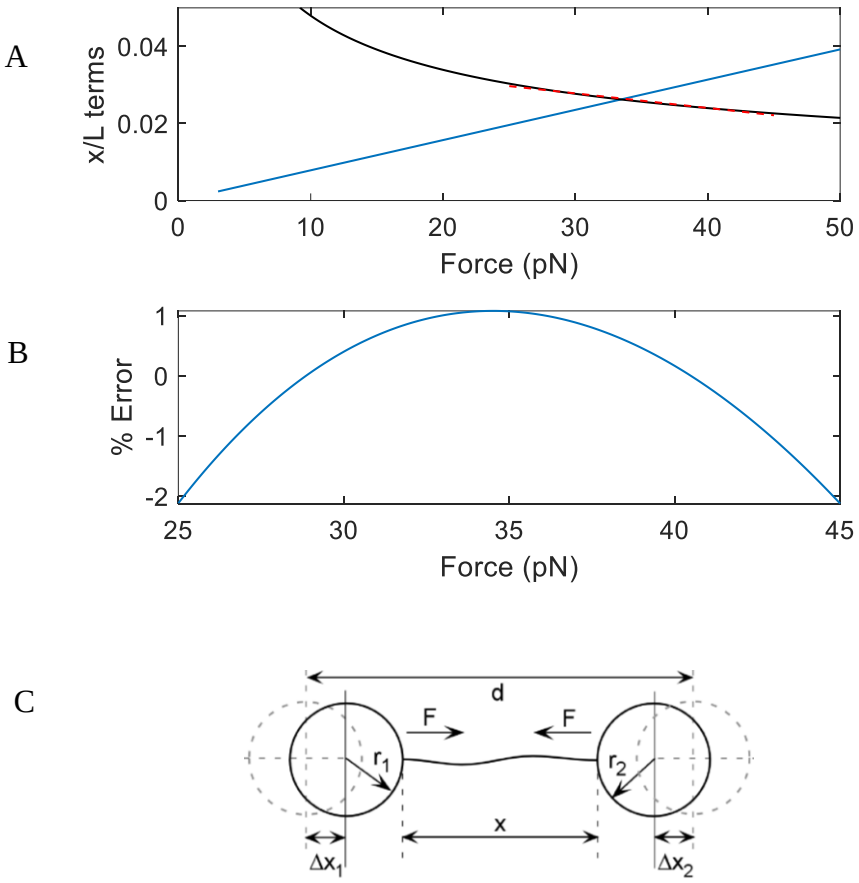

Figure 4.1 **(A)** Plots of the magnitudes of the two force-dependent terms in Equation (1); the linear term in blue. The square-root term (black) can be accurately approximated by a line over the range from F=25 to 45 pN (red dashed line). **(B)** % error made in the square-root term by the linear approximation. **(C)** Schematic illustration of the variables involved in force-extension measurements of DNA stretched between two optically trapped microspheres. The distance between the trap centers is *d*, the end-to-end extension of the DNA is *x*, the radii of the microspheres are r_1 and r_2, the force exerted by the tensioned DNA on the microspheres is *F*, and the displacements of the microspheres from the trap centers are Δx_1 and Δx_2.

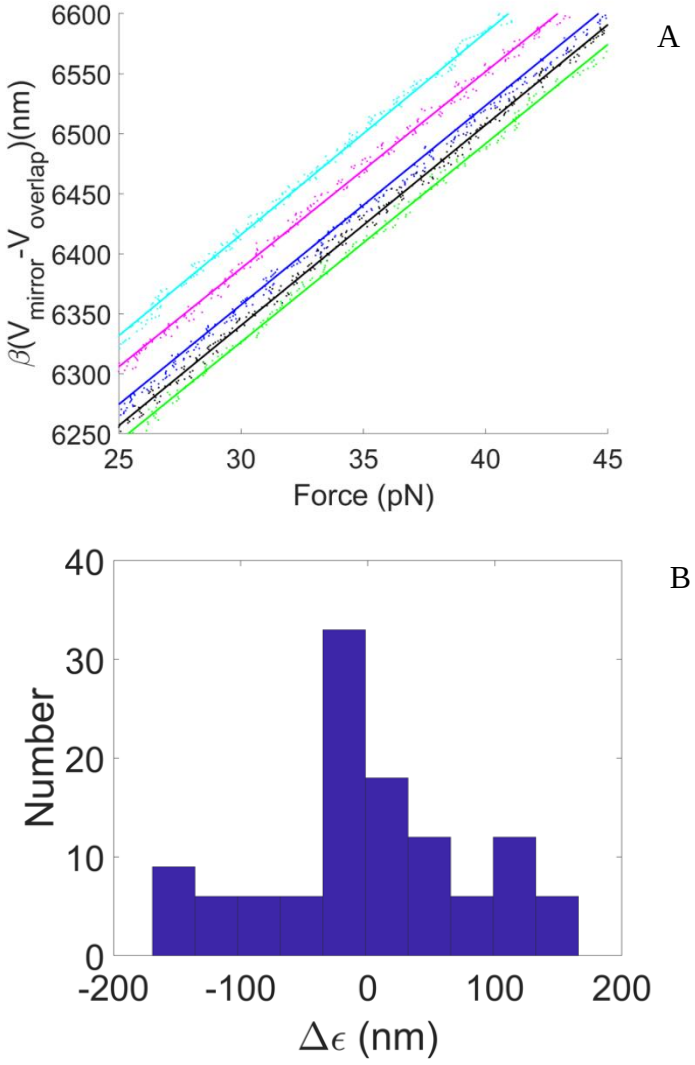

Figure 4.2 (A) Examples of plots of $\boldsymbol{\beta}(\boldsymbol{V_{mirror} - V_{overlap}})$ (proportional to trap separation) vs. F for data recorded when stretching the DNA molecules between the two optically trapped microspheres. The points are experimental measurements and the lines are fits to Equation (4), used to determine trap compliances, microsphere size variations, and average displacement offset factor. **(B)** Histogram of variations in $\boldsymbol{\epsilon}$ values determined by the linear fits of the DNA stretching data to Equation (4), which characterizes the effect of variations in microsphere sizes.

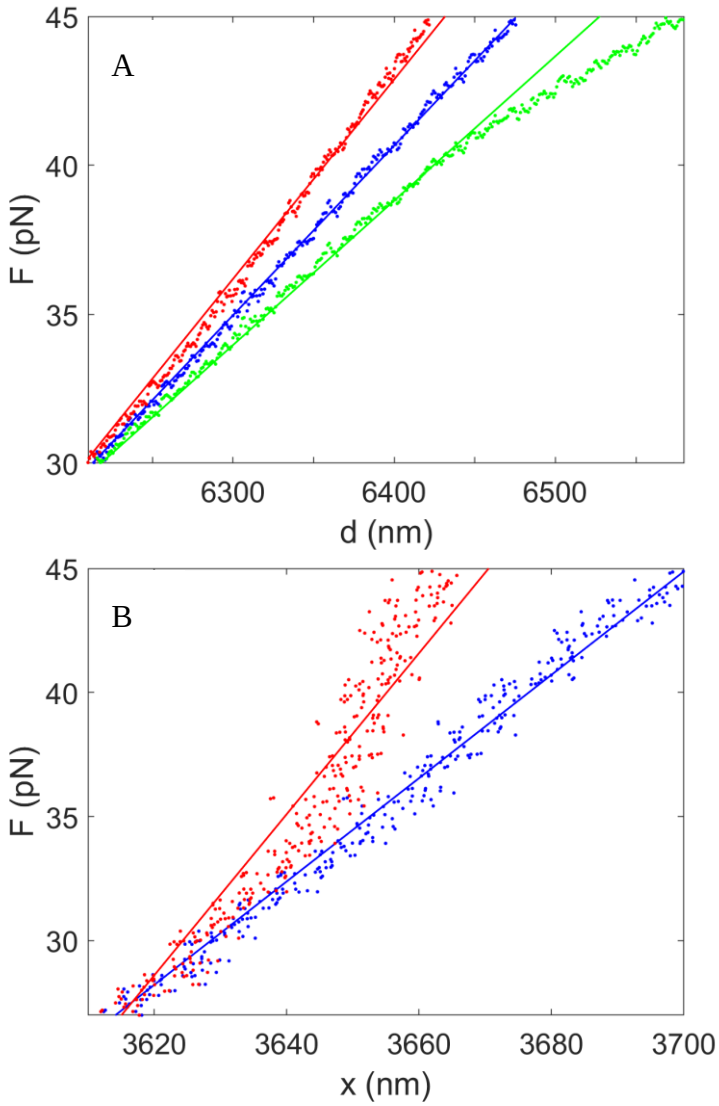

Figure 4.3 Analyses of the predicted effects of nonlinear errors in system responses. (A) The blue points are recorded force vs. separation data and the blue line shows a linear fit, which describes the data well. The red points predict the effect of a quadratic error term affecting the separation control, which causes detectible curvature in the plot. The red line is a linear fit to these points. Similarly, the green points predict the effect of a quadratic error term affecting the force determination, which again causes detectible curvature. The green line is a linear fit to these points. **(B)** The blue points are recorded force vs. DNA extension measurements and the blue line shows a linear fit. The red points predict the effect of a quadratic error in the displacement of the microspheres from the trap centers ($\Delta \boldsymbol{x}$), which causes detectible curvature, and the red line is a linear fit to these points.

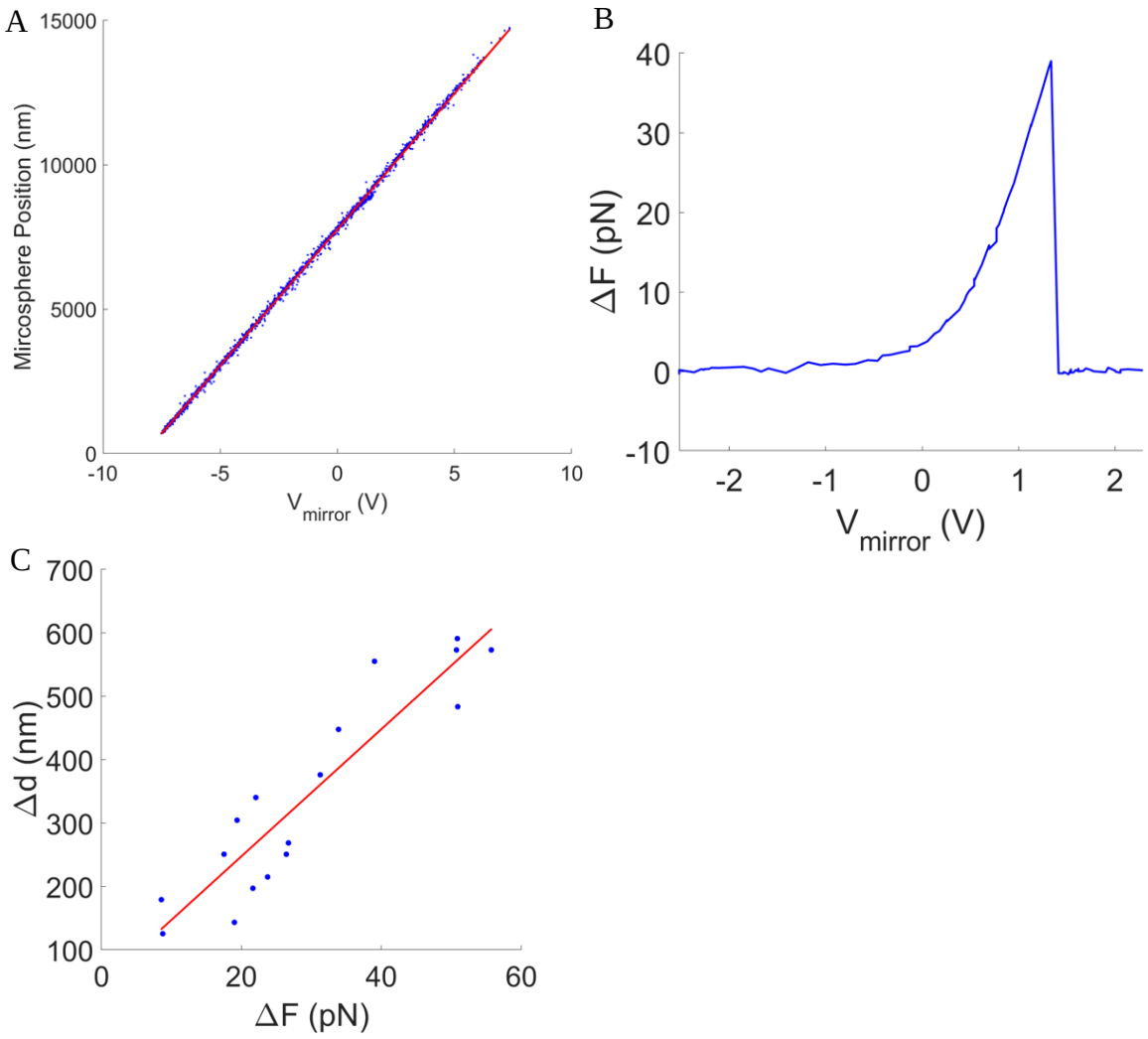

Figure 4.4 (A) Microsphere position, determined by image centroid tracking, versus mirror control signal (blue points). The red line shows a linear fit. **(B)** Example of a measured DNA detachment event where the separation of the traps is increased (by increasing V_{mirror}) and the force is measured to suddenly drop to zero. **(C)** Measurements of the movement of the microsphere back to the trap center, determined by image centroid tracking, after DNA detachment events that occurred at different force levels. The red line is a linear fit to the data.

Supplementary Materials for 'Determining Trap Compliances, Microsphere Size Variations, and Response Linearities in Single DNA Molecule Elasticity Measurements with Optical Tweezers' by Y. Mo et al.

Optical Tweezers Instrument

A schematic diagram of the instrument is shown below and additional details on the design and construction are given in Ref 21. A 1064-nm linearly polarized CW fiber laser (IPG Photonics, Inc.) is expanded and collimated by a telescope (T_1) before being split into two orthogonal polarization by a polarizing beam splitter (PB_1). One beam is reflected by a stationary mirror, while the other reflects off a mirror tilted by a computer-controlled piezoelectric actuator (PM). The two beams are re-directed to follow the same path by a second polarizing beam splitter (PB_2) and relayed by a second telescope (T_2), to make the plane of the rotating mirror conjugate to the back focal plane of a 60x 1.2-NA water-immersion objective (O_1; Olympus, Inc.), so that steering the mirror moves one trap in the sample plane. This objective focuses the beams to form the two traps in the fluid chamber (FC), each ~150 mW in the sample plane. An identical objective (O_2) collects the exiting beams and a final polarizing beam splitter (PB_3) separates the beams by polarization so that each can be directed to a position-sensing detector (PSD_1 and PSD_2, On-Trak, Inc). Relay lenses are used to make the back focal plane of O_2 conjugate to the detector surface so that the PSDs give signals proportional to the beam deflections and thus the transverse forces acting on the microspheres. A ~10 mW blue LED ('light source'), imaged onto the back focal plane of O_2,

provides illumination for imaging by O_1 of the sample plane onto a video camera (Watec WAT-902H2).

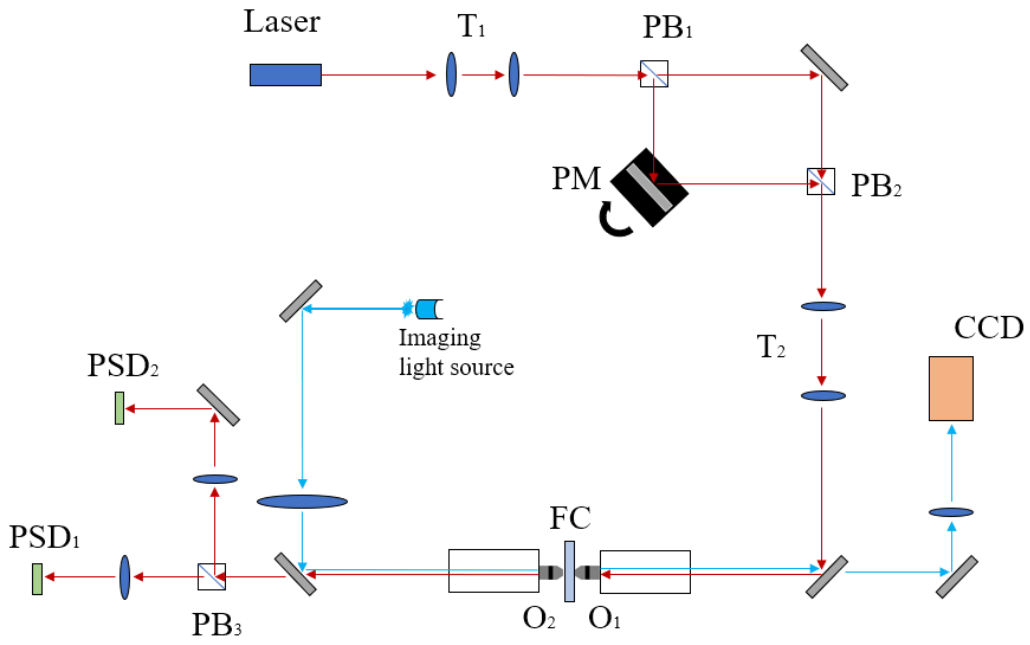

Figure 4.5 Diagram of the customized optical tweezers system in our lab.

Determining force and displacement scale factors

The force scale factor α was determined by measuring the force where the DNA undergoes the overstretch transition. As described in detail in Ref. 22, we measured the PSD voltage at which this occurs by finding the midpoint between the intersection points of a linear fit to the overstretch plateau and two polynomial fits to the low-force stretching regime and the high-force stretching regime after the overstretch transition. The displacement scale factor β was also determined as described in detail in Ref. 22. Briefly, the values of V_{mirror} that correspond to the special value $F = 33.4$ pN, where $x = L$, were determined for the two different DNA construct lengths (V_1 and

V_2). At this special force value Eq. (6) gives $\beta = (L_2 - L_1)/(V_2 - V_1)$, where L_1 and L_2 are the known contour lengths of the two DNA constructs.

Preparation of DNA constructs and microspheres

DNA constructs were prepared by PCR from lambda phage DNA using biotin and anti-digoxygenin labeled primers (IDT DNA, Inc.) and protocols given in Ref. 41. The 10.7 kbp DNA construct used the forward primer Biotin-5'-CATCATCATGCAGAACATGCGTGACGAAGAGCTG-3' and the reverse primer dig-5´-ATACGCTGTATTCAGCAACACCGTCAGGAACACG-3'. The 25.3 kbp DNA construct used the forward primer Biotin-5'-CTGATGAGTTCGTGTCCGTACAACTGGCGTAATC-3' and the reverse primer dig-5'-ATCCGATCTGCGTTACCGAATGGATGGATG-3'. DNA-coated microspheres were prepared by incubating 10 μL of 0.5% w/v 2.1 μm streptavidin-coated microspheres (Spherotech, Inc.) in 1x PBS (Phosphate Buffered Saline) with ~70 ng of DNA and 20 μg BSA (Bovine Serum Albumin) (Sigma-Aldrich, Inc.) for 20 minutes. Digoxygenin-coated microspheres were prepared by incubating 5 μL of 5% w/v 2.3um protein G-coated microspheres (Spherotech, Inc.) suspended in 1x PBS with 200 ng of anti-digoxygenin (Sigma-Aldrich, Inc.) for 45 minutes, as described in Ref. 41 and in the PhD Thesis of Nicholas Keller (Univ. of California, San Diego, 2016).

DNA Force-Extension Measurements

Multiple force-extension curves for the two different DNA construct lengths were collected. Tethers were formed by briefly bringing the trapped DNA- and anti-DIG-coated microspheres

together and then separating them while checking for the increase in PSD signal indicative of increasing force on the microspheres. Once a tether was formed, a force-extension curve was measured by increasing V_{mirror} at a fixed rate of 500 nm/s between values V_1 and V_2, where V_1 is the largest control voltage such that the PSD signal does not differ from background, and V_2 corresponds to a microsphere separation distance large enough that the DNA tether will either detach or complete an overstretch transition before reaching. These values were determined empirically. PSD data were recorded at a rate of 1 kHz. The background PSD signal as a function of $V_{mirror}$was subtracted from each force-extension dataset.

Error in the determination of the series trap compliance

We analyzed uncertainties/error sources in the determination of the compliance parameter γ by considering the effect of various factors on the use of Equation (4) to fit the DNA force-extension datasets:

i) Wenner et al. (Ref. 35) report an uncertainty of 2 nm in the DNA persistence length (P), which through our Equations (1) and (2) results in an uncertainty of 0.9% in the parameter B. Through Equation (4) this results in an uncertainty of 0.2% in the determination of γ.

ii) Wenner et al. (Ref. 35) report an uncertainty of 217 pN in the DNA stretch modulus (S), which results in an uncertainty of 11.6% in the parameter B. Through Equation (4) this results in an uncertainty of 3.8% in the determination of γ.

iii) Uncertainty in the force measurements is caused by an uncertainty of 2% in the calibration factor α, due to uncertainty in the DNA overstretch transition force plateau reported by Wenner et al. (Ref. 35) and random measurement errors in our measurements of the overstretch transition (as

discussed by delToro et al. (Ref. 22)). Through Eq. (4) this uncertainty in the force measurements results in an uncertainty of 1.5% in the determination of γ.

iv) There is uncertainty in the trap displacement values caused by an uncertainty of ~0.7% in calibration factor β as discussed by delToro et al. (Ref. 22). Through Eq. (4) this uncertainty results in an uncertainty of 0.95% in the determination of γ.

v) Our assumption that the DNA force-extension relationship is linear in the range from 25-45 pN, when it actually has a small nonlinearity, causes an error of 0.96% in the determination of γ.

Note: The five sources of uncertainty listed above do not have completely independent effects. By systematically varying all these parameters together, we calculated that the maximum induced uncertainty in gamma from these factors is ~6%. In addition, we considered the effect of random measurement errors in the force vs. extension measurements, as follows:

vi) Noise in the force measurements, due to Brownian and instrumental noise, also causes uncertainty in the determination of gamma from the ensemble of force-extension measurements. We investigated this by generating an ensemble of simulated datasets with the DNA force law fixed according to Eq. (1) and random Gaussian noise added to each dataset. The signal-to-noise ratio for the simulated datasets was set equal to that in the measured force-extension datasets. By fitting these simulated datasets to Eq. (4) we determined that this noise causes 0.3% uncertainty in the determination of γ.

References

[1] Moffitt JR, Chemla YR, Smith SB, Bustamante C. Recent advances in optical tweezers. Annu Rev Biochem. 2008;77:205-28.

[2] Fazal FM, Block SM. Optical tweezers study life under tension. Nature Photonics. 2011;5:318-21.

[3] Jones PH, Maragò OM, Volpe G. Optical tweezers: Principles and applications: Cambridge University Press; 2015.

[4] Shabestari MH, Meijering A, Roos W, Wuite G, Peterman E. Recent advances in biological single-molecule applications of optical tweezers and fluorescence microscopy. Methods in enzymology: Elsevier; 2017. p. 85-119.

[5] Polimeno P, Magazzu A, Iati MA, Patti F, Saija R, Boschi CDE, Donato MG, Gucciardi PG, Jones PH, Volpe G, Marago OM. Optical tweezers and their applications. Journal of Quantitative Spectroscopy and Radiative Transfer. 2018;218:131-50.

[6] Robertson-Anderson RM. Optical tweezers microrheology: from the basics to advanced techniques and applications. ACS Publications; 2018.

[7] Choudhary D, Mossa A, Jadhav M, Cecconi C. Bio-molecular applications of recent developments in optical tweezers. Biomolecules. 2019;9:23.

[8] Berns MW. Laser scissors and tweezers to study chromosomes: a review. Frontiers in Bioengineering and Biotechnology. 2020;8.

[9] Wang MD, Yin H, Landick R, Gelles J, Block SM. Stretching DNA with optical tweezers. Biophysical journal. 1997;72:1335.

[10] Bustamante C, Macosko JC, Wuite GJL. Grabbing the cat by the tail: manipulating molecules one by one. Nature Reviews Molecular Cell Biology. 2000;1:130-6.

[11] Gemmen GJ, Millin R, Smith DE. DNA looping by two-site restriction endonucleases: heterogeneous probability distributions for loop size and unbinding force. Nucleic acids research. 2006;34:2864-77.

[12] Tsay JM, Sippy J, DelToro D, Andrews BT, Draper B, Rao V, Catalano CE, Feiss M, Smith DE. Mutations altering a structurally conserved loop-helix-loop region of a viral packaging motor change DNA translocation velocity and processivity. J Biol Chem. 2010;285:24282-9.

[13] Chemla YR, Smith DE. Single-molecule studies of viral DNA packaging. In: Rao V, Rossmann MG, editors. Viral Molecular Machines. Boston, MA: Springer; 2012. p. 549-84.

[14] Keller N, Grimes S, Jardine PJ, Smith DE. Repulsive DNA-DNA interactions accelerate viral DNA packaging in phage phi29. Phys Rev Lett. 2014;112:248101.

[15] Migliori AD, Keller N, Alam TI, Mahalingam M, Rao VB, Arya G, Smith DE. Evidence for an electrostatic mechanism of force generation by the bacteriophage T4 DNA packaging motor. Nature Communications. 2014;5:4173.

[16] Pongor CI, Bianco P, Ferenczy G, Kellermayer R, Kellermayer M. Optical trapping nanometry of hypermethylated CPG-island DNA. Biophysical journal. 2017;112:512-22.

[17] Lehmann K, Shayegan M, Blab GA, Forde NR. Optical tweezers approaches for probing multiscale protein mechanics and assembly. Frontiers in Molecular Biosciences. 2020;7.

[18] Bustamante C, Alexander L, Maciuba K, Kaiser CM. Single-molecule studies of protein folding with optical tweezers. Annual review of biochemistry. 2020;89:443-70.

[19] Abbondanzieri EA, Greenleaf WJ, Shaevitz JW, Landick R, Block SM. Direct observation of base-pair stepping by RNA polymerase. Nature. 2005;438:460-5.

[20] Moffitt JR, Chemla YR, Izhaky D, Bustamante C. Differential detection of dual traps improves the spatial resolution of optical tweezers. Proceedings of the National Academy of Sciences of the United States of America. 2006;103:9006-11.

[21] delToro D. Doctoral Dissertation, University of California San Diego. 2015.

[22] delToro D, Smith DE. Accurate measurement of force and displacement with optical tweezers using DNA molecules as metrology standards. Appl Phys Lett. 2014;104:143701.

[23] Smith SB, Cui Y, Bustamante C. Optical-Trap Force Transducer that Operates by Direct Measurement of Light Momentum. Methods in Enzymology. 2003;361::134.

[24] Rickgauer JP, Fuller DN, Smith DE. DNA as a metrology standard for length and force measurements with optical tweezers. Biophys J. 2006;91:4253-7.

[25] Neuman KC, Block SM. Optical trapping. Rev Sci Instrum. 2004;75:2787-809.

[26] Berg-Sørensen K, Flyvbjerg H. Power spectrum analysis for optical tweezers. Review of Scientific Instruments. 2004;75:594-612.

[27] Tolić-Nørrelykke SF, Schäffer E, Howard J, Pavone FS, Jülicher F, Flyvbjerg H. Calibration of optical tweezers with positional detection in the back focal plane. Review of scientific instruments. 2006;77:103101.

[28] Viana N, Mazolli A, Maia Neto P, Nussenzveig H, Rocha M, Mesquita O. Absolute calibration of optical tweezers. Applied physics letters. 2006;88:131110.

[29] Jahnel M, Behrndt M, Jannasch A, Schäffer E, Grill SW. Measuring the complete force field of an optical trap. Optics letters. 2011;36:1260-2.

[30] Sarshar M, Wong W, Anvari B. Comparative study of methods to calibrate the stiffness of a single-beam gradient-force optical tweezers over various laser trapping powers. Journal of biomedical optics. 2014;19:115001.

[31] Bui AA, Kashchuk AV, Balanant MA, Nieminen TA, Rubinsztein-Dunlop H, Stilgoe AB. Calibration of force detection for arbitrarily shaped particles in optical tweezers. Scientific reports. 2018;8:1-12.

[32] Yale P, Konin J-ME, Kouacou MA, Zoueu JT. New Detector Sensitivity Calibration and the Calculation of the Interaction Force between Particles Using an Optical Tweezer. Micromachines. 2018;9:425.

[33] Melo B, Almeida F, Temporão G, Guerreiro T. Relaxing constraints on data acquisition and position detection for trap stiffness calibration in optical tweezers. Optics Express. 2020;28:16256-69.

[34] Bustamante C, Smith SB, Liphardt J, Smith D. Single-molecule studies of DNA mechanics. Current opinion in structural biology. 2000;10:279-85.

[35] Wenner JR, Williams MC, Rouzina I, Bloomfield VA. Salt dependence of the elasticity and overstretching transition of single DNA molecules. Biophysical journal. 2002;82:3160-9.

[36] Smith SB, Cui Y, Bustamante C. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. Science (New York, NY). 1996;271:795-9.

[37] Neuman KC, Nagy A. Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. Nat Methods. 2008;5:491-505.

[38] Odijk T. Stiff chains and filaments under tension. Macromolecules. 1995;28:7016-8.

[39] Marko JF, Siggia ED. Stretching dna. Macromolecules. 1995;28:8759-70.

[40] Lipfert J, Skinner GM, Keegstra JM, Hensgens T, Jager T, Dulin D, Köber M, Yu Z, Donkers SP, Chou FC, Das R. Double-stranded RNA under force and torque: similarities to and striking differences from double-stranded DNA. Proceedings of the National Academy of Sciences. 2014;111:15408-13.

[41] Fuller DN, Gemmen GJ, Rickgauer JP, Dupont A, Millin R, Recouvreux P, Smith DE. A general method for manipulating DNA sequences from any organism with optical tweezers. Nucleic Acids Res. 2006;34:e15.

[42] Farré A, van der Horst A, Blab GA, Downing BPB, Forde NR. Stretching single DNA molecules to demonstrate high-force capabilities of holographic optical tweezers. Journal of biophotonics. 2010;3:224-33.

[43] Bormuth V, Jannasch A, Ander M, van Kats CM, van Blaaderen A, Howard J, Schäffer E. Optical trapping of coated microspheres. Optics express. 2008;16:13831-44.

[44] Keller N, delToro D, Smith DE. Single-Molecule Measurements of Motor-Driven Viral DNA Packaging in Bacteriophages Phi29, Lambda, and T4 with Optical Tweezers. In: Lavelle C, editor. Molecular Motors Methods and Protocols. New York, NY: Humana Press; 2018.

[45] Ortiz D, delToro D, Ordyan M, Pajak J, Sippy J, Catala A, Oh CS, Vu A, Arya G, Feiss M, Smith DE. Evidence that a catalytic glutamate and an ‘Arginine Toggle’ act in concert to mediate ATP hydrolysis and mechanochemical coupling in a viral DNA packaging motor. Nucleic Acids Research. 2018;47:1404-15.

[46] Berndsen ZT, Keller N, Smith DE. Continuous Allosteric Regulation of a Viral Packaging Motor by a Sensor that Detects the Density and Conformation of Packaged DNA. Biophys J. 2015;108:315-24.

[47] Richardson AC, Reihani SNS, Oddershede LB. Non-harmonic potential of a single beam optical trap. Optics express. 2008;16:15709-17.

[48] Ueberschär O, Wagner C, Stangner T, Gutsche C, Kremer F. A novel video-based microsphere localization algorithm for low contrast silica particles under white light illumination. Optics and Lasers in Engineering. 2012;50:423-39.

Chapter 5

Bacteriophage Lambda DNA Packaging: Efforts to Detect Translocation Steps

Abstract

Viral dsDNA packaging motors are members of the Additional Strand Catalytic E (ASCE) superfamily of NTPase proteins, other members of which also participate in protein translocation/degradation, ATP synthesis, DNA translocation and recombination, etc. To investigate the mechanism by which these motors translocate DNA we used optical tweezers to measure single DNA molecule packaging in the bacteriophage λ, driven by the so called λ terminase motor complex. Here we conduct high spatial and temporal resolution measurements of the DNA translocation under conditions which the motor is slowed by a chemical inhibitor, in an effort to see if we can resolve a smallest quantized translocation step size. Previous studies found evidence that the phage φ29 motor translocates in rapid bursts of four 2.5 bp steps, attributed to translocations made by four motor subunits after they each hydrolyze one ATP, adding up to a rapid 10 bp quantized translocation. However, it is unclear whether this mechanism is universal for other viral motors. The phage lambda differs in many ways from the φ29 phage in that it infects a different host cell (*E. coli* vs. *B. Subtilis*) and has a ~3 times larger genome length. The lambda motor also lacks an RNA component that φ29 has, has an endonuclease function, and translocates DNA ~3 times faster than the φ29 motor. Our

preliminary results and preliminary data analysis for the lambda motor provide possible evidence for a step (or step burst) size of around 5 bp rather than 10 bp, but do not reject the possibility of a 2.5 bp individual subunit step size. We also provided new evidence of a package-slip-pause behavior for the λ motor where a transient ~5 bp backwards slip occurs right after a ~5 bp translocation step, followed by a long pause. We propose a possible mechanism for such events.

5.1 Introduction

Additional Strand Catalytic E (ASCE) group and Kinase-GTPase (KG) group are two main structural groups of P-loop NTPases. The ASCE superfamily has a similar structurally conserved core with motifs which participate in ATP binding and hydrolysis. The subfamilies of ASCE, such as RecA-like ATPases, Ftsk/HerA dsDNA translocases, AAA+ enzymes, and Superfamily 2 (SF2) helicases, are involved in many biological procedures, for instance, protein translocation/degradation, ATP synthesis, DNA translocation and recombination, etc.[1-7] The phage DNA packaging motors are good examples of these motors, that can be studied in detail using single-molecule techniques and thus may help shed light on the ASCE superfamily of motor proteins in general.

Two major viral proteins, the Ftsk/HerA family and the terminase large subunit (TLS) family, have been studied by structural methods and sequence homologies.[8-17] Single bp resolution single-molecule optical tweezers methods were also employed to investigate the motor dynamics and mechanisms of the phage φ29's motor, in which the a ringed Ftsk/HerA NTPase composed by five gp16 proteins subunits translocated dsDNA into phage's procapsid. It was found by Bustamante and cooperators that five gp16 proteins subunits translocated DNA in 10

base-pairs (bp) increments consisting of ‘bursts’ of four rapid and successive 2.5 bp steps.[18] The studies suggest that each 2.5 bp step follows hydrolysis of one ATP by one of the motor subunits, and in each cycle one of the five subunits does not translocate so there are only four steps. The action of the subunits is tightly coordinated such that they usually do not hydrolyze ATP and translocate until all five subunits bind ATP and then there is a rapid burst of four steps causing a 10 bp DNA translocation. However, it is not clear if this translocation stepping dynamic is universal for all viral or phage motors, such as for the phage λ motor we study here.

Additionally, T4 exhibits frequent unslipping (sometimes called ‘unpackaging’) that happens especially when there are large load forces on the motor or when ATP concentration is low. In of the φ29 studies the viral DNA translocation stepping size was acquired by measuring the distance between two sequential pauses between the bursts of steps but the probability of DNA slipping out before pausing has not been discussed.

With φ29 it was also found that addition of orthovanadate (VO_4^{3-}, from added sodium orthovanadate) causes pauses in DNA translocation and sometimes clusters of multiple pauses separated by 10 bp are also observed.[1] This it thought to occur because orthovanadate is a phosphate (Pi) analog that can form stable complexes with ADP and delays dissociation of ADP from the binding pocket. In this case, for φ29, a single burst of four 2.5 bp translocation steps occurs separated by pauses that are easier to detect because they are much longer than the short dwells that normally occur between bursts when packaging in ATP alone. Therefore, as a preliminary effort to investigate potential similarities or differences between the φ29 and lambda motors, we conducted studies, described in this chapter, where we added sodium orthovanadate to the solutions during packaging measurements and looked for similar pause clusters.

. The preliminary data and analysis find some evidence for a possible different step size for λ than for φ29, although more work may be needed to verify this preliminary conclusion. We also provided some evidence for a new package-slip-pause dynamic where we observe small backwards 5 bp steps occurring right after a small translocation step and followed by a pause.

5.2 Methods

5.2.1 DNA Construct

5.2.1.1 Dual Labeled λ Packaging DNA with *cos* Site

A 13,881bp plasmid, named as pJM1, containing the full λ cos site was prepared by the Feiss Lab (University of Iowa). Labeled PCR primers were bought from IDT Inc. The sequences of the forward and backward primers were Dig-5'-TCGATAATCGTGAAGAGTCGGCGAGCCTGGTTAG-3' and Biot-5'-TACGTCGAAGTGACCAACTAGGCGGAATCGGTAG-3' respectively. Dual labeled DNA was PCRed by the same protocol as described in Section 2.2.1.2.

5.2.1.2 Production of λ Terminase and Procapsids

Lambda terminase from an *E. coli* cell extract and purified procapsids used in these studies were obtained from the Feiss Lab (University of Iowa) as described in Ref. [19].

5.2.2 Pre-Stalled Packaging Protocol

5.2.2.1 Beads

Prepare SA Beads

1. Spin down 20µL SA beads and remove the supernatant;
2. Wash 2X in 20µL 1X TM;
3. Resuspend in 20µL 1X TM;
4. Add 0.3µL 100mg/mL BSA;
5. Use as described below. Extra can be stored in the 4℃ refrigerator for use another day.

Prepare Antibody Beads

1. Spin down 75µL protein G beads and remove the supernatant;
2. Wash 2X with 50µL 1X PBS;
3. Resuspend with 15µL 1X PBS;
4. Add 8µL purified λ anti-sera. Mix well by gently pipetting up and down and/or gently flicking;
5. Incubate on rotator for 30 minutes at room temperature;
6. Spin down and remove the supernatant;
7. Wash twice with 50µL 1X PBS;
8. Wash with 50µL 1X TM;
9. Resuspend in 15µL 1X TM;
10. Store at 4℃ until ready to use. Can be used over multiple days.

5.2.2.2 Buffer

10X TM Buffer (250 mM Tris pH 7.5, 50 mM $MgCl_2$):

1. 700µL H_2O
2. 250µL 1M Tris pH7.5
3. 50µL 1M $MgCl_2$

5 mL Packaging Buffer:

1. 4.83mL H_2O
2. 12.5µL 1M Tris Ph7.5
3. 25µL $MgCl_2$
4. 2.5µL BSA (100mg/mL)
5. 25µL 100mM ATP
6. 1.25µL 100mM Na_3VO_4 (Only for stepping experiment)

5.2.2.3 Syringe Solution

Lambda Antibody Beads (300µL)

1. 270µL H_2O
2. 30µL 10X TM
3. ~3µL λ antibody beads

Complexes or SA Beads (250µL)

1. 200µL H_2O
2. 25µL 10X TM
3. 0.25µL BSA (100mg/mL)
4. ~4µL Complexes or SA Beads

5.2.2.4 Stalled Complexes Preparation

See Table 5.1

5.2.2.5 Complexes Bead Preparation

1. Gently add 0.15 µL stalled complex to 7 µL washed SA beads by slowly pipetting up and down multiple times and swirling the pipet tip;

2. Incubate these beads with complexes solution at room temperature on the rotator for 20 mins;

3. Immediately put the remaining complexes on ice.

5.2.3 Data Processing

5.2.3.1 Segmentation with SIC

The edge of each pause section was found by using a Schwarz Information Criterion (SIC) analysis.[20] The SIC score was computed by

$$score = -2log\mathcal{L}(\theta) + plog(n)$$

where $\mathcal{L}(\theta)$ was the likelihood function of the statistical model, θ was the vector of parameters, *p* was the number of parameters and *n* was the number of data points. In practice, we broke a piece of packaging data into two sections then computed the mean of both sections. A step function was employed to fit the data where the horizontal line was at the separation point and the y axis values were the mean of each section. The score of this fitting was calculated by applying the SIC formula. The separated point with maximum score to assess was picked to segment the packaging signal. Then the segmentation repeated until the score were maximized.

5.2.3.2 Data Section Alignment and Pairwise Distance Distribution Analysis

The pause data was cropped by the segmentation algorithm and only the segments with slope < 28 bp/s (statistically consistent with zero velocity) were marked as pause sections for pairwise distance statistics. The positions of the considered pause sections were calculated as the mean (or median) of these sections. A cluster of pauses was identified to be a group of pause sections for which the maximum distance between each two pauses was less than 30 bp. The baseline reference length of a cluster was taken to be the mean of the longest pause in that

cluster. Then we aligned all clusters of pauses along their baselines (with the baseline length subtracted).

5.3 Preliminary Results

We identified N = 55 Lambda packaging events (out of 75 'single tether' packaging events) in which the pause clusters of the type described above were detected. Within these, 104 pause clusters were manually cropped for data analyses. Examples of pause clusters were shown in Fig. 5.1 and Fig. 5.4.

5.3.1 Detection of Discrete Translocation Steps

All pause sections of duration > 0.01 second were segmented and their baselines were computed by averaging over all points of the corresponding sections as we described in 5.2.3, and the pairwise distances between every two pauses in the same cluster were calculated. The histogram of the distribution of pairwise distances is plotted in Fig. 5.2. There is a clear major peak centered at ~5 bp and, unlike what was found for φ29 there was no significant peak at ~10 bp. This preliminary data and preliminary analysis suggests there might be a fundamental difference between the Lambda and φ29 motors, although the mechanism is not completely certain. One possibility, if the Lambda motor does operate in a similar manner as the φ29 motor, in which there is tightly coordinated actions of four subunits that translocate the DNA in a burst

of four steps, is that the size of the burst for Lambda is ~5 bp instead of 10 bp. However, a second possibility is that, under these conditions, the Lambda motor does not operate in a similar way making coordinated bursts of steps, but instead the ~5 bp steps we detect are individual discrete steps that individual subunits make after hydrolyzing a single ATP (although we cannot rule out that they could be pairs of two 2.5 bp steps made by two subunits). Close inspection of Fig 5.2 indicates that there may be a secondary peak at ~2.7 bp. A Gaussian distribution curve was applied to fit both peaks such that we acquired the height (maximum) and full width at half maximum (FWHM) for both peaks. The fitting indicated the FWHM of the primary peak was 2.75 bp which was slightly greater than the distance between the centers of two peaks and larger than the single motor subunit step size for the phage φ29 motor found by G. Chistol and S. Liu et al. in Ref. [21]. The resolution of our current measurements cannot rule out the possibility that each subunit of the Lambda motor might make either a 5 bp or 2.5 bp step per ATP.

We also tabulated the pauses duration for N = 2424 pauses and a histogram is presented in Fig. 5.3. We find that the distribution can be reasonably well fit by an exponential function. The observation of a pause distribution following an exponential distribution is similar to that found with the pauses observed for φ29 with Na_3VO_4, which suggests that in both cases the pauses are caused by the translocation of individual subunits being inhibited due to delayed ADP release that randomly follow an exponential (interval) probability distribution function in time (as expected for a Poisson process).

5.3.2 Detection of Small Package-Slip-Pause Events

A new behavior, different than any reported previously for φ29, was also discovered while performing λ packaging measurements with added Na_3VO_4. A 'package-slip-pause' pattern, such as that shown in the red segment in Fig. 5.4, was observed. Specifically, a spike in which a small amount of DNA was packaged before rapidly slipping out again, often appeared between two pauses separate by the ~5 bp steps. These features were smoothed by the data averaging and thus ignored in the data analysis method described in the previous sections. However, upon closer inspection, we were able to detect 75 such spikes and compute statistics on them. (Fig. 5.5) The average height of spike was 4.8±3.5 bp, suggesting that a subunit of λ motor, in each red segment, translocated ~10 bp (or possibly made two ~5 bp steps) but then this was immediately followed by a slip of ~5 bp. Although more measurements would be needed to verify this finding, we propose one possible speculative model that could possibly explain this effect (see Fig. 5.6).

(a) Subunit #1 hydrolyzes ATP and translocates 5 bp and signals Subunit #2

(b) Subunit #2 hydrolyzes ATP and translocates 5 bp

(c) Subunit #3 has ADP + vanadate bound and grips DNA, but only for a short time and then slips (explains the +5 bp spike)

(d) The DNA slips backwards and after slipping back 5 bp is gripped again by Subunit #2 (which it comes back into the same alignment it had with before) which can grip it because it has bound a new ATP

(e) However now there is a pause because Subunit #2 does not receive another signal from Subunit #1 because subunit #1 would not signal it again (until subunit #1 is able to hydrolyze and translocate again after all other subunits have translocated).

(f) Instead of firing in the pattern 1-2-3-4, the motor has misfired in the pattern 1-2-3-2 because subunit #3 had an ADP+vanadate bound.

Additionally, experiment of T4 packaging dsDNA with Na_3VO_4 were also conducted and those are described in Chapter 7 of this thesis. In this case we were not able to detect any pause clusters with pauses separated by small, discrete distances. Instead, all packaging episodes were interrupted by slipping events with variable lengths.

5.4 Acknowledgements

We thank Damian del Toro for his collaboration in conducting measurements for this project. We also thank Michael Feiss and Jean Sippy for providing the λ reagents and Jeffrey Moffitt for providing additional, useful information about the single-molecule φ29 stepping analysis.

Chapter 5, in part is currently being prepared for submission for publication of the material by Youbin Mo and Douglas E. Smith. We thank Damian del Toro for his collaboration in conducting measurements for this project. The dissertation author was the primary investigator and author of this material.

Table 5.1 λ stalled complexes preparation.

CP1 Total	5 μL
H_2O	1.625 μL
10X TM	0.5 μL
13.7kbp DNA	0.625 μL
Terminase (add by slowly pipetting up and down while gently swirling)	0.5 μL
ATP (5mM)	0.5 μL
Incubate at room temperature for 5 min	
γ-S-ATP (5mM)	0.5 μL
Incubate at room temperature for 5 min	
λ prohead (stalled at 4°C)	0.5 μL

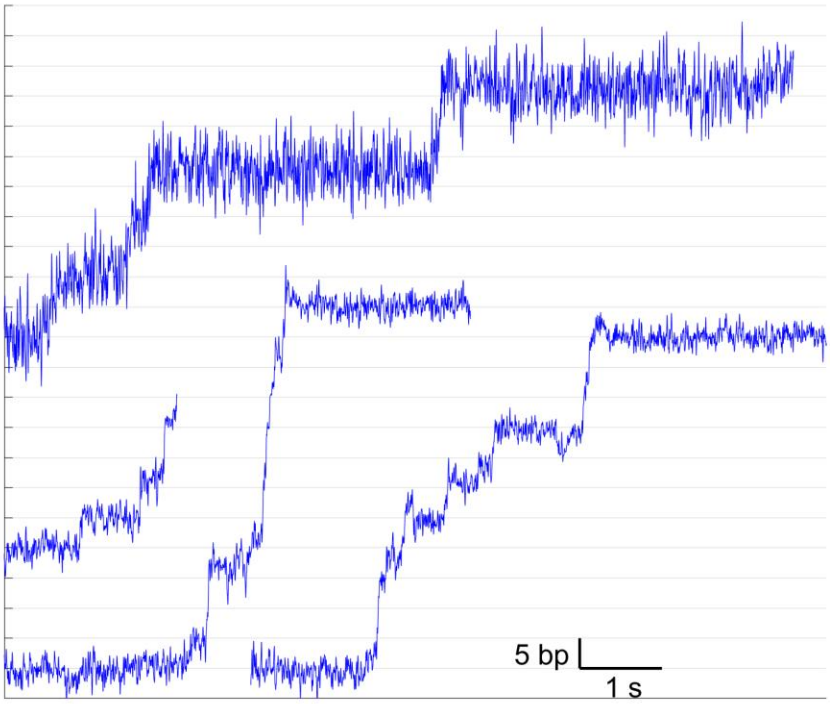

Figure 5.1 Orthovanadate (VO_4^{3-}) induces pausing events. Examples of measurements of length of DNA packaged when λ is packaging pJM1 DNA measured at 100 kHz sampling rate in 500 μM ATP and 25μM Na_3VO_4. N = 102 individual complexes were measured and long pauses (> 1 second) were segmented to look for pausing clusters. In this figure, four multiple stepping events were selected out of 418 pausing clusters. The traces were decimated and smoothed by averaging every 100 points in a sliding window. The horizontal lines in this plot indicate 5 bp distance steps.

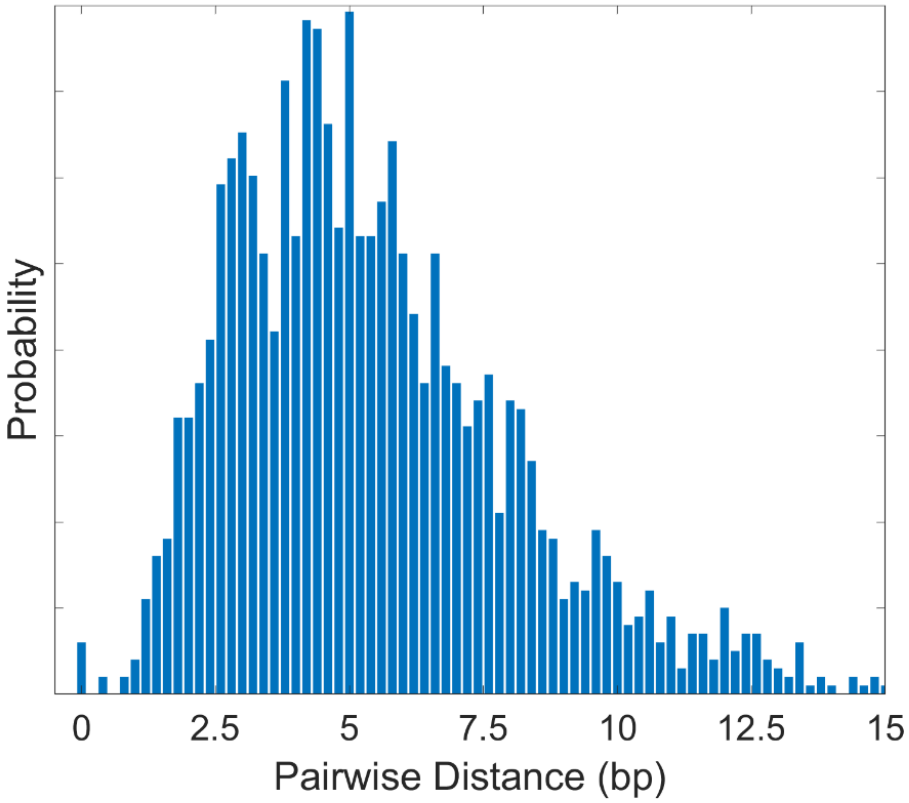

Figure 5.2 The frequency histogram obtained by counting the distance of two adjacent pauses. (A) Pausing clusters were segmented from the data shown in Figure 5.1. The data was smoothed and decimated by averaging every 100 points in a sliding window. We use the SIC algorithm to find each single pause from pause clusters then represent a single pause by the median of this segment. A single peak appears at 5 bp and the Full Width at Half Maximum (FWHM) is 5.5 which means the spatial resolution of our instrument is 2.75 bp. If there is 2.5 bp stepping, we cannot detect it.

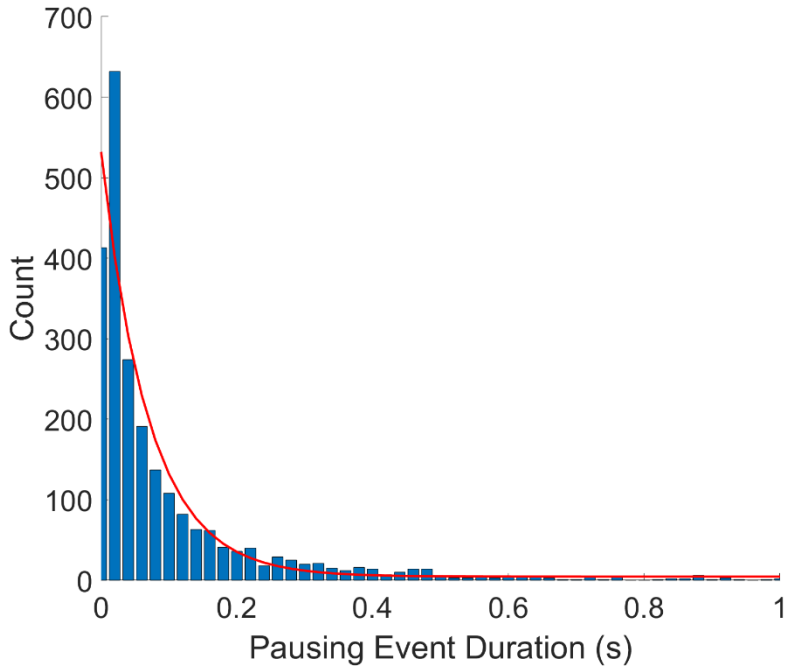

Figure 5.3 Histogram of the measured durations of single pauses. N = 2424 single pauses were found from the data shown in Fig. 5.2. The height of bar represents the count of pause in the duration intervals and the red line is a single exponential decay fit. The gap between the first and second bar (representing 0 - 0.02 and 0.02 - 0.04 sec) is present because the SIC method as described cannot find pauses of duration shorter than 0.01 second.

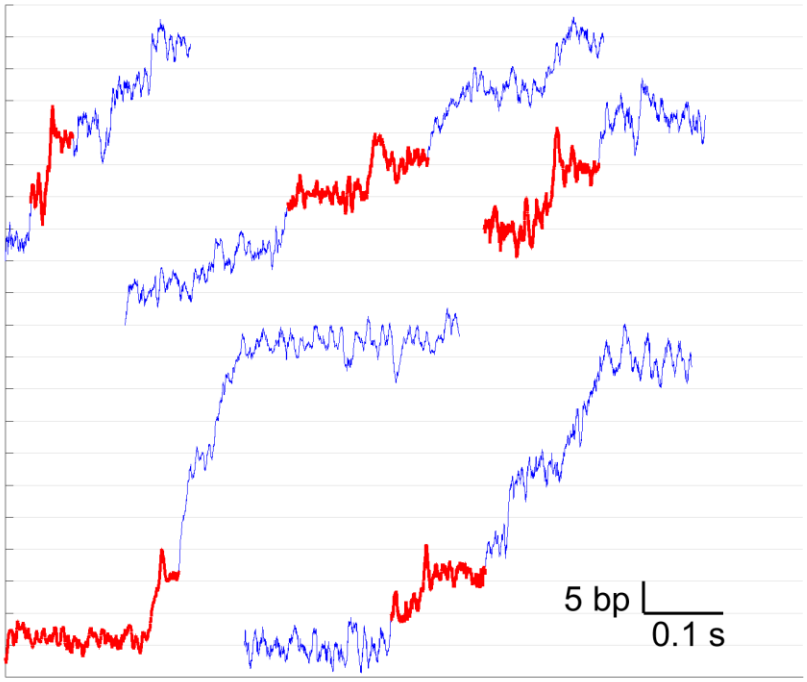

Figure 5.4 Examples of short slipping 'spikes' occurring between pauses. We observed N = 75 such spikes in 418 pause cluster segments. The sections of data where this slipping was detected is plotted in red. The slipping generally happens after bacteriophage λ terminase experiences a short pause. The motor packages the DNA with the regular motor velocity but suddenly stops, slips out a few base pair then pauses.

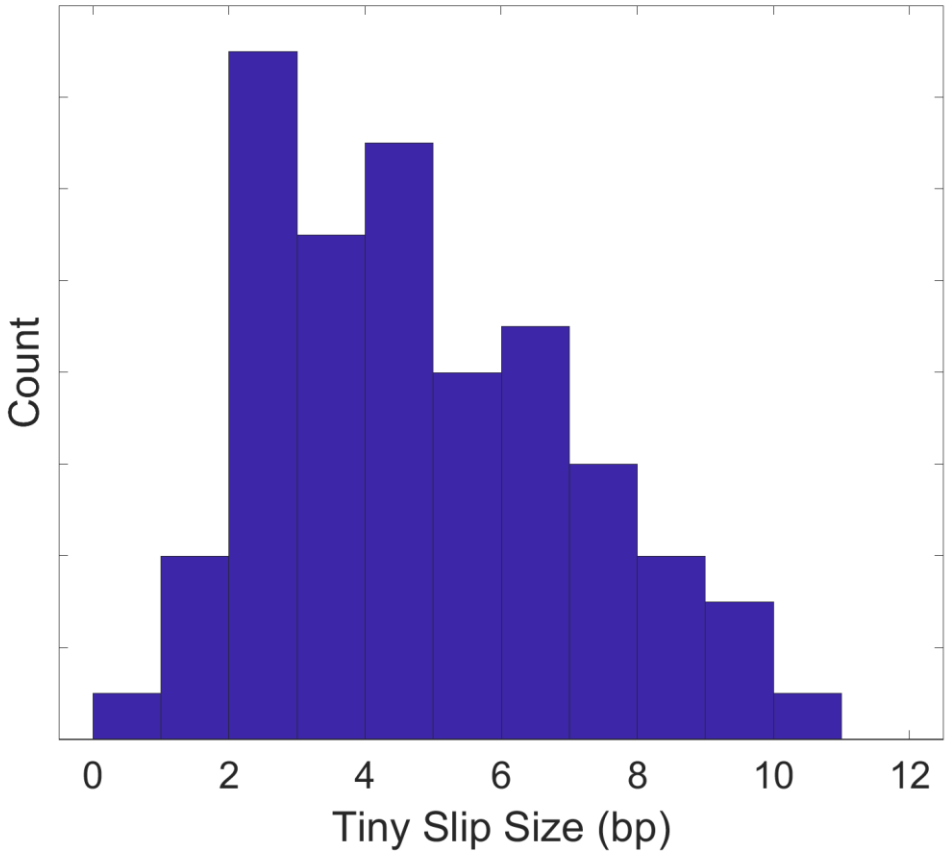

Figure 5.5 Histogram of the tiny slipping 'spike' sizes. We define the slipping-size as the distance from the peak value that DNA has been translocated to the pause position just next to the slip. The pause position is the median of the pausing section. N = 75 events were counted and the mean slipping size was 4.8 bp.

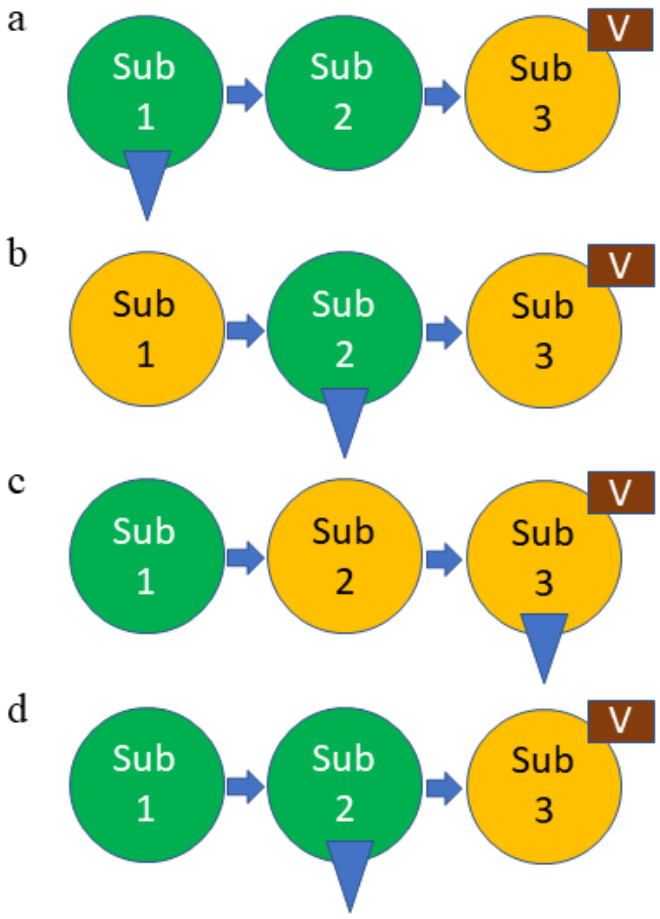

Figure 5.6 Illustration of possible λ slipping models. Green circle represents the subunit bound with ATP. Yellow circle represents the subunit bound with ADP. Brown rectangle with 'V' is the vanadate, and blue triangle indicates the subunit gripping DNA. (**a**) Subunit #1 hydrolyzes ATP and translocates 5 bp and signals Subunit #2. (**b**) Subunit #2 hydrolyzes ATP and translocates 5 bp. (**c**) Subunit #3 has ADP + vanadate bound and grips DNA, but only for a short time and then slips (explains the +5 bp spike). (**d**) The DNA slips backwards and after slipping back 5 bp is gripped again by Subunit #2 (which it comes back into the same alignment it had with before) which can grip it because it has bound a new ATP.

References

1. Snider, J., & Houry, W. (2008). AAA+ proteins: diversity in function, similarity in structure. *Biochemical Society Transactions*, *36*(1), 72-77.
2. Fairman-Williams, M. E., Guenther, U. P., & Jankowsky, E. (2010). SF1 and SF2 helicases: family matters. *Current opinion in structural biology*, *20*(3), 313-324.
3. Lyubimov, A. Y., Strycharska, M., & Berger, J. M. (2011). The nuts and bolts of ring-translocase structure and mechanism. *Current opinion in structural biology*, *21*(2), 240-248.
4. Wendler, P., Ciniawsky, S., Kock, M., & Kube, S. (2012). Structure and function of the AAA+ nucleotide binding pocket. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, *1823*(1), 2-14.
5. Ye, J., Osborne, A. R., Groll, M., & Rapoport, T. A. (2004). RecA-like motor ATPases—lessons from structures. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, *1659*(1), 1-18. 164.
6. Snider, J., Thibault, G., & Houry, W. A. (2008). The AAA+ superfamily of functionally diverse proteins. *Genome Biol*, *9*(4), 216.
7. Barkow, S. R., Levchenko, I., Baker, T. A., & Sauer, R. T. (2009). Polypeptide translocation by the AAA+ ClpXP protease machine. *Chemistry & biology*, *16*(6), 605-612.
8. Rao, V. B., & Mitchell, M. S. (2001). The N-terminal ATPase site in the large terminase protein gp17 is critically required for DNA packaging in bacteriophage T4. *Journal of molecular biology*, *314*(3), 401-411.
9. Mitchell, M. S., Matsuzaki, S., Imai, S., & Rao, V. B. (2002). Sequence analysis of bacteriophage T4 DNA packaging/terminase genes 16 and 17 reveals a common ATPase center in the large subunit of viral terminases. *Nucleic acids research*, *30*(18), 4009-4021.
10. Mitchell, M. S., & Rao, V. B. (2004). Novel and deviant Walker A ATP-binding motifs in bacteriophage large terminase–DNA packaging proteins. *Virology*, *321*(2), 217-221.
11. Sun, S., Kondabagil, K., Draper, B., Alam, T. I., Bowman, V. D., Zhang, Z., Hegde, S., Fokine, A., Rossmann, M.G., & Rao, V. B. (2008). The structure of the phage T4 DNA packaging motor suggests a mechanism dependent on electrostatic forces. *Cell*, *135*(7), 1251-1262.
12. Zhao, H., Christensen, T. E., Kamau, Y. N., & Tang, L. (2013). Structures of the phage Sf6 large terminase provide new insights into DNA translocation and cleavage. *Proceedings of the National Academy of Sciences*, *110*(20), 8075-8080.
13. Kainov, D. E., Mancini, E. J., Telenius, J., Lísal, J., Grimes, J. M., Bamford, D. H., Stuart, D.L., & Tuma, R. (2008). Structural basis of mechanochemical coupling in a hexameric molecular motor. *Journal of Biological Chemistry*, *283*(6), 3607-3617.
14. El Omari, K., Meier, C., Kainov, D., Sutton, G., Grimes, J. M., Poranen, M. M., Bamford, D.H., Tuma, R., Stuart, D.I., & Mancini, E. J. (2013). Tracking in atomic detail the

functional specializations in viral RecA helicases that occur during evolution. *Nucleic acids research, 41*(20), 9396-9410.

15. Hilbert, B., Hayes, J., Stone, N., Duffy, C., Sankaran, B., & Kelch, B. (2015). The Structure and Mechanism of the ATPase that Powers Viral Genome Packaging. *The FASEB Journal*, *29*(1 Supplement), LB156.
16. Iyer, L. M., Makarova, K. S., Koonin, E. V., & Aravind, L. (2004). Comparative genomics of the FtsK–HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucleic acids research*, *32*(17), 5260-5279.
17. Burroughs, A., Iyer, L., & Aravind, L. (2007). Comparative genomics and evolutionary trajectories of viral ATP dependent DNA-packaging systems.
18. Moffitt, J. R., Chemla, Y. R., Aathavan, K., Grimes, S., Jardine, P. J., Anderson, D. L., & Bustamante, C. (2009). Intersubunit coordination in a homomeric ring ATPase. *Nature*, *457*(7228), 446-450.
19. Fuller D. N., Raymer D. M., Rickgauer J. P., Robertson R. M., Catalano C. E., Anderson D. L., Grimes S.& Smith D. E. (2007) 'Measurements of single DNA molecule packaging dynamics in bacteriophage λ reveal high forces, high motor processivity, and capsid transformations'. *Journal of Molecular Biology 373***,**1113-22.
20. Kalafut, B., & Visscher, K. (2008). An objective, model-independent method for detection of non-uniform steps in noisy signals. *Computer Physics Communications, 179*(10), 716-723.
21. Chistol, G., Liu, S., Hetherington, C.L., Moffitt, J.R., Grimes, S., Jardine, P.J. and Bustamante, C. (2012) High degree of coordination and division of labor among subunits in a homomeric ring ATPase. *Cell, 151*, 1017-1028.
22. Kottadiel, V. I., Rao, V. B., & Chemla, Y. R. (2012). The dynamic pause-unpackaging state, an off-translocation recovery state of a DNA packaging motor from bacteriophage T4. *Proceedings of the National Academy of Sciences, 109*(49), 20000-20005.
23. Klein, M., Andersson, M., Axner, O., & Fällman, E. (2007). Dual-trap technique for reduction of low-frequency noise in force measuring optical tweezers. *Applied optics*, *46*(3), 405-412.

Chapter 6
Development and Testing of Methods for Determining DNA Translocation Velocity and Pauses and Slips (and Exploration of A Novel Signal Processing Algorithm Based on Data Sorting)

Abstract

Time series data analysis is widely used to modeling dynamic behaviors in physics, biology, business, and engineering. A special pattern of time series data – monotonic increasing underlying pattern with a high level of added "noise" – is commonly to be seen in daily life but it attracts few attentions in studies of signal analysis or is mentioned in data mining research. In our work studying DNA translocating molecular motor we found a situation where sections of data (of length vs. time measurements) are presumed to contain a monotonically increasing signal (and approximately linear over a short time range) while other sections contained constant or decreasing signals. These signals also contained fairly high levels of instrumental measurement noise, in part caused by Brownian fluctuations. For this work it was important to both identify the different data segments and to determine the slopes of the linearly increasing regions. Here, the application of three segmentation methods of regulation identification we tested are discussed. A novel sorting-matched algorithm was also proposed to serve as a tool that could detect different regions where the signal is increasing or not, and to determine slope. A

general conclusion is that these methods may have advantages for segmenting regions of data, although we do not find that the slope determination method (applied to a defined segment of simulated data) is more accurate than a standard linear regression method.

6.1 Introduction

Time series data, a sequence of number mapping to its sampling time, occurs in many financial, engineering and industrial processes.[1-7] Biological examples may include growth of cell densities in an incubator, or in my own research, DNA length packaged in bacteriophage shell.[8-13] Time series data is commonly collected by electric sensors which detect events or changes, transform them into digital numbers and send information to other electrics.[14, 15] Nowadays, for example, an inexpensive positing sensing photodetector can achieve sampling rates of at least 5 kHz such that a 1-minute experiment yields 300,000 data points. Generally, the magnitude of data points of 1 event is ~10^5 -10^9 that brings a burden for devices when they are processed. On another hand, it is inconvenient to figure out a universal model of the observation if a great amount of data is analyzed at the same time. One feasible solution for those two problems is segmentation, which is an algorithm partitioning signal into multiple segments.[2, 16-21]

It is usually to be seen that a set of time series data or signal, over a short time scale and neglecting any measurement noise, behaves as a monotonic increasing function, for instance, the cumulative number of infection cases, the DNA length packaged into phage prohead shell, the distance from a subway train to its terminal and the number of files processed by a computer.[22-

26] In a short enough window, such a monotonic signal may also be approximated as a linearly changing signal. The sequential data we wish to consider has a characteristic that any number sampled later, if there were not noise, would not be less than the earlier one. Other practical examples might be a subway train pulling backward to align the doors with the platform or a computer merging similar files after processed. These local adjustments are called ‘a local regulation of the system’, which happen randomly, locally, in a short interval and do not reverse the monotonic increasing trend of time series data but constitute a regulation feature of the system. As result, local regulation in the research of system error should be considered as an unignorable property which was smoothed as noise.[30] Moreover, locally monotonic data segmentation and regression are heat topics of studies in Machine Learning optimization problem.[32]

In this chapter, we discuss the application of three different types of segmentation methods of regulation identification. A sorting-matched algorithm was proposed to separate out a region of monotonic increasing signal. Additionally, a relationship between matched probability and slope of signal is demonstrated to be a novel method to characterize the regulation.

6.2 Classification of Pauses and Slips

6.2.1 Terminology

6.2.1.1 Increasing Signal and Decreasing Signal

'Increasing Signal', as so called 'Ascending Signal', refers to a one-dimension data set that the values of data points have increasing (ascending) tendency as the same as its sampling order. The sampling order could be the order of time or the order of sample position, etc. We call those variables which mark the order samples as 'Sequence Stamp' or 'Stamp'. Vise verse, 'Decreasing Signal' or 'Descending Signal' refers to the 1D data set that the values of data points have decreasing (descending) tendency as the same as its sampling order.

'Monotonic Signal' is the advance version of increasing (deceasing) signal that describes all values in the dataset obeys the increasing (decreasing) rule. Mathematically, the monotonic increasing signal is defined as

$$\forall 0 < i < j \leq N, y_i \leq y_j$$

which means for all of i, j satisfied 0<i<j<N, the later inequation is always true.

In practice, data is collected by sensers that always result in at least some noise added to the signal. It is almost impossible to get a set of data strictly and rigorously satisfied the mathematical definition. To apply our method proposed below, there is another definition of this monotonic signal as

$$\forall x_i < x_j, \exists \varepsilon \; that \; \overline{y_i} \leq \overline{y_j}$$

where $\overline{y}$ is the mean of y over $(x - \varepsilon, x + \varepsilon)$.

In this chapter, all increasing signals or monotonic signals refers to the second type of definition.

6.2.1.2 Pause Signal

A ‘Pause Signal’ is what we call a sequence of data points that have the same values in an interval of stamps. The mathematical definition is

$$\forall x_1 \leq x \leq x_2, \left|\overline{y_i} - \overline{y_j}\right| < \varepsilon$$

where ε is a small enough number.

6.2.2 Simulation Experiment

All of the simulation experiments and the data analyses in this chapter were conducted on Matlab. The source signal was generated by simulated signal generator which created a 2.2 second ascending signal with slope = 1. The sampling rate of signal generator and receiver were both 1 kHz. In the simulation experiments regarding signals with pauses (zero slope regions), another 1 second length pause signal with the same sampling rate was inserted into the ascending signal at 1.2 second (Fig. 6.1A black solid line). A non-integer insertion position was used to introduce a phase shift to avoid the coherence of signal segmentation. A signal sampled by a ‘real world’ sensor was mimicked by adding Gaussian noise which had 0 mean value and σ^2 variance onto the source signal. The source signal and collected signal are plotted in Fig. 6.1A as blue solid line.

In the simulation experiment regarding signals with slips (negative slope regions), the 1 second length pause signal was replaced by a 1 second length decreasing signal with slope = -0.5 and a 1 kHz sample rate. (Fig. 6.2A. Black and blue solid line represented the source and collected signal, respectively.)

6.2.3 Sliding Window Algorithm

A standard sliding 'boxcar' window method, cropping the entire collected signal by the same window, has been proved an efficient method on 1D signal segmentation.[2] For some applications in our lab studying single DNA molecule translocation by molecular motors, we identified classes of those pieces of cropped signal by some criteria. It is a common method to segment collected data as was done in the most of our previous viral packaging experiment data analyses. In detail, the simulation signal was analyzed in a sliding 0.5 second window in which its slope was calculated by standard linear regression (using the Matlab *fitting* function with parameter *'poly1')*. The slope was considered as the DNA packaging rate or instantaneous motor speed within this window.

For the purposes of section classification, to distinguish a pause from translocation, slope (packaging rate) was compared with a threshold depending on noise level (since due to noise the slope will never be exactly zero even during a true pause). Those sections with slope greater than threshold were classified as 'packaging'. Otherwise, they would be treated as pauses or slips which depended on the type of simulation experiment. Fig. 6.1B and Fig. 6.2B showed the results of the signal classification of pause and slip using the sliding window method. The black

solid line was raw signal. The blue dots were sections of collected signal classified as packaging. The red dots represented sections of collected signal classified as non-packaging (pause or slips).

6.2.4 K-means Clustering Algorithm

K-means clustering is another method used on our data analyses. K-means is a method of grouping and labeling random distributed data points into 'K clusters', which was developed from signal processing and aiming to minimize the within-cluster variances (squared Euclidean distances) of each group. This algorithm results in a partitioning of the data space into Voronoi cells.[29] Since K-means is an unsupervised learning algorithm that it does not give any label to each cluster, K-means clustering result only reflects the nearest relationship among data points but not the consistence with users' target. Fig. 6.1C and Fig. 6.2C are results of clustering simulation signal by K-means algorithm of K=3. In fact, K-means is doing the following steps for clustering:

(1) Randomly picks K points in dataset and marks them as K different centers of clusters.

(2) Computes the squared distance from the centers of clusters to the rest of data points, i.e. $d^2 = (x - x_c)^2 + (y - y_c)^2$.

(3) Each data point is marked by the same label as its nearest center.

(4) Recompute the center of each cluster as the way of computing center of mass.

(5) Repeat step 2 to 4 until the centers of each cluster do NOT change.

In this chapter, the K-means algorithm was implemented by the Matlab *kmeans* function with parameter K = 3.

6.2.5 Maximum Distance Algorithm

Maximum distance is a novel method we propose and explore here to identify the target section in the collected data. The plots in Fig. 6.1D and Fig. 6.2D illustrate how it can find the target section containing a monotonic increasing signal. If one makes a line from the starting point of raw signal to the end point (red solid line), it can be proved that the turning points have the maximum distance to the line. Therefore, finding the farthest points to the line and cropping the collected signal at turning point becomes a feasible way to segment the experimental data. In practice, we took the starting point and the end points to be the mean of the first and the last 50 data points, respectively. Then distance from points to line is computed by $d = (kx - y + b)/\sqrt{k^2 + 1}$. Generally, a positive and a negative distance would be found from the calculation. The positive distance implies the data point is above the line and the negative point implies under. The non-increasing segment is between the point contributing the maximum distance and the point contributing the minimum distance.

6.2.6 Confusion Matrices

A so called ‘confusion matrix’, also known as an error matrix, is a 2-by-2 table which quantitatively describes the accuracy and performance of a classification algorithm. Each row of

the matrix represents the instances in an actual class, meanwhile, each column represents the instances in a predicted class (or vice versa).[30] 4 numbers in confusion matrix are the score (ratio) of true positive, false positive, false negative and true negative. True Positive, short as ‘TP’, represents the pause signal being classified as pause class; False Positive, short as ‘FP’, represents the increasing signal being classified as pause class; False Negative, short as ‘FN’, represents the pause signal being classified as increasing class; True Negative, short as ‘TN’, represents the increasing signal being classified as increasing class. The score of each class, which sum up to 1 for all methods, is the ratio of the count of that class vs. total number of data points. To test this method, I ran a simulation program, with simulated signals with noise added, 1000 times with changing parameters. The final scores are listed in Table 6.1 & 6.2 by averaging over 1000 simulations.

6.2.7 Accuracy

Confusion matrices of the simulation of pause/slipping are visualized in Fig. 6.3A and 6.3B. The blue, red and yellow bars indicate the score of segment classification by the methods of sliding window, maximum distance and K-means, respectively. The X-axis indicates what kind of score each of those bars refers to. In the comparison of target classification, TP, it is notable that the ‘maximum distance’ algorithm we proposed achieves higher score than other two algorithms for both pause and slipping simulations. Meanwhile, the ‘maximum distance’ method has lower error rate (score of FP and FN) than other two methods.

Sliding Window

The accuracy of the ‘sliding window’ method hugely depends on the threshold slope used to classify as target section. In the simulation experiment conducted in this chapter, a larger threshold would introduce to the algorithm a bias of classifying segment as pause/slipping. As result, it is expected to obtain a low score of false positive and a high score of false negative. And it is the opposite situation with a smaller threshold. On another hand, accuracy of sliding window method also depends on the position of windows. In this simulation, the window slides from x=0 and moves 0.5 s for each step. I chose to have no overlapping sections for these window sections. If a cropped signal contains both packaging part and pause/slipping, this cropped signal will probably lead to a wrong classification, false positive or false negative. For example, we make a 0.2 second shift in the first part of packaging signal that the cropped signal of 1.0-1.5 second and 2.0-2.5 second are mixed by packaging segment and target segment. The algorithm picks either class as the classification result but causes error. According to the simulation (Fig. 6.1B and 6.2B), the former signal would be classified as target section and the latter one would be treated as packaging (increasing).

K-means

The accuracies of the K-means method regarding FP and FN occurrences are better than other two methods. At the beginning, the label given to the non-packaging segment is determined without supervision by its initial label. It is possible to label a wrong section as target and cause an error. The length of the simulation data brings an error by change the weight of each component. Here three sections (pre-target, target, post-target) have similar length such that the

component on y-axis does not play a leading role. In some extreme situations, with super long target sections for instance, the classified pre-target and post-target sections contain a lot of data points from target.

Maximum Distance

The accuracy of maximum distance algorithm relies on the signal noise ratio (SNR) and data distribution. A large noise (small SNR) would make the neighboring data points of maximum or minimum distance point be mis-judged which would affect accuracy of finding turning points. On another hand, length of target section and slope of packaging section affects the absolute value of distance from turning points. Both could reduce the accuracy of maximum distance algorithm.

To have an overall comparison, we add the scores for true positive and true negative to get a measure of the "overall accuracy" of all three algorithms for analysis of the simulated data with pauses or slips. (Fig. 6.3C) The sliding window algorithm and the maximum distance algorithm have about 83% identification accuracy for the pause datasets and about 90% accuracy for the slipping datasets. Notably, the accuracy of K-means algorithm is much lower, 28% and 35% lower respectively, than other two methods for pause and slipping identification.

6.3 Methods for Data Cropping/Segmentation

In section 6.2, we discussed how to identify a pause/slip section within a piece of monotonic increasing signal. Simulation results shows the classification accuracy of three methods applied to simulated data with certain noise levels. However, there is an assumption of our simulation which is that the collected signal explicitly contains three sections: pre-target, target and post-target. In a general case, a signal that needs to be segmented could have different unpredictable properties, like containing multi-pause and multi-slipping sections, transition of slope of increasing section, etc., which would introduce further complications and potential problems and errors to data processing. In this section, we propose a cropping method to divide the collected signal into a series of simplified signals carrying only monotonic increasing signal and only one piece of target section as we demonstrated in section 6.2.

6.3.1 Reorder Datapoints

A piece of monotonic increasing signal usually has the same order as its sampling monotonic increasing reference stamp (commonly time, which can often be measured very accurately with very little error/noise).[26, 31, 32] There is a one-to-one mapping relationship between the sampling signal and its time stamp. We have explored a method in which sorting of datapoints for an underlying monotonic increasing signal is applied to analyze experimental data on phage DNA packaging, of the kind discussed in the other chapters of this thesis. A sorting method is adapted to help us finding different sections within the data. All data sorting was conducted by using the Matlab *sort* function with the ‘ascending’ parameter option.

6.3.2 Matching Probability After Sorting Method

We propose a novel method in which we compared the raw sequence data and the sorted data and then counted how many data points in the entire sequence did not change in value after being sorted. These points unchanged were called ‘matched-points’ or ‘fixed-points’. We demonstrate below that calculating the probability of matching is related to slope when the underlying signal of that data section is linearly increasing and can be used for both data segmentation and slope determination. Fig. 6.4 demonstrated the distribution of the match-points of three simulated signals with different modes. Fig. 6.4A was a copy of simulated pause signal in Fig. 6.1. Fig. 6.4B was a larger noise version of A that the amplitude of noise was doubled. Fig. 6.4C was a copy of simulated slipping signal in Fig. 6.2. The red solid line was the sorted signal, and the red circles were the matched points. Empirically, matched frequency and matched-point distribution were associated with the behavior of the underlying signal in different sections of the dataset. The matched-point density within increasing sections was higher than within sections where the underlying signal is a pause (constant signal) or slipping (decreasing signal) section. Although we show that this may be useful for detection of signal behavior when there is added noise, the approach is limited by the fact that signals with more noise were more likely to have less matched density. To quantitatively characterize the match frequency, we defined a few variables below:

k: Slope of monotonic increasing signal

SR: Sampling rate

i: integer index number, from 1 to N

x: sequence of reference stamp, from 1/SR to N/SR

y_0: collection of monotonic increasing signal

y: sampled signal, y = y_0 + noise

y_{sort}: sorted signal of y

noise: white noise, following Gaussian distribution with mean of 0 and variance σ^2.

According to the definition of increasing signal, there is a relationship of y_0:

$$y_0[i-1] < y_0[i] < y_0[i+1]$$

$$prob(y_0[i-1] < y_0[i] < y_0[i+1]) = 1$$

However, this relation is not always true for a signal with added noise. Fortunately, we still have similar relationship on the matched-points, which is stated as 'the i-th data point becomes a matched-point is due to there are (i-1) data points not larger than it'. In theory, the event of the i-th being a matched-point is

$$\sum_{j=1}^{N}(y[j] < y[i]) < i$$

$$\sum_{j=1}^{N}(y[j] \leq y[i]) \geq i$$

And the probability of this event is

$$prob(y[i]\ is\ matched\_point) = prob(y_0[i] + noise[i] == y_{sort}[i])$$

Combined with the properties of the signal, one could derive the probability of $y[j] < y[i]$ is

$$prob(y[j] < y[i]) = prob(y_0[j] + noise[j] < y_0[i] + noise[i])$$

$$= prob(y_0[j] - y_0[i] < noise[i] - noise[j])$$

For a linear raw signal, $y_0[j] - y_0[i] = k \times (j - i)/SR$. The difference of two pure Gaussian noise signals is also Gaussian noise but the variance is doubled. Hence,

$$prob(y[j] < y[i]) = \Omega(\frac{k \times (j - i)}{SR \times 2\sigma^2})$$

where Ω(x) is the cumulative distribution function of normal distribution. The probability that a point becomes a matched-point, which we call 'matched probability', is

$$p = prob\left(\left[\sum_{j=1}^{N}(y[j] < y[i])\right] < i\right) = prob\left[\sum_{j=1}^{N}\Omega(\frac{k \times (j - i)}{SR \times 2\sigma^2}) < i\right]$$

This result implies the probability is a function of signal slope, sampling rate, and noise amplitude.

6.3.3 Simulations

Simulation experiments have been conducted to investigate the relationship among slope k, sampling rate SR and noise parameter σ^2. A piece of simulated source signal of 100 seconds, slope k=50 and 1 kHz sampling rate was generated. Gaussian noise of $\sigma = 0.1$ was added to mimic a real-world recorded signal. This signal was sorted then checked for matched points. The matched-probability was computed by dividing counts of matched-points by 100,000 (100 sec * 1kHz). We re-ran the simulation program with changing the slope and the amplitude of noise then plotted the results in Fig. 6.5A. These results show how determination of matched-probability could be used to estimate slope, and the effect of slope magnitude and noise level on

this estimate. An adapted normal distribution CDF and an adapted sigmoid function were employed to fit this simulated result and shown on Fig. 6.5B and 6.5C, respectively, which could then be used as interpolation functions to deduce slope from matched-probability.

6.4 Negative Probability Hypothesis

6.4.1 Introduction to Negative Probability

In the exploration of the proposed sort and matched probability analysis method discussed above, a potential application of the concept of "negative probability" arose, which we first review briefly here. Probability, as a daily concept, has been introduced to help making decision reasonably in our lives, but few of us interpretate what probabilities are.[38] The orthodox probability was proposed by Andrei Kolmogorv in 1933,[39] which stated that the probabilities of events must be **positive real numbers**. Here we would like to focus on the *frequency interpretation* approach, even though there are a few interpretations of traditional probability, such as *the belief interpretation*, the *support interpretation*, the *logical interpretation* and the *propensity interpretation*.[40] In the *frequency interpretation* approach, the probability of an event is the frequency of its occurrences in an infinite identical repeat observation or experiment. However, occurrences of the observing event and the entire set of events are included based on human realization and long-term experience which has spatial and temporal limitation due to insufficiency of event recognitions. Usually, we call such an unrecognizable event as 'hidden event'. Hidden event is a concept developed from an intermediate stage which was introduced in R. Feynman's probability problem in 1987 [41] that the intermediary step

carried a negative probability and can leads to an interesting interpretation of a physical problem. The interpretation of negative probability was developed and applied in quantum theory and computer science.[42-46] The idea of hidden event (also referring to hidden variable) was widely employed to describe the indetectable events, although the definition of hidden event was unclear. One example is a system comprising 2 bit registers designed by S. Abramsky and A. Brandenburger [47, 48] that the (0,0) event with p(00) = -1/2 can never be observed. The two bit registers experiment briefly illustrated the negative probability theory and application but cannot reveal the universal mechanism of negative probability on: (1) the experiment cannot be extended to a general case; (2) the final probability is equal to 1 that their theory is limited by the entire set of events has to be observable.

In 6.4.2, we would like to propose a scenario of negative probability derivation and its application, as orthodox probability, in which the hidden event (hidden variable) is explicitly recorded in dataset. The exact definition of hidden event is also defined. In 6.4.3, the difference between hidden event versus unobservable event or inactivate event has been identified.

6.4.2 Example of Negative Frequency

In the sorted-match frequency simulation, the frequency of target event (matched-points) was a variable obeying the probability distribution function (PDF) related to generator parameters. It has been proved the match frequency, which defined the match probability, was able to be computed by traditional statistical theory. (Fig. 6.5) Besides, this distribution can be fitting by a CDF on the positive axis side. A normal distribution CDF and sigmoid function are widely used on characterizing the cumulative probability from 0 to 1. They have 2 features:

(1) There are 2 asymptotes, y=0 and y=1;

(2) The function is rotational symmetry about (0, 0.5).

A curious question is what conclusion could be derived if the slope of monotonic increasing signal changes to negative value? To investigate this, we repeat the simulation experiment but vary the slope from -1000 to 1000, and also define the match-probability of descent sorting to be negative. The simulation result and fitting are plotted in Fig. 6.6. The datapoints on the negative axis represent a simulated monotonic decreasing signal which is an opposite trend signal as in the previous simulation experiments. This decreasing signal was 'out' of our recognition because we neither expected it nor designed a sensor (sorted-match algorithm) to detect it. However, it could be easily re-identified by just inverting the detector (invert the axis). That is why the matched frequency obtained by inverting detection algorithm was given a negative sign. Furthermore, the data points can be fitting by a continuous differentiable function as shown in Fig. 6.6.

6.4.3 Negative Probability Interpretation

In the simulation experiments described above, a hidden event which was generated by changing the sign of signal slope could not be detected by the 'previous' detection algorithm. Discarding the value of slope, this monotonic decreasing signal was identified as 0 match frequency by the ascending sort method. However, if a piece of decreasing signal was inserted into an increasing signal as an intermediate step, the system went thru the hidden stage and the final stage would be detectable even though this intermediate stage was still hidden.

On another side, the inverted detection illustrates a solution to uncover the hidden event and to obtain non-zero match frequency. This match frequency was identical to the positive one except has a negative sign. According to this result, we propose a negative probability hypothesis: the negative probability describing the hidden event was the frequency of an event that was the exact opposite of the event defined for observation. For instance, going up is the opposite event of going down. The hidden event associated to negative probability is beyond the original expectations for features in the signal such would not be included into the whole set of events. Although a hidden event is unable to be measured with the original considerations, one can predict its existence by identifying a locally abnormal behavior from a number of repeat experiment or observations.

6.5 Application

A novel classification and segmentation method is developed in Section 6.2 and 6.3, applied to the type of DNA translocation studies described in the other chapters. Here we are planning to estimate viral motor velocity from noisy DNA length packaged vs. time measurements by applying the new method. One set of phi29 packaging trace records is used to get the slope (motor velocity) and matched frequency by the sliding window method. We plot motor velocity vs. matched frequency in Fig. 6.7. The reason for using phi29 packaging data is because the motor velocity decreases while filling level increases. It offers larger range of motor velocities for fitting. When the noise amplitude is constant, there is a linear relationship between motor velocities of phi29 and the matched probability ($R^2 = 0.87$). It implies a potential

application of the novel classification and segmentation method for computing the slope of signal. But we find that it works only when a segment of data contains > 2000 points. Moreover, it is easily to obtain the slope of one cropped signal and the matched probability. One could estimate noise amplitude of the target signal by finding the point on a slope-probability map.

6.6 Discussion

Sliding Window

In this study, three methods of segmentation of pause/slipping data -- sliding window, K-means and maximum distance -- have been discussed. A common but easily implemented method -- **sliding window**, which was the most widely used packaging data processing method, is found to be limited on its performance by its window size, threshold criteria and phase shift.

Window Size: one should trade off the need of enough data points to improve the accuracy of linear fitting vs. the number of segments designated as having different features if applying the method to analyze the dynamics of the viral motors. In addition, one should choose a smaller window size to separate the target section from background.

Threshold: accuracy of classification dramatically relies on the setup of threshold slope. The same threshold setting could make a difference while processing two similar datasets. Empirically, the parameters can be changed manually to get the optimal classification result.

Phase Shift: it has about the same drawback of window size that it can lead to errors when a segment contains multiple classes of signal change. Besides reducing the length of window size, one possible improvement is reducing the step size to identify and vote the same

area many times. We slide a window but we overlap the sliding window that one data point appears in multiple sliding windows. This datapoint would be classified many times then we pick the most voted one. Then it maximizes the likelihood based on the vote.

K-means, a traditional mature machine learning algorithm, was one of options to improve the signal processing. It has been implemented in many coding languages and data analysis software as support packages. K-means is advantageous for quickly and accurately grouping spatially distributed (like tribes) unlabeled data.[29] However, it does not meet our expectation when it clusters the sequential data, for instance the viral packaging data with significant pausing and slipping. One should perform a data normalization before applying the K-means on packaging signal clustering because it should be guaranteed that the weights of all dimensions of vectors are the same.[49]

Another point to be noticed is the K-means algorithm is an unsupervised learning algorithm which labels the clusters only by the order it encounters them. It is better for researchers who use this method on viral packaging data analysis to add an extra fitting step to classify the sections it crops. Another issue to consider is that it uses significant computational power which could be exhausted if a huge amount of data needs to be analyzed.

Maximum distance

Maximum distance is more suitable for packaging data segmentation under low noise conditions. Compared with the sliding window method which classifies the signal piece by piece, the maximum distance method finds the beginning and the end of the target section (pause or slipping) which is advantageous to avoid mis-classification compared to the sliding window

method in long target sections. In addition, the sliding window method requires one to develop a separate algorithm to merge classified sections of different windows, while the maximum distance method is simpler because it does not require this. Compared with the K-mean method which randomly assigns numbers to groups by datapoints, in this method there is a certain criterion (which can be the match-probability) to automatically judge the classes of cropped sections. However, the accuracy/error of this method is related to the quality of signal, i.e., depending on the amplitude of noise. Using a noise reduction procedure, such as smoothing/decimation, could help to minimize the error.

Sorting-Matched Algorithm

The last set of experiments done with simulated data in this chapter is the exploration of a novel sorting-matched algorithm and its application. The idea of matched-points comes from an observation of the unchanged data points while transforming and it is a relatively good solution to divide sequential data. In our case, a sorting transformation is applied then the fixed points are proved to be associated with the parameters of signal. As result, any of the parameters associated with the matched-probability could be estimated once the rest of them are acquired. Noise or signal noise ratio, as a universal unknown feature, therefore, can be computed.

We further propose that negative probability theory is an interesting point of view that could be applied with this method. Generally, probability, which is a theoretical number of predicting the frequency of an event, is always defined in between 0 to 1. This theoretical number could be verified by running repeatable experiments to compute the ratio of count of the target event observations over the total observations. But the unobservable event (the event

beyond the original expectations) cannot be evaluated in the traditional probability theory. Negative probability theory could provide a solution to this problem. The probability in between 0 to 1 is the same representation as usually, whereas the value in between -1 to 0 would describe the probability of the opposite event. In the initially applied method, both an impossible event and unobservable event are characterized as 0% probability, but they could now be separated. The impossible event has 0% probability, but the unobservable event (opposite event) is characterized by a probability of negative value.

Table 6.1 Confusion Matrix of Classification of Pause.

	Method	True Positive	False Positive	False Negative	True Negative	Accuracy
SNR=10	Sliding Window	0.1583	0.0099	0.1542	0.6776	0.8359
	Maximum Distance	0.2668	0.1279	0.0457	0.5596	0.8264
	K-means	0.1024	0.2308	0.2101	0.4567	0.5591
SNR=1	Sliding Window	0.1513	0.0501	0.1612	0.6374	0.7887
	Maximum Distance	0.1684	0.1301	0.1441	0.5574	0.7258
	K-means	0.1050	0.2323	0.2075	0.4552	0.5602
SNR=1, Double	Sliding Window	0.2847	0.0389	0.1912	0.4845	0.7692
	Maximum Distance	0.3041	0.0825	0.1717	0.4409	0.745
	K-means	0.1577	0.1801	0.3181	0.3433	0.501
SNR=1, Half	Sliding Window	0.0419	0.0613	0.1431	0.7526	0.7945
	Maximum Distance	0.0741	0.1496	0.1109	0.6643	0.7384
	K-means	0.0615	0.2709	0.1235	0.543	0.6045

Table 6.2 Confusion Matrix of Classification of Slipping.

	Method	True Positive	False Positive	False Negative	True Negative	Accuracy
SNR=10	Sliding Window	0.2477	0.0615	0.0648	0.626	0.8737
	Maximum Distance	0.2948	0.0812	0.0177	0.6063	0.9011
	K-means	0.0992	0.2337	0.2133	0.4538	0.553
SNR=1	Sliding Window	0.2347	0.0758	0.0778	0.6117	0.8464
	Maximum Distance	0.2306	0.1379	0.0819	0.5496	0.7802
	K-means	0.1018	0.2350	0.2107	0.4525	0.5543
SNR=1, Double	Sliding Window	0.4179	0.0513	0.0579	0.4722	0.8901
	Maximum Distance	0.4208	0.0574	0.0551	0.4661	0.8869
	K-means	0.1685	0.1806	0.3073	0.3428	0.5113
SNR=1, Half	Sliding Window	0.0915	0.0906	0.0934	0.7233	0.8148
	Maximum Distance	0.0958	0.1585	0.0892	0.6554	0.7512
	K-means	0.0646	0.275	0.1204	0.5389	0.6035

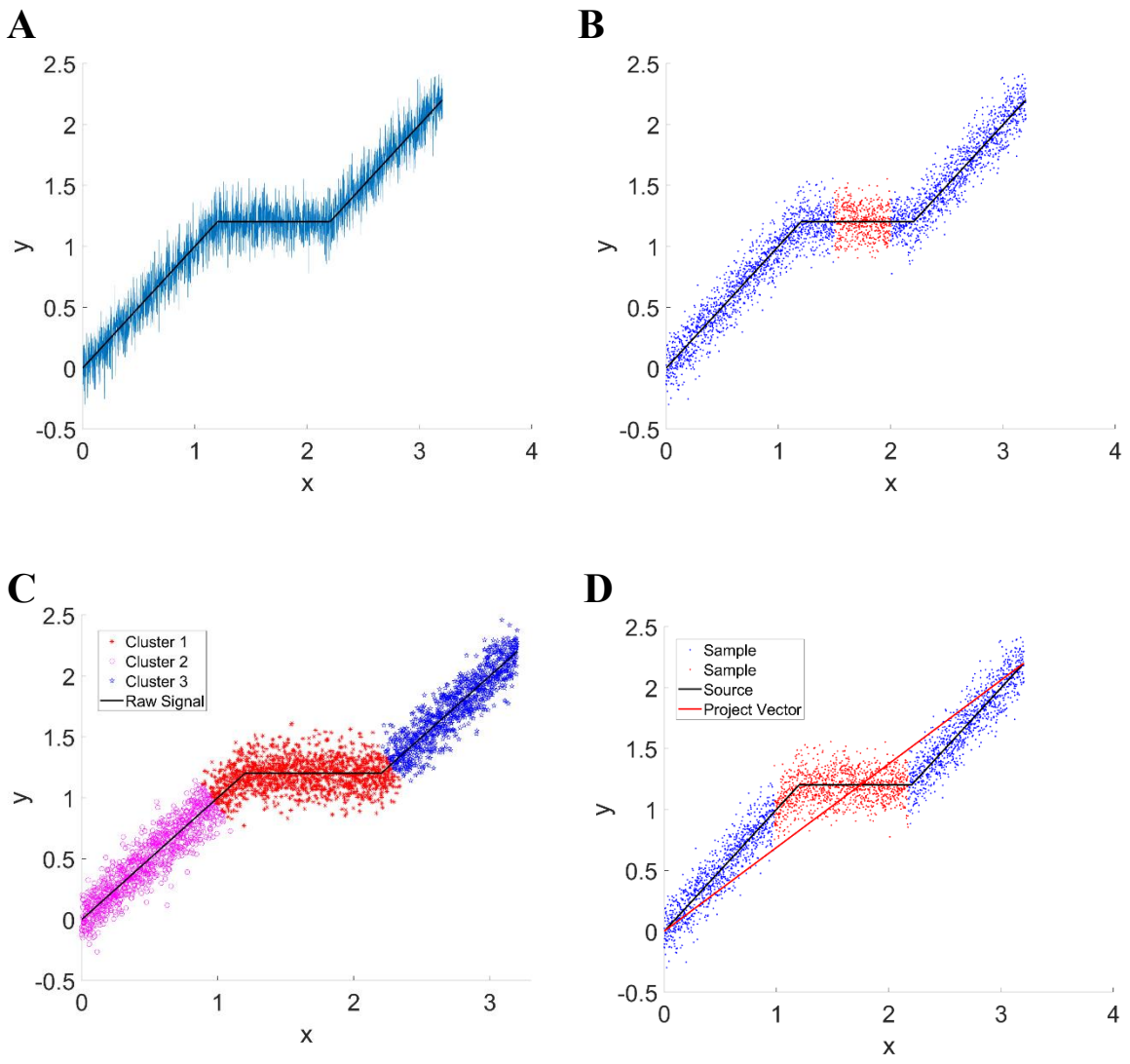

Figure 6.1 A simulated increasing signal with a single pause, with added noise, and three methods to identify the pause section. (A) Illustration of a monotonic increasing signal with 1 second length pause. Sampling rate equals 1 kHz, slope of increasing section equals 1, SNR = 20, phase shift = 0.2 second. The black solid line represents the original signal and the blue line represents the simulated sampling data with noise. **(B)** Identification of the pause section by a sliding window method, window size = 0.5 second. The blue dots are the data points classified as increasing section and the red dots are classified as pause. **(C)** Identification of the pause section by a K-means cluster method with two features. The data points are clustered into 3 groups based on their spatial position. **(D)** Identification of the pause by the Maximum Distance method. The black solid line is the raw signal and the red solid line is the projection vector. The blue dots are the data points classified as increasing section and the red dots are classified as pause.

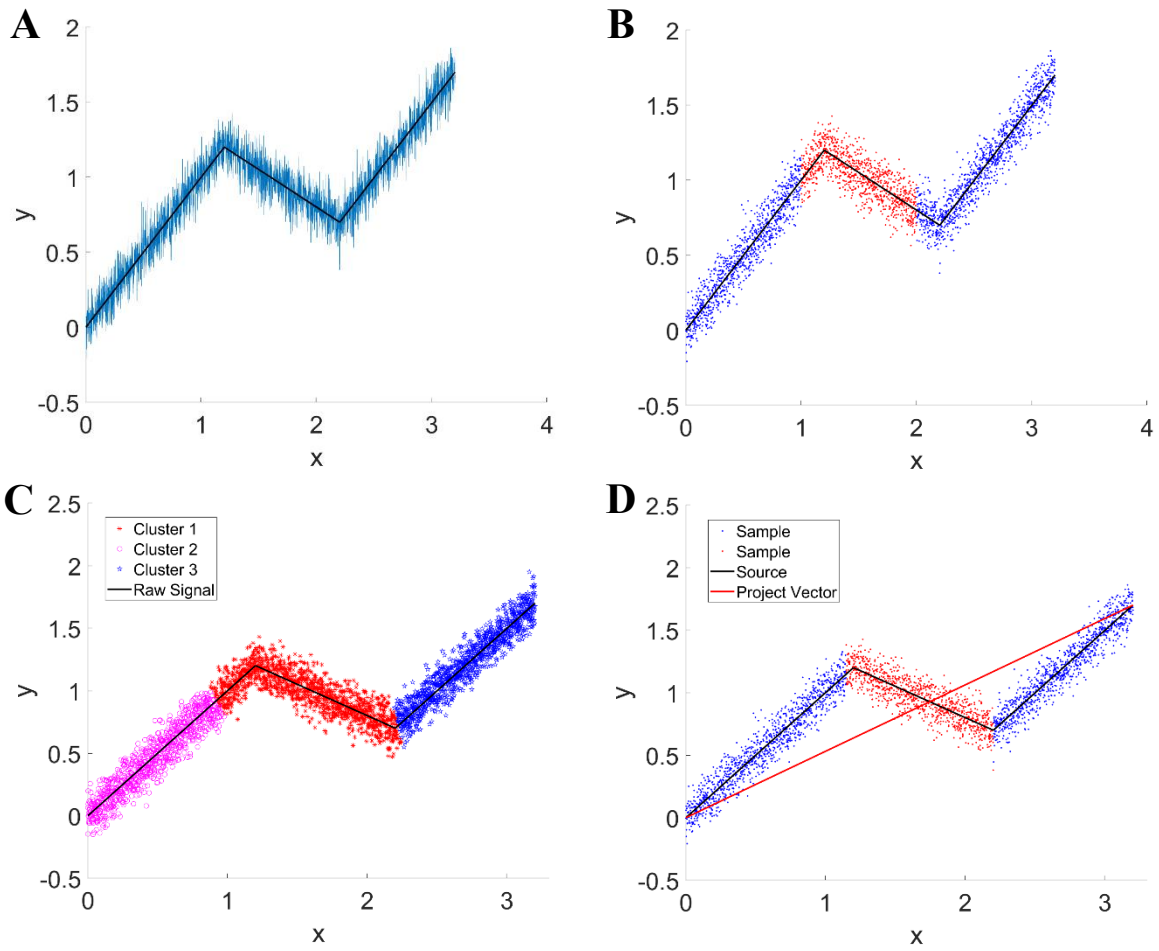

Figure 6.2 A simulated increasing signal with a single decreasing section (slip) and three methods to identify the slipping section. (A) Illustration of a monotonic increasing signal with 1 second length slipping. Sampling rate equals 1 kHz, the slope of increasing section equals 1, the slope of slipping equals -0.5, SNR = 20, phase shift = 0.2 second. The black solid line represents the original signal and the blue line represents the simulated sampling data with noise. **(B)** Identification of the slipping section by a sliding window method, window size = 0.5 second. The blue dots are the data points classified as increasing section and the red dots are classified as slip. **(C)** Identification of the slipping section by a K-means cluster method with two features. The data points are clustered into three groups based on their spatial position. **(D)** Identification of the slipping section by a Maximum Distance method. The black solid line is the raw signal and the red solid line is the projection vector. The blue dots are the data points classified as increasing section and the red dots are classified as slip.

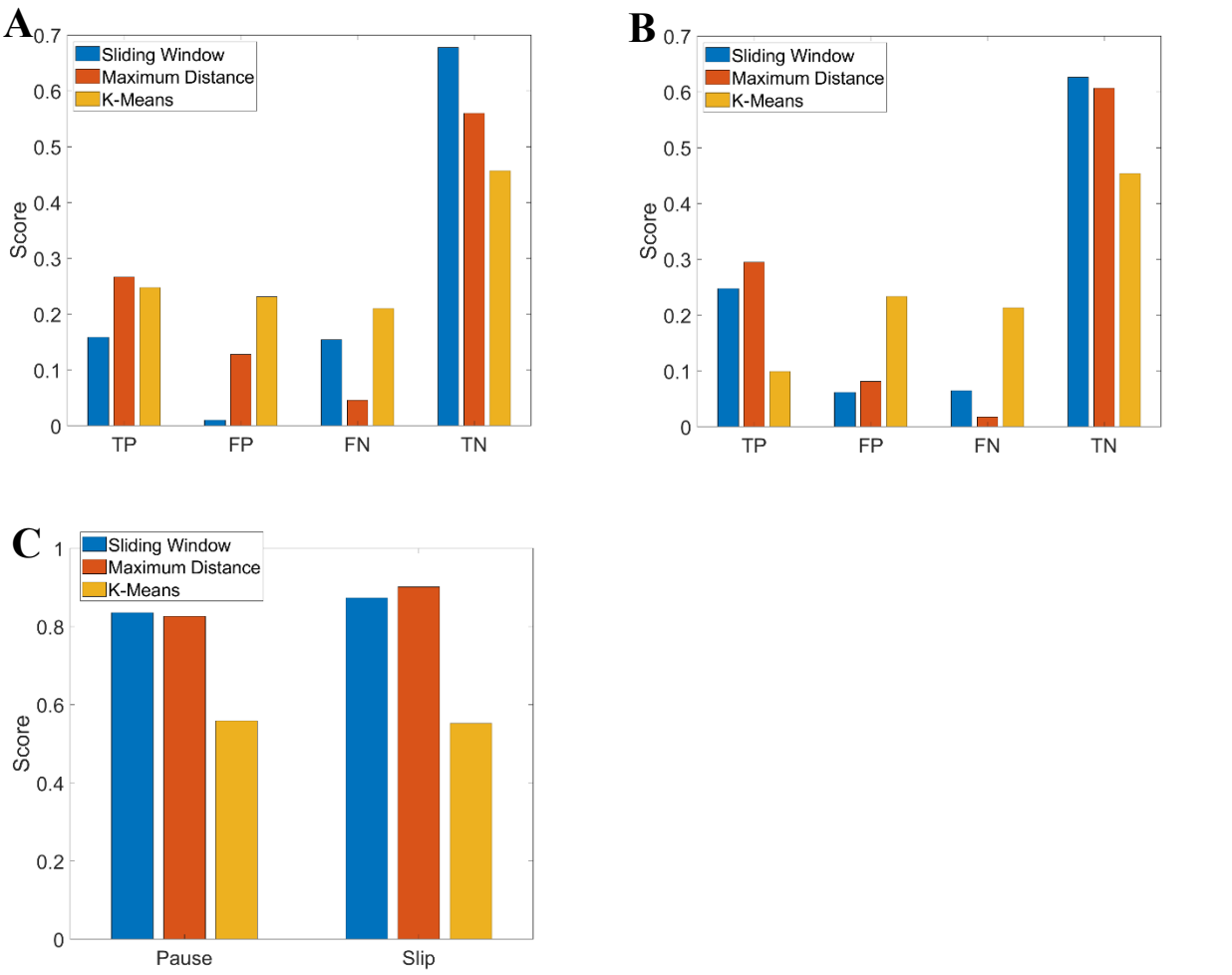

Figure 6.3 Statistics of classification accuracy of different section finding methods. (A) Confusion matrices comparison of different methods for identifying a pause signal. 'TP', short for True Positive, represents the pause signal being classified as pause class; 'FP', short for False Positive, represents the increasing signal being classified as pause class; 'FN', short for False Negative, represents the pause signal being classified as increasing class; 'TN', short for True Negative, represents the increasing signal being classified as increasing class. The score of each class, which sum up to 1 for each method, is the ratio of the count of that class vs. total number of data points. The final score is averaged over 1000 simulations. **(B)** Confusion matrices comparison of different finding methods for detecting a slipping segment. The content's meaning and the number of simulations are the same as for panel A. **(C)** Overall accuracy of each method for pause/slipping section classification. The score is a sum of the True Positive score and True Negative score.

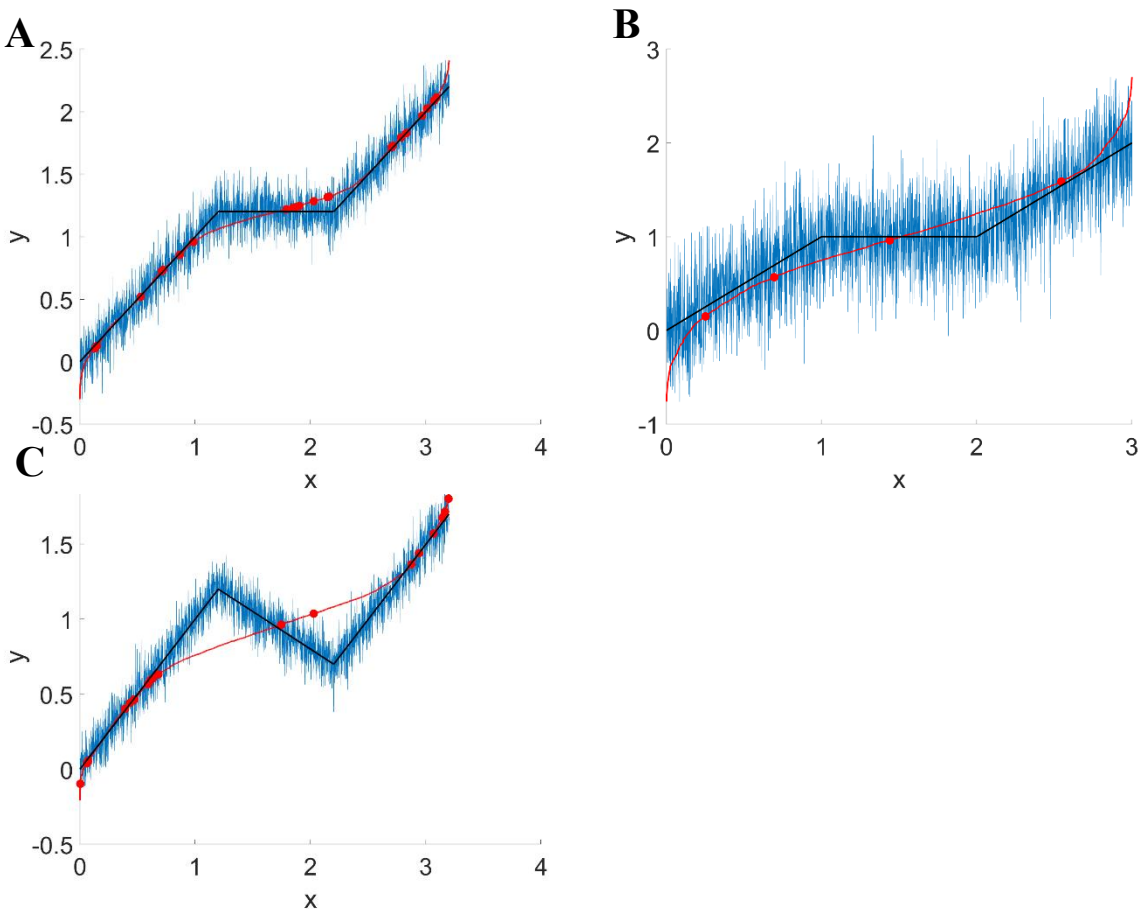

Figure 6.4 A sorting-match algorithm for finding suspect section. (A) Matched points found in a simulated noisy signal with a pause are shown in Fig. 6.1. The red solid line is the sorted data and the red circle dots is the matched points. **(B)** Matched probability reduces to 20% when SNR reduces to 10. **(C)** Matched points found in a noisy simulated signal with a slip are shown in Fig. 6.2.

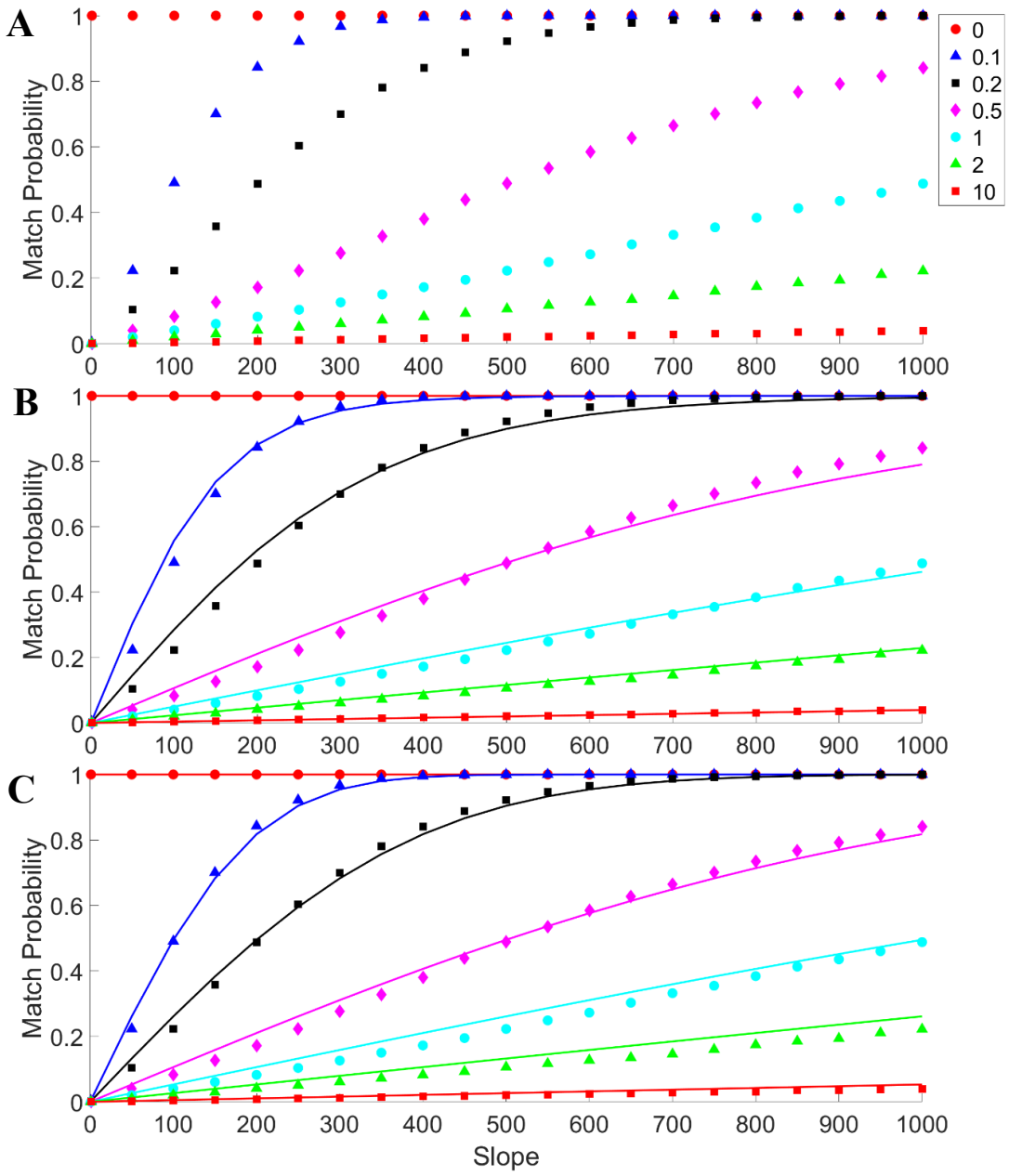

Figure 6.5 Total matched probability relates to the parameters of the simulated monotonically increasing signal. (A) Matched probability varies in response to changing the slope of source signal and the amplitude of noise. Sampling rate is fixed at 1 kHz. X-axis represents is the value of slope and Y-axis is the matched probability. The noise amplitude is indicated by different colors as indicated in the plot legend on the top right. **(B)** Fitting of the matched probability points by a Gaussian cumulative distribution function, $\boldsymbol{prob} = \boldsymbol{CDF}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\sigma^2})$, where the variable x is slope over sampling rate, mu equals 0 and sigma equals 1.5 times of noise amplitude. **(C)** Fitting of the the matched probability points by sigmoid function, $\boldsymbol{prob} = [\mathbf{1} + \mathbf{exp}\left(-\frac{\boldsymbol{k}}{\boldsymbol{SR}\times\boldsymbol{a}}\right)]^{-\mathbf{1}}$, where k is the slope of signal, SR is sampling rate and a is the amplitude of noise.

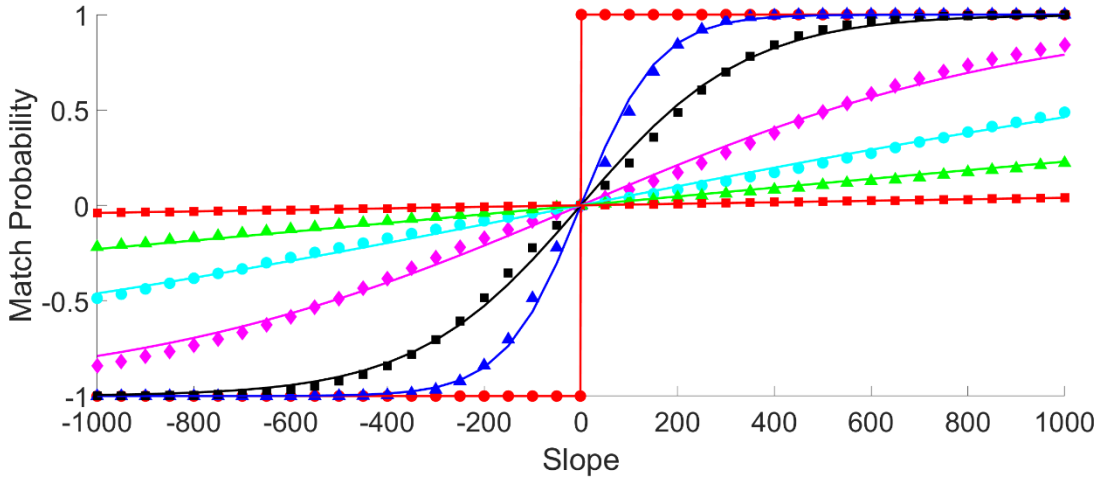

Figure 6.6 Matching probabilities for a simulated increasing and decreasing signal can be related to slope by a sigmoid function. The data points of slope > 0 are the same as in Fig. 6.5. The data points of slope < 0 are computed by sorting the decreasing signal by descending order then matching the original signal. Here we define the match probability to be negative values because we sorted by descending order against ascending order. However, the sigmoid fitting function for fitting both slope > 0 and slope < 0 sections is $\boldsymbol{prob} = [\mathbf{1} + \mathbf{exp}\left(-\frac{\boldsymbol{k}}{\boldsymbol{SR}\times\boldsymbol{a}}\right)]^{-\mathbf{1}}$.

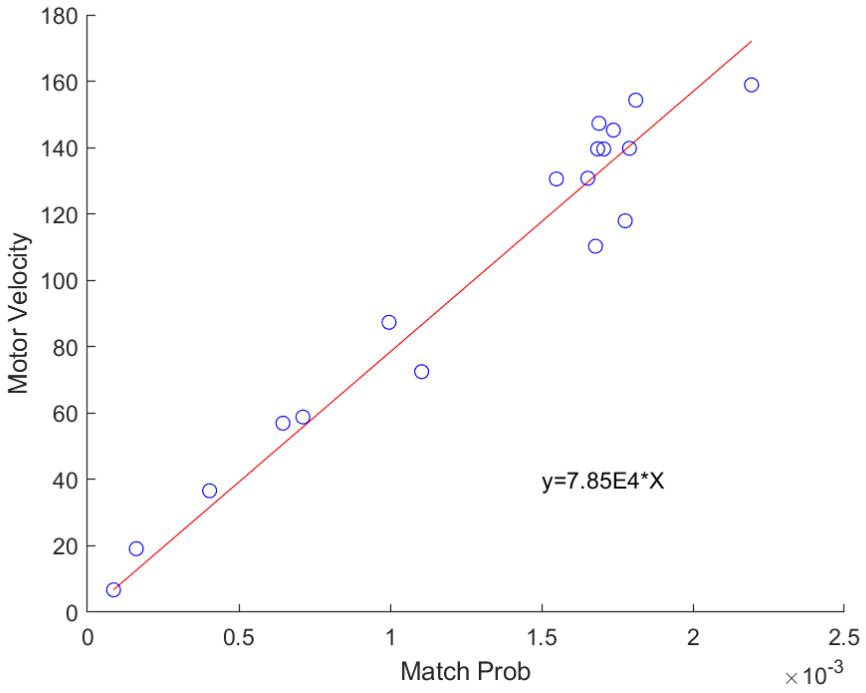

Figure 6.7 Applying the match probability method on computing the slope of experimental data. The system noise is assumed to be white noise and the amplitude of noise is constant. A 2-second window slides over the packaging trace to compute the motor velocities and match probabilities. Blue circles are the matching probability vs. slope of linear fitting of the data points in the same sliding window. The red line is a linear fitting of blue circles.

References

1. Galati, D. G., & Simaan, M. A. (2006). Automatic decomposition of time series into step, ramp, and impulse primitives. *Pattern Recognition*, 39(11), 2166-2174.
2. Hubert, P., Padovese, L., & Stern, J. M. (2018). A sequential algorithm for signal segmentation. *Entropy*, 20(1), 55.
3. Han, J., Gong, W., & Yin, Y. (1998, August). Mining Segment-Wise Periodic Patterns in Time-Related Databases. *In KDD* (Vol. 98, pp. 214-218).
4. Härdle, W., Lütkepohl, H., & Chen, R. (1997). A review of nonparametric time series analysis. *International Statistical Review*, 65(1), 49-72.
5. Fu, T. C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1), 164-181.
6. Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, 53, 16-38.
7. Li, M. (2010). Fractal time series—a tutorial review. *Mathematical Problems in Engineering*, 2010.
8. Jiang, D., Pei, J., & Zhang, A. (2003, March). DHC: a density-based hierarchical clustering method for time series gene expression data. In *Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings*. (pp. 393-400). IEEE.
9. Vilar, J. A., Alonso, A. M., & Vilar, J. M. (2010). Non-linear time series clustering based on non-parametric forecast densities. *Computational Statistics & Data Analysis*, 54(11), 2850-2865.
10. Okedi, T. I., & Fisher, A. C. (2021). Time series analysis and Long Short-Term Memory (LSTM) network prediction of BPV current density. *Energy & Environmental Science*.
11. Hanhijärvi, K. J., Ziedaite, G., Bamford, D. H., Hæggström, E., & Poranen, M. M. (2017). Single-molecule measurements of viral ssRNA packaging. *RNA*, 23(1), 119-129.
12. Roux, S., Chan, L. K., Egan, R., Malmstrom, R. R., McMahon, K. D., & Sullivan, M. B. (2017). Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nature communications*, 8(1), 1-12.
13. Keogh, E. J., & Pazzani, M. J. (1998, August). An Enhanced Representation of Time Series Which Allows Fast and Accurate Classification, Clustering and Relevance Feedback. In *Kdd* (Vol. 98, pp. 239-243).
14. Bennett, S. (1993). A history of control engineering, 1930-1955 (No. 47). *IET*.
15. Ko, M., Mendecki, L., Eagleton, A. M., Durbin, C. G., Stolz, R. M., Meng, Z., & Mirica, K. A. (2020). Employing conductive metal–organic frameworks for voltammetric detection of neurochemicals. *Journal of the American Chemical Society*, 142(27), 11717-11733.

16. Grützmacher, F., Beichler, B., Hein, A., Kirste, T., & Haubelt, C. (2018). Time and memory efficient online piecewise linear approximation of sensor signals. *Sensors*, 18(6), 1672.

17. Kim, S., & Choi, J. H. (2019). Convolutional neural network for gear fault diagnosis based on signal segmentation approach. *Structural Health Monitoring*, 18(5-6), 1401-1415.

18. Guo, Y., Liu, Y., Georgiou, T., & Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7(2), 87-93.

19. Moccia, S., De Momi, E., El Hadji, S., & Mattos, L. S. (2018). Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics. *Computer methods and programs in biomedicine*, 158, 71-91.

20. Witkin, A. P. (1987). Scale-space filtering. *In Readings in Computer Vision* (pp. 329-332). Morgan Kaufmann.

21. Witkin, A. (1984, March). Scale-space filtering: A new approach to multi-scale description. In ICASSP'84. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 9, pp. 150-153). IEEE.

22. Olito, C., White, C. R., Marshall, D. J., & Barneche, D. R. (2017). Estimating monotonic rates from biological data using local linear regression. *Journal of Experimental Biology*, 220(5), 759-764.

23. Güçlü, Y. S. (2018). Alternative trend analysis: half time series methodology. *Water Resources Management*, 32(7), 2489-2504.

24. Asfaw, A., Simane, B., Hassen, A., & Bantider, A. (2018). Variability and time series trend analysis of rainfall and temperature in northcentral Ethiopia: A case study in Woleka sub-basin. *Weather and climate extremes*, 19, 29-41.

25. Susto, G. A., Schirru, A., Pampuri, S., Beghi, A., & De Nicolao, G. (2018). A hidden-Gamma model-based filtering and prediction approach for monotonic health factors in manufacturing. *Control Engineering Practice*, 74, 84-94.

26. Gibson, G. M., Leach, J., Keen, S., Wright, A. J., & Padgett, M. J. (2008). Measuring the accuracy of particle position and force in optical tweezers using high-speed video microscopy. *Optics express*, 16(19), 14561-14570.

27. Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H. & Chawla, N. V. (2019, July). A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 1409-1416).

28. Harguess, J., & Aggarwal, J. K. (2009, November). Semantic labeling of track events using time series segmentation and shape analysis. In *2009 16th IEEE International Conference on Image Processing (ICIP)* (pp. 4317-4320). IEEE.

29. https://en.wikipedia.org/wiki/K-means_clustering

30. Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, *9*(1), 40-50.

31. YE, F., & LUO, J. Q. (2005). A multi-parameter synthetic signal sorting algorithm based on BFSN clustering [J]. *Radar & Ecm*, *2*.

32. Ankerst, M., Breunig, M., Kriegel, H. P., Ng, R., & Sander, J. (2008). Ordering points to identify the clustering structure. In *Proc. ACM SIGMOD* (Vol. 99).

33. Palacios, A. R., & Bovik, A. C. (1994). On the statistical optimality of locally monotonic regression. *IEEE transactions on signal processing*, *42*(6), 1548-1550.

34. Feynman, R. P. (1987). Negative probability. *Quantum implications: essays in honour of David Bohm*, 235-248.

35. Bartlett, M. S. (1945, June). Negative probability. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 41, No. 1, pp. 71-73). Cambridge University Press.

36. Scully, M. O., Walther, H., & Schleich, W. (1994). Feynman's approach to negative probability in quantum mechanics. *Physical Review A*, *49*(3), 1562.

37. Harris, A. J., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition*, *110*(1), 51-64.

38. Lyon, A. (2010). Philosophy of probability. *Philosophies of the Sciences: a guide*, 92-125.

39. Kolmogorov, A. N. (1933). Foundations of Probability. (1950) Chelsea Publishing Company.

40. Burgin, M. (2010). Interpretations of negative probabilities. *arXiv preprint arXiv:1008.1287*.

41. Feynman, R. P. (1987). Negative probability. *Quantum implications: essays in honour of David Bohm*, 235-248.

42. Abramsky, S., Brandenburger, A., & Savochkin, A. (2014). No-signalling is equivalent to free choice of measurements. *arXiv preprint arXiv:1412.8523*.

43. Bartlett, M. S. (1945, June). Negative probability. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 41, No. 1, pp. 71-73). Cambridge University Press.

44. de Barros, J. A., Oas, G., & Suppes, P. (2014). Negative probabilities and Counterfactual Reasoning on the double-slit Experiment. *arXiv preprint arXiv:1412.4888*.

45. Burgin, M. (2009). Extended probabilities: mathematical foundations. *arXiv preprint arXiv:0912.4767*.

46. Lowe, D. (2004/2007) Machine Learning, Uncertain Information, and the Inevitability of Negative `Probabilities', *Machine Learning Workshop 2004*, Sheffield, England (http://videolectures.net/mlws04_lowe_mluii/)

47. Abramsky, S., & Brandenburger, A. (2014). An operational interpretation of negative probabilities and no-signalling models. In *Horizons of the mind. A tribute to Prakash Panangaden* (pp. 59-75). Springer, Cham.

48. Abramsky, S., Brandenburger, A., & Savochkin, A. (2014). No-signalling is equivalent to free choice of measurements. *arXiv preprint arXiv:1412.8523*.

49. Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524.

Chapter 7
Preliminary Studies of Coupling of the ATP Hydrolysis Cycle of The Bacteriophage T4 Packaging Motor to DNA Translocation

7.1 Introduction

The phage motors are powered by the energy released during ATP hydrolysis, where the motor catalyzes the reaction in which ATP binds to a motor subunit and is broken down into the products of ADP and inorganic phosphate (denoted Pi) which are released. Pi is a phosphate ion PO_4^{3-} (which in aqueous solution primarily forms a combination of hydrogen phosphate $[H_2PO_4]^-$ and dihydrogen phosphate $[HPO_4]^{2-}$ ions). Other important information has come from structural studies showing that the phage T4 and Phi29 motors are multimeric complexes consisting of five monomeric protein subunits each of which can catalyze ATP hydrolysis.[5-7] Initial prior efforts to understand the coupling of the ATP hydrolysis cycle to DNA translocation focused on the phage Phi29 motor and investigated the effects of changing the solution ATP, ADP, and Pi concentrations, as well as identifying compounds that could inhibit the cycle.[1,8] Lowering [ATP] sufficiently was found to slow the motor because diffusion of ATP to the motor becomes rate limiting. Increasing [ADP] sufficiently was found to slow motor because ADP can bind to the motor and interfere with ATP binding. Addition of high concentrations of Pi, however, did not affect the packaging rate, suggesting that Pi does not tend to rebind to the motor and that Pi release after hydrolysis is essentially irreversible. Very high resolution

measurements of Phi29 packaging found evidence that packaging occurs in bursts of four ~2.5 bp translocation steps, adding up to a net 10 bp DNA translocation increment, and was interpreted to indicate that the actions of four motor subunits are tightly coordinated and each one translocates DNA by ~2.5 bp per ATP hydrolyzed.[1,2,8,9] It is of interest to know whether the behavior is the same for other viral motors but this has proven difficult to measure for the T4 motor due to its faster translocation rate and more frequently observed backwards slipping when [ATP] is lowered.[10] However, with Phi29 it was also found that addition of orthovanadate (VO_4^{3-}, from added sodium orthovanadate) causes pauses in DNA translocation and sometimes clusters of multiple pauses separated by 10 bp are also observed.[1] This it thought to occur because orthovanadate is a Pi analog that can form stable complexes with ADP and delays dissociation of ADP from the binding pocket. Therefore, as a preliminary effort to investigate potential similarities or differences between the Phi29 and T4 motors, we conducted studies, described in this chapter, where we tested the effect on the T4 motor of adding phosphate ions or sodium orthovanadate to the solutions during packaging measurements.

7.2 Methods and Preliminary Results

7.2.1 T4 Packaging with Added Sodium Phosphate

T4 packaging experiments were conducted as described in Chapter 2 except the contents of the packaging buffer were changed. Sodium phosphate solution (from a phosphate buffer solution) was added into the T4 packaging buffer in order to increase the concentration of phosphate (['Pi'] or $[PO_4^{3-}]$), of which the solute normally released by the motor as a product of

ATP hydrolysis along with ADP. According to the bacteriophage packaging model and the ATP hydrolysis equation

$$ATP \rightarrow ADP + Pi\ ,$$

it is conceivable that the Pi and/or ADP release steps could be inhibited such that Pi and/or ADP could re-bind to the motor protein and, if so, this could interfere with ATP binding and potentially slow the ATP hydrolysis cycle and also slow DNA translocation. For the Phi29 motor it was found that high concentrations of ADP indeed did slow translocation, but the similar behavior was not found with high concentrations of Pi, suggesting that Pi release is essentially irreversible. Here, in our T4 studies the concentration of phosphate was varied from 50 μM to 5 mM and a small ensemble of packaging events were recorded for each condition, and the results are listed in Table 7.1. The motor velocities in the solutions with 0 μM (zero added Pi), 50 μM, 500 μM and 5 mM PO_4^{3-} were not significantly different to within the experimental measurement uncertainties. The pause durations and pause frequency were also measured in the same experiment and listed in Table 7.1. Neither the average pause frequency nor pause duration showed any evidence that the PO_4^{3-} induced more pauses while packaging.

The highest phosphate concentration we used of 5 mM, is estimated to be at least 1000 times greater than any stray Pi that might be present in the standard packaging buffer (which did not contain any added Pi). This quantity of 5 mM added Pi is the same amount that was added in the studies of the phage Phi29 motor that concluded that the Pi release step is essentially irreversible, since it indicates that the equilibrium constant for phosphate release is very large.[8] We can thus conclude from our measurements here that the same conclusion applies to the T4 motor.

7.2.2 T4 Packaging with Added Na_3VO_4

Additional T4 packaging experiments were done with added 2.5 μM Na_3VO_4, 3.5 μM Na_3VO_4 since prior studies with Phi29 found evidence that this induces pauses that sometimes occur in clusters separated by ~10 bp, which was understood in that system to correspond to a quantized burst of four ~2.5 bp translocation steps of the same kind which occur repeatedly during normal operation of the motor when only ATP is added. Here the T4 packaging experiment was conducted with inhibitors to test whether the T4 motor performs similar behavior or not, and if a quantized motor step-burst size could be discerned. In fact, instead of observing a bunch of pauses (stepping), we found that the T4 motor experiences more slipping and re-packaging. Results of preliminary measurements of T4 motor packaging are shown in Fig. 7.1. 105 prohead-DNA complexes were measured and N = 24 of them exhibited repeated episodes of packaging with slipping. With the added vanadate we rarely see continuous packaging interrupted by pauses. Instead, there was very limited translocation and many backwards slips. Usually only a few hundred bp DNA at most was packaged then slipping prevented further progress. In addition, the average packaging rate was reduced by ~10-fold. We can thus conclude that the added vanadate does indeed result in strong inhibition of motor translocation, presumably because it interacts with ADP to hinder ADP release. However, unlike for Phi29 in which the motor usually retained grip on the DNA in this state, with T4 it usually induced slipping. After the first slip and subsequent slips occurred packaging frequently resumed but usually only ~20-100 bp was packaged before another slip occurred (Figs. 7.1 and 7.2). Due to this slipping it was difficult to use this approach to try to detect a quantized motor step (or step

burst) size, as in the Phi29 studies, because that relies on long pauses (at least 0.1 second) which are reference lengths that can be accurately measured occurring before and after each step. Statistics on the slipping that occurred between episodes of packaging are presented in Fig. 7.3.

7.3 Future Work

Our results show the ATP hydrolysis was affected by phosphate and the phosphate analog thought to inhibit ADP release, however we caution that the results in this chapter must be considered preliminary because the dataset contains lower-than desired statistical repeats, and thus relatively high uncertainty. It also consists only of data recorded with specific samples and solutions during only a few days of experiments, and thus should be repeated to better establish reproducibility. In addition, another direction of interest that may be informative would be studies of translocation with mixtures of ATP and ADP to investigate whether, as found for Phi29, ADP could significantly re-bind to the motor (competing with ATP).

Table 7.1 Summary of result for T4 packaging with added phosphate.

Phosphate Concentration	Package Rate (bp/s)	Motor Velocity (bp/s)	Pause Count (count/kbp)	Pause Duration (second/kbp)	# of Events
0 mM	499 ± 99	671 ± 81	0.4016 ± 0.0172	0.6574 ± 0.0241	20
50 μM	606 ± 87	659 ± 76	0.2046 ± 0.0109	0.2020 ± 0.0101	16
500 μM	489 ± 56	543 ± 49	0.2793 ± 0.0076	0.3565 ± 0.0092	26
5 mM	590 ± 50	622 ± 44	0.1898 ± 0.0041	0.1625 ± 0.0045	34

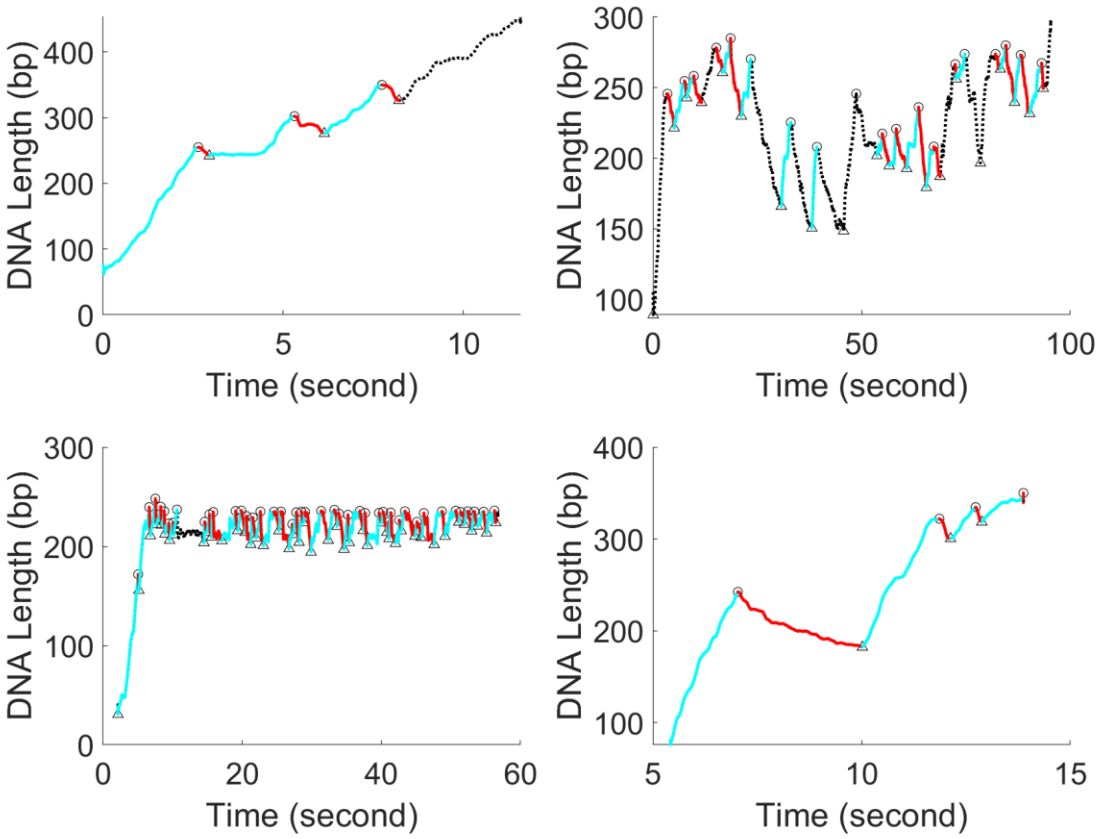

Figure 7.1 Examples of T4 packaging with added Na_3VO_4. Black dash line represents the DNA length packaged into prohead. Black circles and black triangles represent the starts of slipping and packaging, respectively. Cyan solid lines are the sections classified as packaging, and red solid lines are the sections classified as slipping. T4 motor translocated a few hundred bp DNA then interrupted by slipping. Short pauses are seen in some packaging complexes but unable to be identified by detection algorithm.

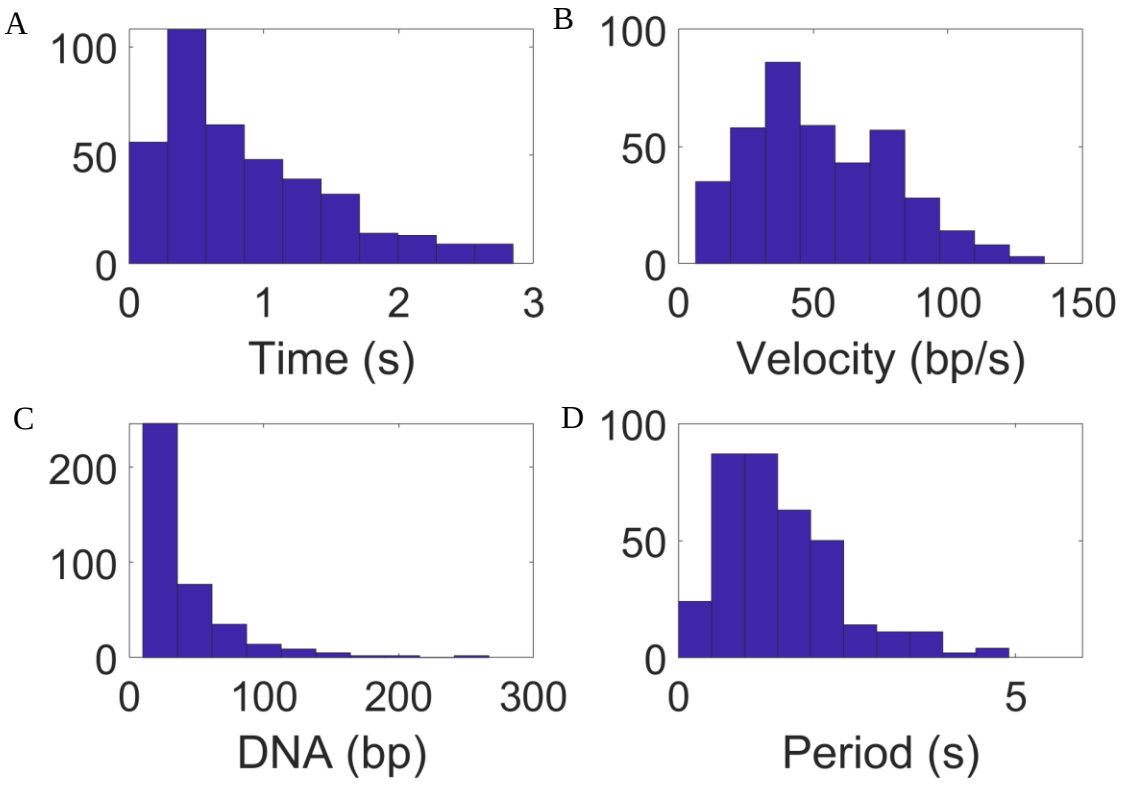

Figure 7.2 Quantification of T4 multiple packaging events occurring between episodes of back slipping with 2.5 µM added Na_3VO_4. Sections of data where episodes of packaging occurred in between slips were marked and the lengths, durations, and velocities of packaging quantified. (**A**) Histogram of durations of packaging. (**B**) Histogram of average motor velocities. (**C**) Histogram of the length of DNA packaged into the prohead in each packaging interval. (**D**) Time intervals between packaging segments.

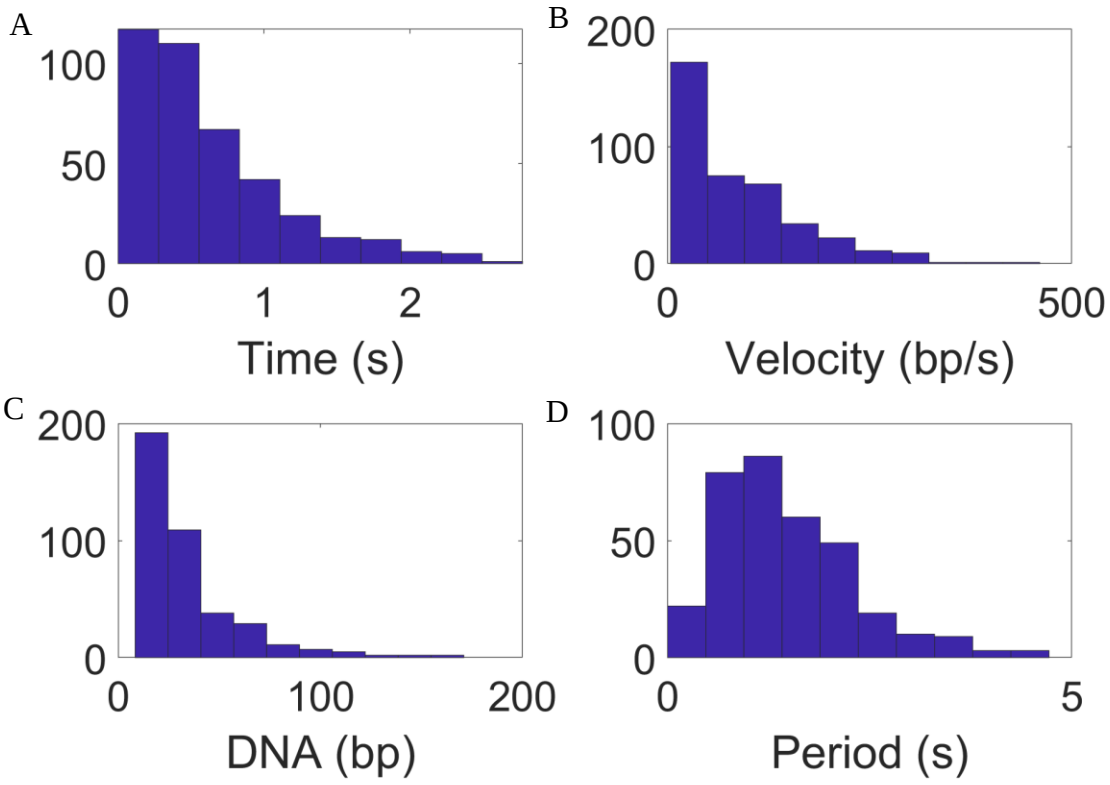

Figure 7.3 Quantification of slipping events occurring during the in T4 packaging measurements with 2.5 μM Na_3VO_4. Sections of data where episodes of slipping occurred in between episodes of packaging were marked and the lengths, durations, and velocities of packaging were quantified. (**A**) Histogram of durations of slipping episodes. (**B**) Histogram of reverse velocities during slipping. (**C**) Histogram of the lengths of DNA slipped out from the prohead in each slipping interval. (**D**) Time intervals between slipping events.

References

1. Chistol, G., Liu, S., Hetherington, C.L., Moffitt, J.R., Grimes, S., Jardine, P.J. and Bustamante, C. (2012) High degree of coordination and division of labor among subunits in a homomeric ring ATPase. *Cell*, **151**, 1017-1028.
2. Liu, S., Chistol, G., Hetherington, C.L., Tafoya, S., Aathavan, K., Schnitzbauer, J., Grimes, S., Jardine, P.J. and Bustamante, C. (2014) A Viral Packaging Motor Varies Its DNA Rotation and Step Size to Preserve Subunit Coordination as the Capsid Fills. *Cell*, **157**, 702-713.
3. Sun, S., Kondabagil, K., Draper, B., Alam, T. I., Bowman, V. D., Zhang, Z., Hegde, S., Fokine, A., Rossmann, M.G., & Rao, V. B. (2008). The structure of the phage T4 DNA packaging motor suggests a mechanism dependent on electrostatic forces. *Cell*, *135*(7), 1251-1262.
4. Ordyan, M., Alam, I., Mahalingam, M., Rao, V. B., & Smith, D. E. (2018). Nucleotide-dependent DNA gripping and an end-clamp mechanism regulate the bacteriophage T4 viral packaging motor. *Nature communications*, *9*(1), 1-9.
5. Alam, T. I., Draper, B., Kondabagil, K., Rentas, F. J., Ghosh‐Kumar, M., Sun, S., Rossmann, M. G., & Rao, V. B. (2008). The headful packaging nuclease of bacteriophage T4. *Molecular microbiology*, *69*(5), 1180-1190.
6. Mahler, B. P., Bujalowski, P. J., Mao, H., Dill, E. A., Jardine, P. J., Choi, K. H., & Morais, M. C. (2020). NMR structure of a vestigial nuclease provides insight into the evolution of functional transitions in viral dsDNA packaging motors. *Nucleic acids research*, *48*(20), 11737-11749.
7. Woodson, M., Pajak, J., Zhao, W., Zhang, W., Arya, G., White, M. A., Jardine, P.J. & Morais, M. C. (2020). A viral genome packaging motor transitions between cyclic and helical symmetry to translocate dsDNA. *bioRxiv*.
8. Chemla, Y. R., Aathavan, K., Michaelis, J., Grimes, S., Jardine, P. J., Anderson, D. L., & Bustamante, C. (2005). Mechanism of force generation of a viral DNA packaging motor. *Cell*, *122*(5), 683-692.
9. Yu, J., Moffitt, J., Hetherington, C. L., Bustamante, C., & Oster, G. (2010). Mechanochemistry of a viral DNA packaging motor. *Journal of molecular biology*, *400*(2), 186-203.
10. Kottadiel, V. I., Rao, V. B., & Chemla, Y. R. (2012). The dynamic pause-unpackaging state, an off-translocation recovery state of a DNA packaging motor from bacteriophage T4. *Proceedings of the National Academy of Sciences*, *109*(49), 20000-20005.