Supplementary Information

Substrate Interactions and Promiscuity in a Viral DNA Packaging Motor

K. Aathavan, Adam T. Politzer, Ariel Kaplan, Jeffrey R. Moffitt, Yann R. Chemla, Shelley Grimes, Paul J. Jardine, Dwight L. Anderson, and Carlos Bustamante
**Supplementary Table 1: List of Insert Oligos**

| Substrate     | Strand   | Sequence (from 5'-3')                                                                 |
|---------------|----------|---------------------------------------------------------------------------------------|
| 5 bp ds-MeP   | Top      | GAGAACTGATACTCACACTGCCTCAGTGATGTCAGTTCTCTTC                                          |
|               | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 9 bp ds-MeP   | Top      | GAGAACTGATACTCACACTGCCTCAGTGATGTCAGTTCTCTT                                          |
|               | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 10 bp ds-MeP  | Top      | GAGAACTGATACTCACACTGCCTCAGTGATGTCAGTTCTCTT                                         |
|               | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 11 bp ds-MeP  | Top      | GAGAACTGATACTCACACTGCCTCAGTGATGTCAGTTCTCTT                                         |
|               | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 15 bp ds-MeP  | Top      | GAGAACTGATACTCACACTGCCTCAGTGATGTCAGTTCTCTT                                         |
|               | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 30 bp ds-MeP  | Top      | GAGAACTGATACTCACACTGCCTCAGTGATGTCAGTTCTCTT                                         |
|               | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 5'-3' 15 bp   | Top      | GAGAACTGATACTCACACTGCCTCAGTGATGTCAGTTCTCTT                                         |
| Hybrid-MeP    | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 5'-3' 30 bp   | Top      | GAGAACTGATACTCACACTGCCTCAGTGATGTCAGTTCTCTT                                         |
| Hybrid-MeP    | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 3'-5' 30 bp   | Top      | GAGAACTGATACTCACACTGCCTCAGTGATGTCAGTTCTCTT                                         |
| Hybrid-MeP    | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| Linker        | Top      | GAGAACTGAT-cyclohexanediolalamiidepolyether--CCTCAGTGATGTCAGTTCTCTT                |
|               | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 10-PO4-Abasic | Top      | GAGAACTGAT-10xphosphatebackboneonly-CCTCAGTGATGTCAGTTCTCTT                          |
|               | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 5'-3' 10 nt   | Top-Left | CCGTACGTCAGTGATGTCAGTTCTCTT                                                       |
| ss-gap        | Top-Right| CCGTACGTCAGTGATGTCAGTTCTCTT                                                        |
|               | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 5'-3' 20 nt   | Top-Left | CCGTACGTCAGTGATGTCAGTTCTCTT                                                       |
| ss-gap        | Top-Right| CCGTACGTCAGTGATGTCAGTTCTCTT                                                        |
|               | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 3'-5' 20 nt   | Top      | CCGTACGTCAGTGATGTCAGTTCTCTT                                                       |
| ss-gap        | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 5'-3' 10 nt   | Top      | CCGTACGTCAGTGATGTCAGTTCTCTT                                                       |
| bulge         | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 5'-3' 20 nt   | Top      | CCGTACGTCAGTGATGTCAGTTCTCTT                                                       |
| bulge         | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 3'-5' 10 nt   | Top      | CCGTACGTCAGTGATGTCAGTTCTCTT                                                       |
| bulge         | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |
| 3'-5' 20 nt   | Top      | CCGTACGTCAGTGATGTCAGTTCTCTT                                                       |
| bulge         | Bottom   | ATCAGTGGAGTCAGTGGTACGTTCTCTT                                                        |

Modifications are in **red**. MeP in **red/bold**. The top strand is defined as the 5'-3' strand in the direction of packaging.
## Supplementary Table 2: Summary of results from all inserts

| Modification | Substrate Name | Force (pN) | Total # phages | Successful Traversal | Traversal Probability* | 95% c.i.** | Mean Pause (s) | SEM (s) |
|--------------|----------------|------------|----------------|----------------------|------------------------|------------|----------------|---------|
| **No Charges** |                |            |                |                      |                        |            |                |         |
| 5 bp ds-Mep  | 5              | 12         | 12             | 0.93                 | 0.79 - 1.00            | 0.1        | 0.1            |         |
| 5 bp ds-Mep  | 16 ± 2*        | 11         | 11             | 0.92                 | 0.77 - 1.00            | 1.5        | 0.3            |         |
| 5 bp ds-Mep  | 20             | 11         | 11             | 0.92                 | 0.77 - 1.00            | 0.6        | 0.2            |         |
| 9 bp ds-Mep  | 5              | 20         | 16             | 0.77                 | 0.58 - 0.93            | 1.2        | 0.3            |         |
| 10 bp ds-Mep | 5              | 103        | 84             | 0.81                 | 0.73 - 0.88            | 3.3        | 0.7            |         |
| 10 bp ds-Mep | 11± 4*         | 9          | 5              | 0.55                 | 0.27 - 0.81            | 14         | 6              |         |
| 10 bp ds-Mep | 20             | 27         | 7              | 0.28                 | 0.13 - 0.45            | 7          | 3              |         |
| 11 bp ds-Mep | 5              | 26         | 12             | 0.46                 | 0.29 - 0.65            | 4          | 1              |         |
| 15 bp ds-Mep | 2.5            | 11         | 4              | 0.38                 | 0.15 - 0.65            | 5          | 3              |         |
| 15 bp ds-Mep | 5              | 33         | 11             | 0.34                 | 0.20 - 0.50            | 35         | 18             |         |
| 15 bp ds-Mep | 10             | 21         | 5              | 0.26                 | 0.10 - 0.45            | 31         | 14             |         |
| 30 bp ds-Mep | 5              | 22         | 0              | 0.04                 | 0.00 - 0.13            | N/A        | N/A            |         |
| 30 bp ds-Mep | 0.5            | 19         | 2              | 0.14                 | 0.02 - 0.33            | 3          | 3              |         |
| 30 bp ds-Mep | 1              | 44         | 3              | 0.09                 | 0.02 - 0.19            | 25         | 21             |         |
| 10 bp ds-Mep, 25uM ATP | 5 | 43 | 22 | 0.51 | 0.37 - 0.65 | 6 | 2 |
| 5'-3' 30 bp hybrid MeP | 5 | 38 | 3 | 0.1 | 0.02 - 0.22 | 9 | 8 |
| 5'-3' 30 bp hybrid MeP | 1 | 23 | 12 | 0.52 | 0.33 - 0.71 | 11 | 7 |
| 3'-5' 30 bp hybrid MeP | 5 | 25 | 22 | 0.85 | 0.69 - 0.97 | 0.2 | 0.1 |
| 3'-5' 30 bp hybrid MeP | 20 | 17 | 9 | 0.53 | 0.31 - 0.74 | 0.2 | 0.1 |
| 5'-3' 15 bp hybrid MeP | 5 | 18 | 8 | 0.45 | 0.25 - 0.66 | 6 | 3 |
| 10 Phosphate-Abasic | 5 | 27 | 25 | 0.9 | 0.76 - 0.99 | 12 | 9 |
| 10 Phosphate-Abasic | 10 | 11 | 9 | 0.77 | 0.51 - 0.96 | 74 | 25 |
| 10 Phosphate-Abasic | 20 | 12 | 1 | 0.14 | 0.00 - 0.38 | 112 | N/A |
| **Gaps** |                |            |                |                      |                        |            |                |         |
| 5'-3' 20 nt ss-gap | 5 | 17 | 8 | 0.47 | 0.26 - 0.69 | 20 | 12 |
| 5'-3' 20 nt ss-gap | 10 | 13 | 6 | 0.47 | 0.23 - 0.71 | 36 | 24 |
| 3'-5' 20nt ss-gap | 5 | 27 | 21 | 0.76 | 0.59 - 0.90 | 1.3 | 0.7 |
| 3'-5' 20nt ss-gap | 10 | 25 | 21 | 0.81 | 0.65 - 0.94 | 9 | 2 |
| 5'-3' 10 nt ss-gap | 5 | 9 | 7 | 0.73 | 0.44 - 0.95 | 24 | 20 |
| 5'-3' 10 nt bulge | 5 | 8 | 7 | 0.8 | 0.51 - 1.00 | 1.7 | 0.5 |
| 5'-3' 10 nt bulge | 5 | 11 | 7 | 0.62 | 0.35 - 0.85 | 7 | 3 |
| 5'-3' 10 nt bulge | 20 | 11 | 9 | 0.77 | 0.51 - 0.96 | 2.6 | 0.6 |
| 5'-3' 20 nt bulge | 20 | 13 | 6 | 0.47 | 0.23 - 0.71 | 30 | 8 |
| 3'-5' 10 nt bulge | 5 | 14 | 11 | 0.75 | 0.52 - 0.93 | 0.2 | 0.1 |
| 3'-5' 20 nt bulge | 5 | 25 | 12 | 0.48 | 0.30 - 0.67 | 0.9 | 0.2 |
| 3'-5' 10 nt bulge | 20 | 4 | 4 | 0.83 | 0.54 - 1.00 | 0.1 | 0.1 |
| 3'-5' 20 nt bulge | 20 | 12 | 7 | 0.57 | 0.32 - 0.81 | 11 | 5 |
| **Bulges** |                |            |                |                      |                        |            |                |         |
| Linker | 5 | 33 | 32 | 0.94 | 0.83 - 1.00 | 0.5 | 0.1 |
| Linker | 20 | 24 | 23 | 0.92 | 0.78 - 1.00 | 0.7 | 0.1 |
| Linker | 30 | 27 | 19 | 0.69 | 0.51 - 0.84 | 1.5 | 0.1 |

*Traversal probability calculating using the Laplace estimator
**95% confidence interval calculated using the adjusted Wald method
# These measurements were made using a high resolution dual trap instrument without force feedback. The error bar on the force is the standard deviation in the average force at the insert
**Supplementary Table 3: Predicted force dependence of pauses**

| Modification     | Contour length (nm) | Force-1 (pN) | Extension-1 (nm)* | Pause-1 (s) | Force-2 (pN) | Extension-2 (nm) | Pause-2 (s) | Measured pause ratio | Predicted Pause ratio** |
|------------------|---------------------|--------------|-------------------|-------------|--------------|-----------------|-------------|----------------------|------------------------|
| 10bp ds-Mep      | 3.4                 | 5            | 3.1               | 3.3         | 20           | 3.3             | 7.4         | 2±1                  | 16,000                 |
| 15bp ds-Mep      | 5.1                 | 5            | 4.7               | 34.8        | 10           | 5.0             | 30.6        | 0.9±0.6              | 160                    |
| 5'-3', 20nt ssDNA| 14.7                | 5            | 6.3               | 19.8        | 10           | 8.9             | 36.1        | 2±2                  | 290,000                |
| 3'-5', 20nt ssDNA| 14.7                | 5            | 6.3               | 1.3         | 10           | 8.9             | 8.9         | 7±4                  | 290,000                |
| ds-10-PO4-Abasic | 7.7                 | 5            | 1.0               | 12.4        | 10           | 1.9             | 74.0        | 6±5                  | 19                     |
| ds-linker        | 4.1                 | 5            | 0.5               | 0.5         | 30           | 2.1             | 1.5         | 3.4±0.8              | 568,650                |

* An estimate of the minimum value of the extensions of inserts was calculated from the extensible worm-like chain model. The persistence length and stretch modulus of DNA neutralized with spermine, 40 nm and 1200pN, were used as an estimate of the properties of ds-Mep DNA. The values for PEG, 0.35nm and 1560pN, were used as an estimate for the abasic phosphate and the abasic linker inserts. Finally, the ssDNA values of 0.75nm and 800pN were used for the ssDNA gaps.

**Predicted using eq1 in Supplementary discussion**

**Supplementary Table 4: Summary of high resolution dynamics**

|                  | Mean (s) | STD (s) | SEM (s) | N   | N_Slip | N_Amount | N_Pass | N_Recover | N_Form | P_Slip* | P_Amount* | P_Pass* | P_Recover* |
|------------------|----------|---------|---------|-----|--------|----------|--------|-----------|--------|---------|----------|---------|-----------|
| Initial Pause    | 1.00     | 0.96    | 0.08    | 129 | 101    | 28       | N/A    | N/A       | N/A    | 0.78    | 0.22     | N/A    | N/A       |
| Attempt Pause    | 0.079    | 0.069   | 0.013   | 28  | 23     | 5        | N/A    | N/A       | N/A    | 0.80    | 0.20     | N/A    | N/A       |
| Slips            | N/A      | N/A     | N/A     | 124 | N/A    | 4        | N/A    | N/A       | N/A    | 0.96    |           |         | 0.96 (0.92,0.99) |

*Traversal probability calculating using the Laplace estimator; in parenthesis are 95% confidence intervals calculated using the adjusted Wald method.

**Supplementary Table 5: Encoding of selected inserts for logistic regression analysis**

| Modification     | Extension (nm) | Phosphates 5'-3' | Sugars/Bases 5'-3' | Special Phosphates | Native Structure | Molecules Traversed | Total |
|------------------|----------------|------------------|--------------------|--------------------|------------------|---------------------|-------|
| 5 bp ds-Mep      | 1.6            | -5               | 0                  | 0                  | 0                | 12                  | 12    |
| 10 bp ds-Mep     | 3.2            | -10              | 0                  | 0                  | 0                | 84                  | 103   |
| 11 bp ds-Mep     | 3.5            | -11              | 0                  | 0                  | 0                | 12                  | 26    |
| 5'-3' MeP30 hybrid | 9.6         | -30              | 0                  | 0                  | -2               | 3                   | 38    |
| ds-10-PO4-Abasic | 1.0            | 0                | -10                | 0                  | -1               | 25                  | 27    |
| ds-linker        | 0.5            | -10              | -10                | 0                  | -1               | 32                  | 33    |
### Supplementary Table 6: Logistic Regression Models of Data collected at 5pN

| Function | Model               | K | -2 log(L) | AICc | Δ   | w   |
|----------|---------------------|---|-----------|------|-----|-----|
| C. Log-Log | P53,P35,B35,SP     | 4 | 45.99     | 55.80| 0.00| 0.21|
| C. Log-Log | S,P53,P35,SP      | 4 | 46.81     | 56.63| 0.82| 0.14|
| C. Log-Log | E,S,P53,P35,SP    | 5 | 44.35     | 57.21| 1.40| 0.11|
| C. Log-Log | E,P53,P35,B35,SP  | 5 | 44.53     | 57.39| 1.59| 0.10|
| C. Log-Log | S,P53,P35,B53,SP  | 5 | 44.68     | 57.54| 1.73| 0.09|
| C. Log-Log | E,P53,P35,B35,SP  | 5 | 45.22     | 58.08| 2.28| 0.07|
| C. Log-Log | P53,P35,B35,SP    | 5 | 45.67     | 58.53| 2.72| 0.05|
| C. Log-Log | E,S,P53,P35,B53,SP| 6 | 42.74     | 58.94| 3.14| 0.04|
| C. Log-Log | E,S,P53,B35,SP    | 6 | 43.41     | 59.61| 3.80| 0.03|
| C. Log-Log | E,P53,P35,B35,SP  | 6 | 44.05     | 60.25| 4.45| 0.02|
| Probit   | E,P53,P35,B35,SP  | 5 | 47.49     | 60.35| 4.54| 0.02|
| Probit   | E,S,P53,P35,SP    | 5 | 47.99     | 60.85| 5.04| 0.02|
| C. Log-Log | S,P53,P35,B53,SP  | 6 | 44.68     | 60.88| 5.08| 0.02|
| Probit   | E,S,P53,P35,B35,SP| 6 | 45.46     | 61.66| 5.85| 0.01|
| C. Log-Log | E,S,P53,P35,B35,SP| 7 | 42.69     | 62.58| 6.78| 0.01|
| Probit   | E,S,P53,P35,B35,SP| 6 | 46.39     | 62.59| 6.79| 0.01|
| Probit   | E,P53,P35,B35,SP  | 6 | 46.98     | 63.18| 7.38| 0.01|
| Probit   | P53,P35,B35,SP    | 4 | 53.67     | 63.49| 7.68| <0.01|

Shown are the best 18 models, ranked and sorted by their AICc (Small sample Akaike Information Criteria). Function – Link function for the generalized linear model. K – Number of parameters in the model. L – Likelihood function, Δ – Difference in AICc with respect to the best model. The Akaike Weights $w = \exp(-\Delta/2)/\sum \exp(-\Delta/2)$, represent the likelihood of the specific model, given the data set.
**Supplementary Discussion**

**Force dependence of pause durations**

The effect of force on pause duration may be used to rule out purely diffusive traversal of inserted modifications. In the diffusive traversal scenario, the motor does not engage with the insert and therefore diffuses on this region of the substrate. If it diffuses backwards it re-engages with the unmodified DNA before the insert, preventing further backwards movement. If it diffuses forward to the end of the insert it engages with the unmodified DNA there, and then continues packaging. This scenario is akin to diffusion up a linear potential, and pulling on the DNA would affect the traversal time roughly exponentially,

\[
t = 2 \left( \frac{x^2}{2D} \right) \left( \frac{kT}{Fx} \right)^2 \left\{ \exp\left( -\frac{Fx}{kT} \right) - 1 + \frac{Fx}{kT} \right\}
\]

where \( x \) is the distance that the motor must diffuse against an opposing force \( F \), \( k \) is the Boltzmann constant, \( D \) is the diffusion coefficient, and \( T \) the temperature. The proper choice of \( x \) is not entirely obvious: it could be the force dependent end-to-end extension of the insert, or it could be its contour length. We chose the end-to-end extension, as it predicts a weaker force-dependence of the diffusive traversal time. The diffusion coefficient of the system is unknown, so the absolute traversal times cannot be calculated, but we can calculate ratios of traversal times for different forces. The ratios of the times are given by:

\[
\frac{t_1}{t_2} = \left( \frac{F_2}{F_1} \right)^2 \frac{\exp\left( -\frac{F_1 x(F_1)}{kT} \right) - 1 + \frac{F_1 x(F_1)}{kT}}{\exp\left( -\frac{F_2 x(F_2)}{kT} \right) - 1 + \frac{F_2 x(F_2)}{kT}}
\]
Supplementary Table 4 contains these diffusive traversal time ratios for inserts that were characterized at multiple forces. These ratios are all much larger than the observed pause ratios, ruling out the possibility of a purely diffusive traversal.

Quantitative Agreement between traversal probabilities and high-resolution branching probabilities

The probability of the motor successfully traversing an insert can be calculated from the probabilities of interconversion between the sub-states shown in Figure 3f as follows. Non-terminal slip events return the motor to the upstream pause where it makes a new attempt, so the motor will then keep attempting until it either successfully passes the insert or terminally slips from it. Therefore, the traversal probability is the fraction of these final events which are successful passes relative to terminal failures/slips:

\[
P_{\text{traversal}} = \frac{P_{\text{success}}}{P_{\text{success}} + P_{\text{failure}}}. \tag{3}
\]

The individual probabilities of success or failure, \( P_{\text{success}} \) and \( P_{\text{failure}} \), are given by the products of the probabilities of the sub-states along the path to the final event. There is only one direct path to success—the motor attempts to package from an upstream pause and then successfully passes—thus, the success probability is

\[
P_{\text{success}} = p_{\text{attempt}} p_{\text{pass}}. \tag{4}
\]

There are two direct paths to failure—i) the motor attempts to package from an upstream pause, slips from a downstream pause (it fails to pass), and then does not recover from this slip or ii) the motor slips directly from the upstream pause before it attempts to package and does not recover from this slip—thus, the failure probability is

\[
P_{\text{failure}} = p_{\text{attempt}} (1 - p_{\text{pass}}) p_{\text{terminal-slip}} + p_{\text{upstream-slip}} p_{\text{terminal-slip}}. \tag{5}
\]
For the high resolution packaging experiments on 10 bp of ds-MeP, the motor successfully traversed the insert in 5 out of 9 packaging experiments yielding a traversal probability, $P_{\text{traversal}}$, of 0.55 as listed in Supplemental Table 2. Inserting the interconversion probabilities listed in Figure 2f and Supplemental Table 3 into the equations above yields an estimate of the traversal probability of 0.54 in remarkable agreement with the traversal probability estimated by simply counting the number of successful crossings. Since the interconversion probabilities are calculated via a different analysis—by counting the microscopic slips and attempts—this agreement adds additional support to the validity of our proposed model in Figure 2f.

Moreover, this microscopic model can also predict values for the traversal probability for other lengths of neutral DNA that agree well with our low resolution measurements. Since the upstream pause occurs at the boundary of the charged and neutral DNA, then the attempt probability, $P_{\text{attempt}}$, should be identical for all lengths of neutral DNA simply due to the fact that the motor has no method to detect the length of the neutral DNA. Moreover, since the recovery from a slip occurs upstream of the neutral insert on charged DNA, this recovery probability should also be independent of the insert. Thus, the probability of not recovering, the terminal slip probability $P_{\text{terminal-slip}}$, should also be identical for all neutral DNA. The only probability which actually depends on the length of the neutral region is the probability of slipping from a downstream pause, the pause within the neutral insert.

If we assume that the probability of successfully passing an insert that is 30 bp long requires successfully passing a 10 bp insert three times, then the passage probability $P_{\text{pass}}$ for the 30 bp of ds MeP should be the third power of the passage probability of the
10 bp of ds MeP, i.e. \((0.2)^3 \approx 0.008\). This large decrease in probability is the origin of the dramatic drop in the probability of crossing large regions of neutral DNA. However, even though this probability is small, the motor can attempt many times, amplifying the final probability of successfully traversing the insert. Inserting this passage probability into the above expressions, Eqs. (3) - (5), accounts for these attempts and yields a final traversal probability of 0.04, in agreement with the measured value.

Discussion of apparent differences of ssDNA gap packaging results from previous studies

A previous study\(^{10}\) on the effect of nicks and gaps on \(\varphi 29\) DNA packaging in bulk concluded that nicks had no effect, while gaps on either strand completely stopped the motor\(^{10}\). In agreement we find no effect of nicks (data not shown). However, contrary to their results, we find that the motor can package gaps on either strand, but with reduced efficiencies that depend on length and strand. Our work is not easily compared with their observations because the nature of the gaps and the experimental sensitivity are different in the two studies.

In the previous work, gaps were introduced by extending site-specific nicks in the DNA by exploiting the 3’-5’ exonuclease activity of T7 DNA polymerase in competition with the polymerase activity in the presence of only dTTP. The lengths of the gaps are not homogeneous by experimental design, and the mean gap length was not determined. It is possible that the gaps were much longer than those tested in our study and contained significant secondary structure that may have impeded the motor. Our gaps were of defined length and contained only poly-CA, preventing secondary structure.
Another study that our findings differ from is the in vitro packaging by T4 of small DNA substrates with a variety of modifications\textsuperscript{11}. Using a 100bp DNA substrate, Oram et al found that a 20b gap inhibited packaging. This is in contrast to our finding of packaging of a 20b gap in the middle of a ~8000bp DNA substrate for φ29. However, this discrepancy may be due to the location of the modification in the DNA molecule. For example, modifications close to the end of the molecule may interfere with the process of packaging initiation. Interestingly, nicks affected packaging in T4 when present in a 100bp substrate, but had no effect in a longer 500bp DNA substrate\textsuperscript{11}, lending support to this possibility.
Modeling of modification traversal probability with a logistic regression analysis

The data set of traversal probabilities collected at different forces and from inserts lacking variable numbers and types of DNA chemical moieties, in one or both strands, allows us to address the relative importance of contacts with these moieties in a native packaging context, i.e. when all other moieties are present. Logistic regression (LR) analysis is a statistical method aimed at modeling the relationship between a dichotomous response variable and a set of explanatory variables (predictors)\(^{12}\). With a data set composed of multiple observations (e.g. molecules tested), each corresponding to a certain combination of the predictors (e.g. removal of bases and structure), LR provides a convenient and rigorous way to infer about the effect of each of the predictors separately (e.g. the effect of removing the bases without perturbing the structure). The first step in LR is a transformation of the probability (defined in \([0,1]\)) into a variable defined in \([-\infty, +\infty]\) using one of several sigmoidal functions. We use the term “logistic regression”, somewhat liberally, to denote not only a regression using the logit transformation but also regressions with other link functions appropriate for a binomial response variable, i.e. Probit and Complementary LogLog.\(^{12}\) The transformed probability is then modeled as a linear combination of the explanatory variables. For example, the classical logit function given by

\[
\text{Logit}(p) = \log\left(\frac{p}{1 - p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k
\]  

(6)

where \(p\) is the probability that the response variable equals one of its possible values, e.g. “success”, \(X_1, X_2, \ldots, X_k\) are the explanatory variables and \(\beta_0, \beta_1, \beta_2, \ldots, \beta_k\) their coefficients.\(^{12}\) A regression procedure based on a maximum likelihood criterion is used then to find these coefficients. While the LR model does not necessarily represent the
true physical model, it provides information on the effect of each of the predictors on the response variable. Notably, if different predictors have the same units, then the values for their coefficients show the relative magnitude of their effect on the response variable.

We performed a logistic regression analysis of our traversal probability data, using as potential predictors for the model: 1) The number of phosphate charges removed in each strand (denoted by the variable P53 for the phosphates in the 5’-3’ strand, and P35 for the phosphates in the 3’-5’ strand), 2) The number of bases and sugars removed in each strand (B53 and B35), 3) The number of special phosphates removed every 10 bp (SP), 4) The presence of native double-strand structure (S equals 0 for native DNA or Mep inserts, -1 for the rest of the inserts), and 5) The end-to-end extension of the insert under the applied force (E, a continuous variable calculated using the contour and persistence lengths as in Supplementary table 3). For example, for 5 bp ds-MeP, where 5 phosphates are “removed” from both strands the predictors are P53 = -5 bp, P35 = -5 bp, B53 = 0, B35 = 0, SP = 0 and E = 1.58 nm. Supplementary Table 5 shows additional examples for this encoding method. The unpaired bulge modifications do not “remove” elements from the DNA and therefore were not included in this analysis. The analysis was performed using the statistical package XLSTAT 2009 (Adinsoft) and the statistical toolbox of Matlab (Mathworks).

Our a priori set of explanatory variable includes all the chemical moieties of DNA and additional parameters that vary in the different experiments. Does each one of these parameters, when controlling for the others, really affect translocation? The answer to this question can be found by a methodical process of model selection. A variety of models can be constructed using different subsets of the explanatory variables defined
above, and each is characterized by a different quality of fit to the data (as represented by
the likelihood function, \( L \)). However, despite the improvement in \( L \) achieved by adding
additional parameters, it is important to avoid over-fitting the data, which will lead to
wrong conclusions. Information theory provides a convenient and rigorous way to rank
the different models that can be constructed, in order to select the best model, which
includes only those variables that have an effect on the calculated probability. The (small-
sample) Akaike's information criterion, \( AICc = -2 \log(L) + 2K\left(\frac{n}{n-K-1}\right) \) where \( K \) is the
number of parameters and \( n \) is the number of samples, provides a balance between bias
and parsimony. In order to be included in the model, a certain variable needs to improve
the likelihood of the model in a way that outweighs the additional complexity of a model
with a larger number of variables.

We first performed a regression using all of our data collected against a force of 5
pN. With 7 variables and 3 link functions (Logit, Complementary LogLog given by \( \log(-\log(1-p)) \), and probit, equal to the inverse of the normal cumulative distribution function),
it is possible to construct 381 different models, ranging from those that include only a
single explanatory variable to the complete set. As seen in Supplementary Table 6, the
model with the lowest AICc (i.e. the best model) is a Complementary Log-Log model
that includes the phosphates in both strands(P53, P35), the bases in the 3’-5’ strand(B35),
and contacts with phosphates separated 10 bp on the 5’-3’ strand(SP). The results for the
predictors’ coefficients are summarized in Supplementary Figure 2a. Supplementary
Figure 2b shows the predicted traversal probabilities calculated for all our measurements,
alongside the measured probabilities.
As an additional control for the validity of the parameter SP, we defined additional variables representing the removal of phosphates with periodicities ranging from 6 to 18 bp, and for each of these variables we performed a regression in which the control variable replaced SP. Supplementary Figure 2c shows the calculated AICc’s for these models. The dashed line in the figure indicates the AICc for a model that includes only P53, P35 and B35. Only SP produces a significant improvement on the quality of the model.

The results of the regression performed on the 5pN data set confirm our previous analysis of the neutral ds-MeP experiments, indicating the importance of phosphate contacts every 10 bp on the 5’-3’ strand. They also confirm the existence of contacts with the rest of the phosphates on both strands. Interestingly, the regression also reveals non-charge contacts in the form of contacts with bases or sugars, a subtle feature that may not be obvious from direct inspection of the traversal probabilities. Moreover, since all the perturbations are parameterized with the same units, i.e. the number of bp over which the change is made, the regression provides a ranking of the relative importance of the different contacts at 5 pN. Clearly, contacts every 10 bp on the 5’-3’ strand are more important than other contacts on the same strand, and those in turn are more important than the phosphates, bases and sugars on the 3’-5’ strand. Using the above results, we are now able to predict the traversal probability for a particular perturbation (e.g. removing a phosphate in the 3’-5’ strand), while all the other contacts are not perturbed. We define the “importance” of each contact as the inverse of the length (bp) that need to be perturbed in order to reduce the probability to 50%. These results are summarized in Figure 4 as a “heat map” on DNA.
Supplementary Table 6 also shows the AICc differences, \( \Delta_i = \text{AICc}_i - \text{AICc}_{i}^{\text{best}} \), over all the candidate models. Using these differences we calculate the Akaike Weights \(^9\) given by

\[
    w_i = \frac{\exp[-\frac{1}{2} \Delta_i]}{\sum_r \exp[-\frac{1}{2} \Delta_r]},
\]

where \( R \) is the number of models being considered. These weights (shown in Supplementary Table 6) quantify the plausibility of this model being the actual best model in the set of candidate models, i.e. the likelihood of the model, given the data.

Since the weight of evidence for our selected model being the best model is not highly conclusive (e.g. the weight for the best model is only 1.5 times larger than the weights of the second-best), it is reasonable to explore the predictions of additional models in our set using multi-model inference techniques \(^9\). The 18 models with the highest \( w \)’s comprise 95\% of the model likelihood (i.e. \( \sum_{i=1}^{18} w_i > 0.95 \)), and we use this subset for our multi-model inference. We use the information from the different models in two different ways. First, we compute the model-averaged coefficients by averaging the predicted values for the coefficients from the different models, weighting these values with the corresponding \( w_i \). While the set of average coefficients includes also an effect for the insert extension and structure (data not shown), the values for the coefficients corresponding to the phosphates and bases, including the special phosphates, are in good agreement with our prediction based on one selected model only (Supplementary Figure 2a). A second way for accounting for the information content of the different models is to sum, for each parameter, the \( w \) values of all the models in the set that include this particular parameter.
The result can be viewed as the likelihood of a parameter\(^9\), given the data. These results also support our previous results: All the models in the top 95% likelihood contain the phosphates in both strands and the phosphates separated by 10 bp, while most of them contain also the bases and sugars on the 3’-5’ strand. The likelihood of P53, P35, B35 and SP is 0.95, 0.95, 0.55, and 0.96 respectively.

The regression results indicate that the motor makes contact with different chemical moieties of DNA, and that the traversal probability can be explained in terms of the number of these moieties that are removed. Surprisingly, they also reveal that removing all of these moieties—phosphates, bases, and sugars—will result in a reduced but finite traversal probability, indicating the existence of additional contacts, not included in our models. Notably, this is true not only for the best model, but for all of the models in our candidates set.

Next, we turned our attention to the effect of force on the traversal probability. We concentrate on the effect of force on the phosphate contacts, which the previous regression revealed as the most important specific contacts, and performed a regression on the data collected for ds-MeP and ss-MeP, against different forces. Since different contacts may have a different force dependence, we introduced the force dependence by using as predictors the products of each one of the chemical predictors with the force (e.g. P53*F). This means that for each of these predictors there will now be two terms in the linear function, and the regression will find two coefficients: a “force-independent” coefficient (e.g. the coefficient of P53) and a “force-dependent” one, (the coefficient of P53*F). The results of the model are shown in Supplementary Figure 2d. Remarkably, the entire effect of the special phosphates is a force-dependent one (i.e the coefficient for
SP * F is significant, but the one for SP is not), predicting that without an external force there will be no effect of removing these phosphates. This surprising result is an additional indication to the fact that these contacts belong to the dwell phase, where the only load on the motor is the externally applied force. The rest of the phosphate contacts have both a force-dependent and a force-independent weight, supporting the fact that these phosphates are contacted during the burst phase, when the motor generates force. Comparing the force-dependent weights of the various phosphates shows that the contacts during the dwell exhibit much higher force sensitivity than the contacts during the burst, an indication of the higher “strength” of the special contacts (as the force increases, removing this contact will result in a much larger decrease in probability. This means that without SP the motor cannot sustain a high force, indicating that these are strong contacts). Note that force-independent and force-dependent weights should not be compared directly since the absolute value of the latter depends on the choice of force units.
Supplementary Figures and Legends

Supplementary Figure 1. High Resolution Studies of Traversal of 5 bp of dsMeP.

To probe the identity of the long, upstream pauses observed at high resolution in the crossing of 10 bp of dsMeP, we package shorter regions of neutral DNA, 5 bp of dsMeP. (a) If the motor is situated on the DNA such that one dwell occurs upstream of the modified DNA by more than 5 bp from the start of the neutral insert, the 10-bp burst will position the dwell within the neutral region. If the upstream pause corresponds to a dwell phase interaction with the neutral DNA, we expect to observe a pause of a long duration. (b) However, if the motor is situated such that this dwell occurs less than 5 bp from the start of the neutral DNA, the 10-bp burst will carry the motor across the neutral DNA and the next dwell will be positioned on charged DNA. If the upstream pause corresponds to the dwell phase, then we would expect to observe no discernable pause in packaging. Since there is a slight variability in the 10-bp burst size, we expect that the motors will randomize their position on the DNA across the 4 kb of charged DNA which must be packaged before the insert. Thus, if the upstream pause corresponds to a dwell phase interaction with the neutral DNA, we expect to see only ~50% of the motors pause at the insert. (c) Example packaging traces of 5 bp dsMeP using the high resolution optical tweezers under 1 mM [ATP] and ~15 pN of opposing load. Note that some traces show no long pause distinguishable from the normal distributions of dwell phase. (d) The
mean pause duration, the traversal probability, and the probability of observing a pause 
\( N_{\text{phage}} = 11; N_{\text{pause}} = 5 \). Error bars for the probabilities are 95% confidence intervals 
calculated via the adjusted Wald method and the standard error of the mean for the pause 
duration. The observed pauses occurred at a mean position of \( 3840 \pm 20 \text{ bp} \) (s.d.); and a 
region of DNA 3 times the standard deviation in the position of the pauses was used as 
the candidate position of the insert. To account for the variability in dwell times on the 
DNA during normal packaging, we scored any event with durations of 300 ms or longer 
as a pause, a duration 3-fold larger than the typical dwell time on charged DNA. Because 
the probability of observing a pause is statistically inconsistent with 100%, we conclude 
the upstream pause corresponds to the dwell phase.
Supplementary Figure 2: Logistic regression of the importance of contacts with different DNA moieties. SP—contacts with phosphates every 10 bp on the 5’-3’ strand. P53—contacts with phosphates in the 5’-3’ strand. P35—contacts with phosphates in the 3’-5’ strand. B35—contacts with bases and sugars in the 3’-5’ strand. (a) Logistic regression using all data collected at 5 pN. Red bars—values for the coefficients of the model parameters (error bars are 95% c.i.). Blue symbols—Mean coefficients averaging the best 18 models, totalling 95% of the Akaike Weights. These results provide a ranking of the importance of different contacts during packaging. (b) Experimental measurements (red) and model predictions (green) for the experiments at 5 pN. The predictions were calculated with the coefficients corresponding to the best model as shown in (a). The experimental result of the ds-linker was not included in the regression calculation. (c) Small sample Akaike Information Criteria (AICc) values for regressions using P53, P35, B35 and a control parameter instead of SP. The control parameters represent phosphates removed with periodicities from 6 to 18 bp instead of SP. (d) Regression using MeP inserts, at a variety of forces. The special phosphates have only a force-dependent weight, consistent with those contacts occurring during the dwell. The
rest of the phosphates in both strands have a also a zero-force component, likely a result of being contacted during the burst, when the motor generates force.
Supplementary References

1. Lewis, J., Sauro, J. When 100% Really Isn’t 100%: Improving the Accuracy of Small-Sample Estimates of Completion Rates. *Journal of Usability Studies* 1, 136-150 (2006).
2. Agresti, A. & Coull, B. Approximate Is Better Than" Exact" for Interval Estimation of Binomial Proportions. *The American Statistician* 52(1998).
3. Moffitt, J.R. et al. Intersubunit coordination in a homomeric ring ATPase. *Nature* 457, 446-450 (2009).
4. Wang, M.D., Yin, H., Landick, R., Gelles, J. & Block, S.M. Stretching DNA with optical tweezers Biophys. J 72, 1335-1346 (1997).
5. Baumann, C.G. et al. Stretching of single collapsed DNA molecules. *Biophys J* 78, 1965-78 (2000).
6. Kienberger, F. et al. Static and Dynamical Properties of Single Poly(Ethylene Glycol) Molecules Investigated by Force Spectroscopy. *Single Molecules* 1, 123-128 (2000).
7. Smith, S.B., Cui, Y. & Bustamante, C. Overstretching B-DNA: the elastic response of individual double-stranded and single-stranded DNA molecules. *Science* 271, 795-9 (1996).
8. Howard, J. *Mechanics of motor proteins and the cytoskeleton*, xvi, 367 p. (Sinauer Associates, Pub., Sunderland, Mass., 2001).
9. Burnham, K.P. & Anderson, D.R. *Model selection and multimodel inference: a practical information-theoretic approach*, (Springer, 2002).
10. Moll, W.D. & Guo, P. Translocation of nicked but not gapped DNA by the packaging motor of bacteriophage phi29. *J Mol Biol* 351, 100-7 (2005).
11. Oram, M., Sabanayagam, C. & Black, L.W. Modulation of the packaging reaction of bacteriophage t4 terminase by DNA structure. *J Mol Biol* 381, 61-72 (2008).
12. Agresti, A. *An introduction to categorical data analysis*, (Wiley-Interscience, 2007).