Supplemental Information

MOSAIC: A Modular Single Molecule Analysis Interface for Decoding Multi-state Nanopore Data

Jacob H. Forstater[#,°], Kyle Briggs[†], Joseph W.F. Robertson[#], Jessica Ettedgui[#,°], Olivier Marie-Rose[^], Canute Vaz[#], John J. Kasianowicz[#], Vincent Tabard-Cossa[†], and Arvind Balijepalli[#,*]

[#] Physical Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

[°] Department of Chemical Engineering, Columbia University, New York, NY 10027, USA

[†] Department of Physics, University of Ottawa, Ottawa, Ontario K1N 6N5, Canada

[^] Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Table of Contents

Section 1: MOSAIC Pipeline Architecture

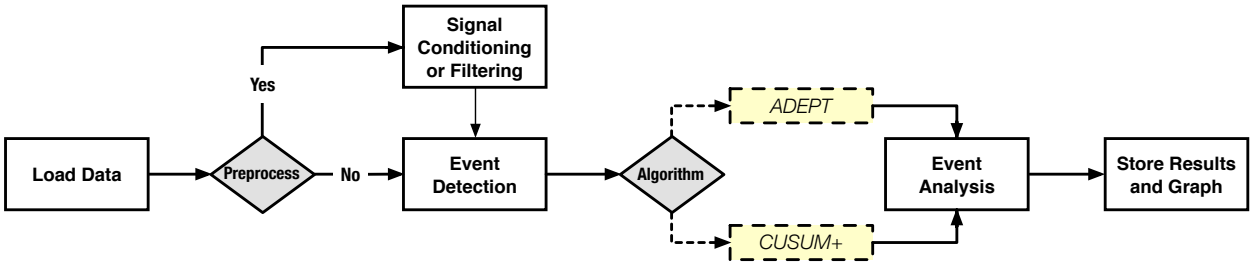


Figure S1. The MOSAIC processing pipeline comprises of five modules that perform well-defined functions. By enforcing a common interface, individual functions can be easily customized in a manner that allows them to interoperate with other parts of the pipeline.

MOSAIC consists of a modular data processing pipeline, shown schematically in Fig. S1. Five self-contained modules provide the building blocks of the pipeline, which together allows the analysis of data from single-molecule nanopore experiments. MOSAIC was designed using object-oriented concepts. This allowed us to leverage polymorphism within individual modules to customize functionality and promote interoperability. Module interoperability is also ensured through a well-defined interface. This approach makes it straightforward for users to implement new features into the software, such as other analysis algorithms or loading custom data formats. In most cases, users interact with MOSAIC using a custom graphical user interface (Fig. 6) that uses the pipeline architecture. Here we briefly describe the key elements of that pipeline architecture.

Load Data: This module allows users to load raw current traces from nanopore experiments into MOSAIC. Axon binary format (ABF) by Molecular Devices Axopatch amplifiers and QUB data files (QDF) are supported natively. In addition, MOSAIC can read a wide variety of raw binary file format for data saved with custom hardware. Data from disk is first loaded into a first-in, first-out (FIFO) queue after applying the specified signal conditioning steps (setting amplifier scale and offsets, correcting for systematic artifacts, etc.). Downstream modules request data in blocks for subsequent processing. The *Load Data* module adds data to the FIFO queue on demand until the end of the data set is reached. Because the logic for the data access is implemented as part of the common interface, new data types can be added to MOSAIC by simply reading in individual files and making the data available to the FIFO queue.

Filtering: MOSAIC allows acquired data to be filtered prior to analysis by applying the optional *Filtering* module. While this module is not a substitute for anti-alias filtering, it allows the data bandwidth and sampling to be controlled in software.[1] High frequency noise, resulting from amplifier noise or charge trapping in solid-state measurements can hinder the measurement of short events (< 100 μs). In some cases, the signal-to-noise ratio (SNR) can be improved by filtering data, albeit with distortion of the signal shape.[2] MOSAIC provides multiple options for data filtering including Bessel which has a well-behaved response to signal transients.[21] Alternatively, arbitrary filters (finite impulse response filters, weighted moving averages, etc.) can be implemented using a tap-delay line.[3] Finally, this module provides experimental support for wavelet-based filtering,[3] which allows a signal to be denoised while preserving SNR and pulse shape.

Event Detection: This module partitions the time-series to detect individual interactions of single molecules with the nanopore. Currently, we have implemented a simple thresholding algorithm that detects deviations of the ionic current from the open channel baseline value.[2,4,5] To discern statistically meaningful events, we heuristically set the current threshold. Typical values range from 2.5 to 6 times the standard deviation of the open channel current,

σ. The end of an event is registered when the ionic current returns to the open-channel baseline. The segment of the time-series corresponding to the event is then processed further using the *Event Analysis* module.

Event Analysis: Detected events are analyzed using the user-selected algorithm as described in the main text. Currently two algorithms—ADEPT and CUSUM+ are available. However, additional algorithms can be added. The parameters and associated metadata of successfully analyzed events are stored in a SQLite database (Figure S2) for future analysis. Events with processing or analysis errors are flagged within the database.

Store Results & Graph: The results generated by the *Event Analysis* module are stored in a SQLite database. The use of a relational database for storage enables rapid, non-platform specific data storage and access and allows users to explore the data and analyze it further using external applications. The database schema for the ADEPT is shown in Fig. S2A. Primary metadata generated from the analysis of individual events in a time-series are stored in the *metadata* table. The *analysisinfo* table holds additional analysis information needed to make the database self-contained. Finally, the *analysissettings* table holds the settings used to run the analysis and the output log of the analysis is stored in the *analysislog* table. A brief description of the metadata generated by the ADEPT algorithm is presented in Fig. S2B. An up-to-date listing of metadata of all algorithms available in MOSAIC is available at https://pages.nist.gov/mosaic/.

A

metadata	[table]
recIDX	INTEGER
ProcessingStatus	TEXT
OpenChCurrent	REAL
NStates	INTEGER
CurrentStep	BLOB
BlockDepth	BLOB
EventStart	REAL
EventEnd	REAL
EventDelay	BLOB
StateResTime	BLOB
ResTime	REAL
RCConstant	BLOB
AbsEventStart	REAL
ReducedChiSquared	REAL
ProcessTime	REAL
TimeSeries	BLOB

analysisinfo	[table]
datPath	TEXT
dataType	TEXT
partitionAlgorithm	TEXT
processingAlgorithm	TEXT
filteringAlgorithm	TEXT
analysisTimeSec	REAL
dataLengthSec	REAL
FsHz	REAL
mosaicVer	TEXT
mosaicBuild	TEXT
recIDX	INTEGER

analysissettings	[table]
settings	TEXT
recIDX	INTEGER

analysislog	[table]
logstring	TEXT
recIDX	INTEGER

B

metadata	
recIDX	Record index
ProcessingStatus	Status of the analysis
OpenChCurrent	Open channel current in pA.
NStates	Number of detected states.
CurrentStep	Blocked current steps in pA.
BlockDepth	BlockedCurrent/OpenChCurrent for each state.
EventStart	Event start in ms.
EventEnd	Event end in ms.
EventDelay	Start time of each state in ms.
StateResTime	Residence time of each state in ms.
ResTime	EventEnd-EventStart in ms.
RCConstant	System RC constant in ms.
AbsEventStart	Global event start time in ms.
ReducedChiSquared	Reduced Chi-squared of fit.
ProcessTime	Analysis time in ms.
TimeSeries	Event time-series data.

Figure S2. (A) Database schema for ADEPT algorithm output in MOSAIC. Primary metadata associated with processed events are stored in the *metadata* table. (B) Descriptions of the different metadata produced by ADEPT.

Section 2: Time Constant Estimation

As described in the paper, the time constant τ varies with ΔR, $\tau = \frac{C_m R_s (R_p + \Delta R)}{R_s + R_p + \Delta R}$. However, the actual difference between the time constants leading up to and following an event are considerably shorter than the sampling rate used in most experiments. For example, the partitioning of DNA into an αHL nanopore, discussed in paper, produces a blockade depth ratio, $\langle i \rangle / \langle i_0 \rangle \approx 0.1$. For typical experimental parameters, the time constant associated with entry and exit of the DNA into the nanopore will differ by by ≈ 293 ns (assuming R_p = 0.8 GΩ, R_s = 50 MΩ, C_m = 2 pF, and ΔR = 7.2 GΩ). This change is even smaller (≈ 237 ps) for solid state nanopores.[6,7] Because these differences are ≈ 1000 times shorter than the duration between sampled points (2 μs) at a sampling rate, F_s=500 kHz, they cannot be resolved by our measurements. We therefore utilize a single fit parameter for τ, which reduces the degrees of freedom in the fit. For cases where these assumptions do not hold, there is an option to override this constraint in the software.

Section 3: Extended Analysis and Comparison of ADEPT and CUSUM+

(A) Solid State Nanopores: dsDNA

As discussed in the text, the observed difference between the output of ADEPT and CUSUM+ results from the limitations of the CUSUM+ algorithm. Because CUSUM+ can only correctly identify states that have reached their steady state value (with residence times > 5τ) we normally exclude events less than 5τ from CUSUM+ analysis.

Relaxing this constraint results in a systematic underestimation of the ionic current when DNA occupies the pore, causing the blockade depth histogram to shift to the right (compare to Fig. 3B in text). This underestimation will also apply to sub-states shorter than 5τ within a longer event.

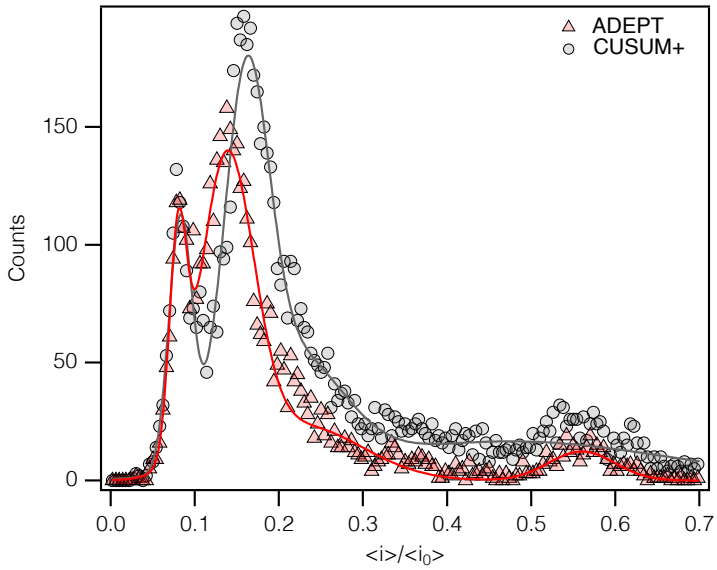


Figure S3. A comparison of ADEPT (*red*) and CUSUM+ (*grey*) blockade depths for dsDNA events with a minimum length of 2τ. Including events in CUSUM+ that do not converge to a steady state (< 5τ) results in a systematic shift in the peak positions.

In some instances, ADEPT's fitting process can fail to converge for events containing more than ~12,500 points (25 ms when F_s = 500 kHz). This is seen in Fig. 3B, where CUSUM+ recovers ≈ 11% more events for the peak at $\langle i \rangle / \langle i_0 \rangle = 0.56 \pm 0.01$, which contains characteristically long events. The overall result can be improved when both algorithms are used in a complementary manner—using ADEPT to fit short events and CUSUM+ to fit long ones, as seen in Fig. S4.

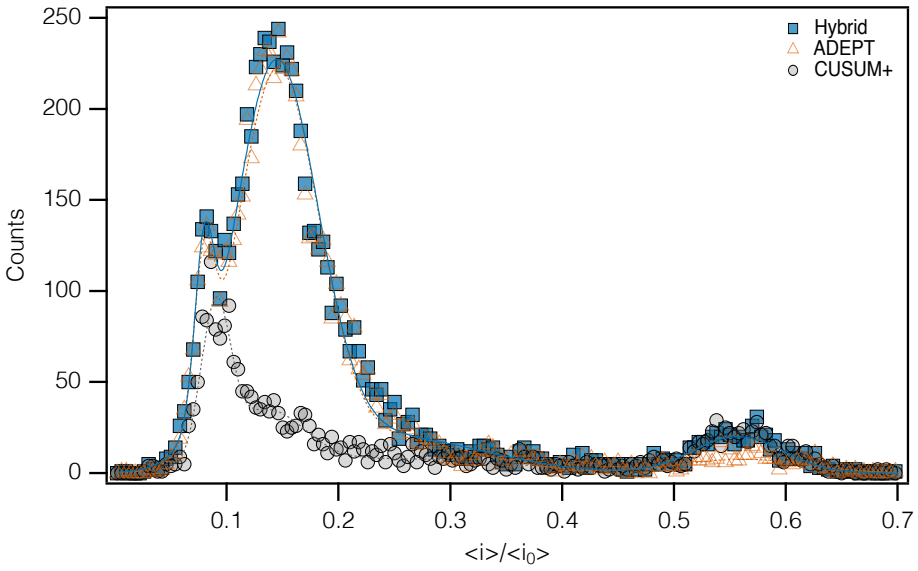


Figure S4. A proof of concept hybrid approach (*Blue*, *rectangles*) that merges the results of ADEPT (*Orange*, *triangles*) and CUSUM+ (*Gray*, *circles*) improves the analysis of the dsDNA data.

(B) Computational Efficiency of Algorithms

The processing times for both ADEPT and CUSUM+ scaled linearly with the length of the event, as seen in Figure S5, suggesting $O(N)$ scaling for each algorithm, as implemented currently.

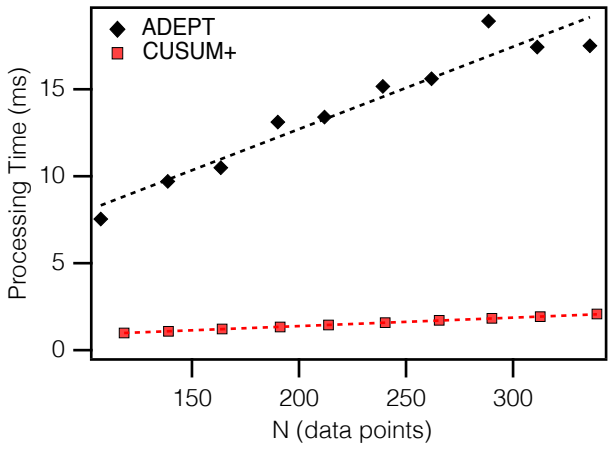


Figure S5. Processing time as a function of number of data points in the event (event length), *N*, for ADEPT (*black*) and CUSUM+ (*red*). CUSUM+ is on average ~10x faster. The processing time for both algorithms scales linearly with N (R^2 = 0.927 and 0.999 for ADEPT and CUSUM+ respectively). Linear fits to data are shown in dotted lines (Slopes are: (47 ± 5) µs/pt and (4.87 ± 0.02) µs/pt for ADEPT and CUSUM+ respectively). The results were obtained from data processed on a computer with an Intel 3.6 GHz i7-4790 CPU, 16 GB RAM and a solid-state hard drive.

(C) Effect of Voltage on ssDNA Blockades as a Function of Polymer Length

We examine the translocation of the homopolymer, poly(dA)$_n$, through the nanopore (*cis*),[8] as a function of hompolymer length and applied potential. The blockade depth histogram produced when the polynucleotides enter from the *cis* side of the pore is shown in Figure S6 as a function of applied potential. At 130 mV, the blockade depth histogram of dA$_{100}$ has three distinct peaks, $\langle i \rangle / \langle i_0 \rangle$ =(0.11 ± 0.01), (0.15 ± 0.03), and (0.51 ± 0.02). The first two peaks (denoted $\langle i \rangle / \langle i_0 \rangle_{3'}$ and $\langle i \rangle / \langle i_0 \rangle_{5'}$) are consistent with the dependence of the blockade depth on the orientation of the leading end of the DNA entering the pore (3' vs 5').[9-11] The location of these two peaks agrees with previous measurements of dA$_{100}$, where two broad overlapping peaks were observed at similar locations.[12]

These differences are not as clearly observed in the shorter ssDNA. As seen in Fig. S6B-C, the position of the 3' peak of dA$_{40}$ and dA$_{20}$ does not change with voltage. The amplitude of the 5' peak decreases substantially for measurements of dA$_{40}$ (Fig. S6B) and is not observed for dA$_{20}$ (Fig. S6C), likely due to both the lower probability of 5' entry as well as the decreasing residence time of shorter ssDNA.[10]

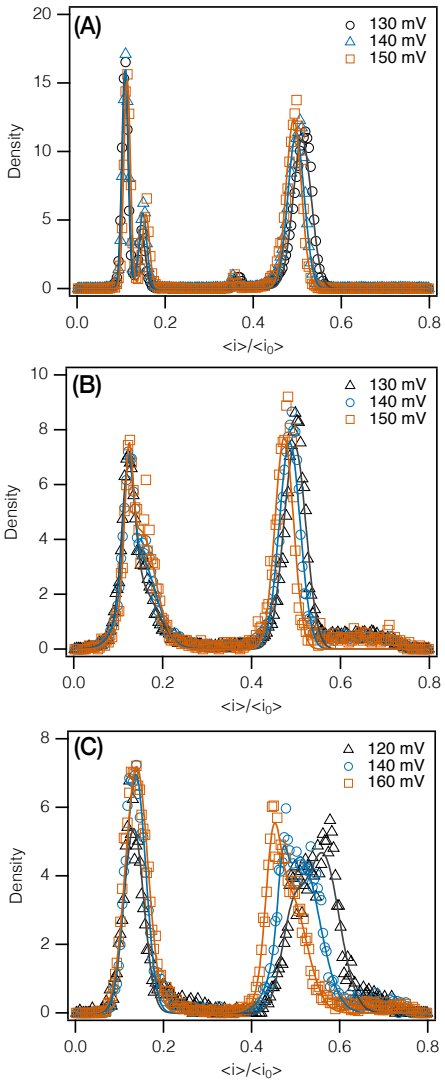
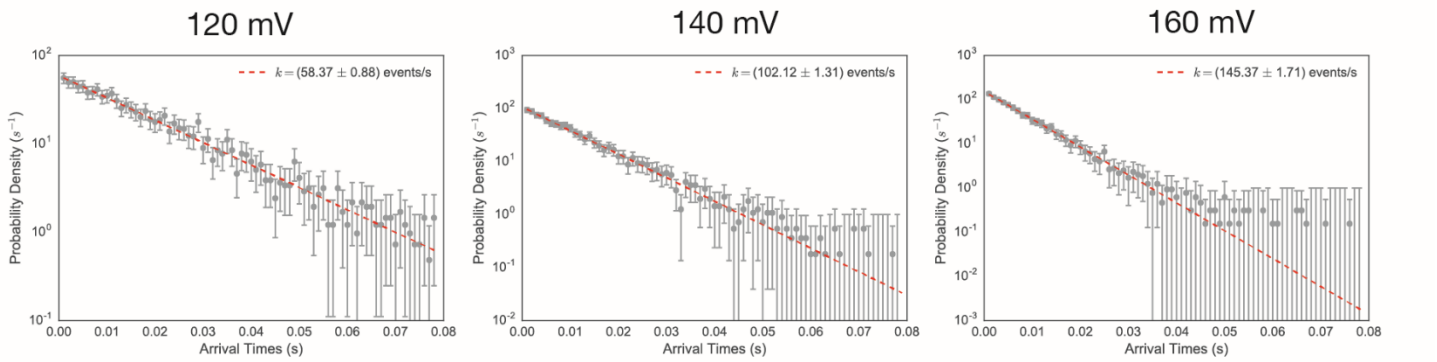


Figure S6. Blockade depth histograms for dA$_{100}$, dA$_{40}$ and dA$_{20}$ ssDNA translocation from the *cis* side of αHL nanopore at as a function of voltage. Peak fits shown were obtained using an error-weighted fit to a sum of Gaussians, as described in the text.

(D) Effect of Voltage on Capture Rate of dA_{20}

To accurately calculate the mean capture rate from the arrival times we include all events with 4 or more data points, regardless of whether the events are successfully fit by ADEPT. The arrival times (defined as the time between the start of sequential events) follow an exponential distribution. In SI Figure 7, we plot histogram of the arrival times for dA_{20} as a function of voltage and fit it to a single exponential decay.



SI Figure 7. Normalized probability distribution of arrival times (log-linear plot). Error bars indicate the standard error for Poisson counting; bins with zero counts were assigned an error of $\sigma = 1$. Mean residence times for each voltage are shown in the plot legend.

(E) SNR Analysis for PEG Measurements.

For the PEG data shown in Fig 7. of the text, both ADEPT and CUSUM+ similarly resolve peaks for polymers with n>17. The difference between the two algorithms becomes apparent for shorter polymers, where the mean residence times are substantially shorter. In SI Figure S6, we directly compare the peak signal to noise ratio of peak, defined as the logarithm of the ratio of the peak amplitude, to the residual fit error. For larger polymers (n>20) the SNR is similar, however for small species (n < 20) the SNR of the CUSUM+ peaks are on average 14.3 ± 0.3% lower than those recovered by ADEPT.

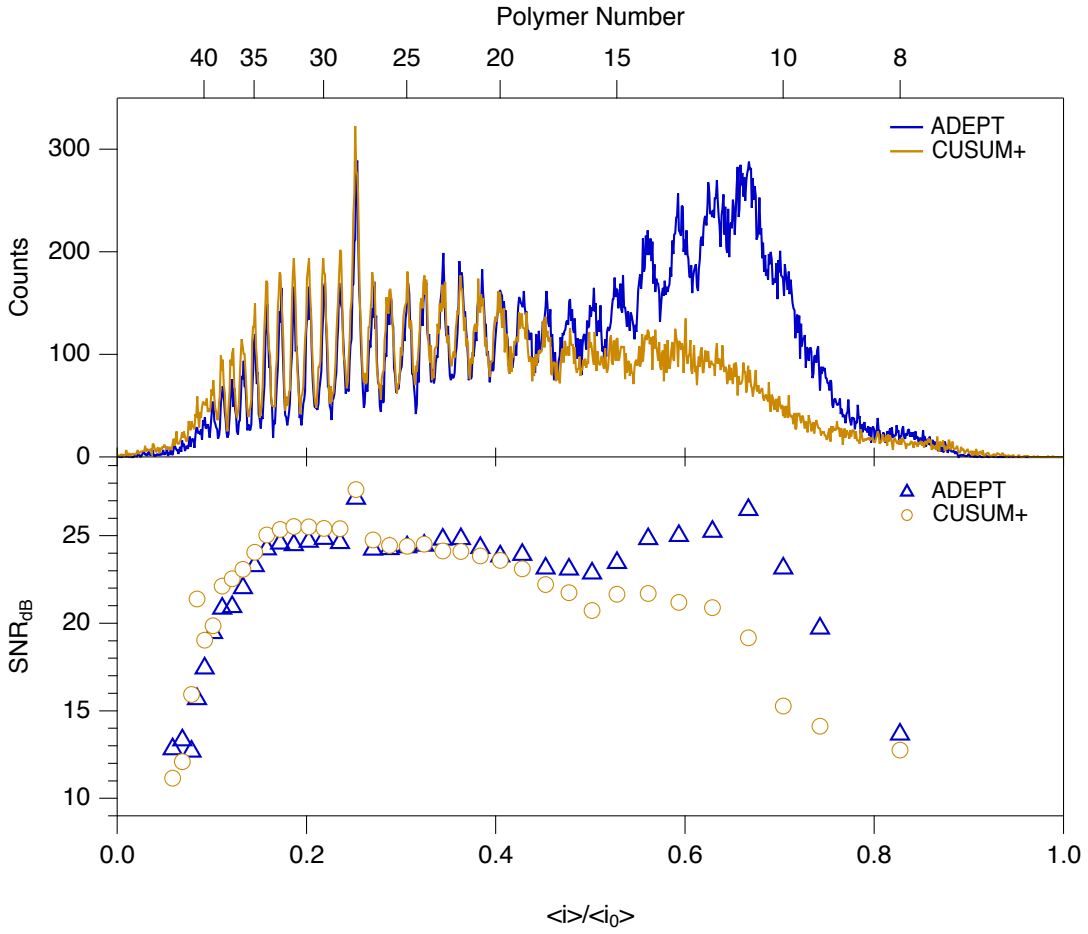


Figure S8. A comparison of ADEPT (blue) and CUSUM+(orange) analysis of a polydisperse PEG solution measured with an αHL nanopore. Blockade depth histogram of events recovered by each algorithm and comparison of signal to noise of each peak. The number of events recovered by CUSUM+ decreases sharply for small polymers (N < 20) that exhibit fast residence times (< 5τ) in the pore.

(F) Tables of Analysis Parameters Used and Fit Parameters Obtained

Parameter	ADEPT 400 mV	ADEPT 800 mV	CUSUM+ 400 mV	CUSUM+ 800 mV
Event Identification Threshold (×σ)	5.0	5.0	5.0	5.0
Min. Event Length (data points)	5	5	5	5
Block Size (s)	2.0	2.0	2.0	2.0
StepSize	9.0	9.0	9.0	18.0
Min. Threshold	–	–	0.1	0.1
Min. State Length (data points)	10	10	10	10
Max. Event Length (data points)	15000	15000	—	—
Fit Iterations	50000	50000	—	—
Fit Tolerance	1e-7	1e-7	—	—

Table S1. Parameters used in the analysis of dsDNA with a solid-state nanopore.

		$\langle i \rangle / \langle i_0 \rangle$	Amplitude
400 mV	**ADEPT**	0.070 ± 0.001	309.0 ± 15.2
		0.488 ± 0.004	11.0 ± 1.1
	CUSUM+	0.070 ± 0.001	301.6 ± 7.4
		0.486 ± 0.004	11.7 ± 1.1
800 mV	**ADEPT**	0.080 ± 0.002	86.5 ± 6.5
		0.138 ± 0.002	128.7 ± 3.9
		0.226 ± 0.018	23.1 ± 1.9
		0.560 ± 0.005	12.3 ± 0.9
	CUSUM+	0.090 ± 0.002	80.8 ± 5.7
		0.134 ± 0.010	31.3 ± 2.7
		0.250 ± 0.030	8.0 ± 0.7
		0.559 ± 0.004	21.6 ± 1.2

Table S2. Blockade depth peaks and amplitude, as a function of voltage and algorithm, for 50bp dsDNA interaction with a solid-state nanopore. Peak locations are obtained from an error-weighted fit of the data to a sum of three Gaussians; individual bin errors were assumed to follow a Poisson error distribution. Expanded uncertainty for individual fit parameters (k=2) is reported.

Parameter	dA_{100}	dA_{40}	dA_{20}
Event Identification Threshold (×σ)	5.0	5.0	5.0
Min. Event Length (data points)	5	5	5
Block Size (s)	0.5	0.5	0.5
StepSize	9	9	9
Min. State Length (data points)	5	5	5
Max. Event Length (data points)	50000	100000	10000
Fit Iterations	5000	5000	5000
Fit Tolerance	1e-7	1e-7	1e-7

Table S3. Parameters used in the analysis of ssDNA with a protein nanopore with ADEPT

		$\langle i \rangle / \langle i_0 \rangle_3$,	$\langle i \rangle / \langle i_0 \rangle_5$,	$\langle i \rangle / \langle i_0 \rangle_a$	$\langle i \rangle / \langle i_0 \rangle_b$
dA_{100}	130 mV	0.11±0.03	0.15±0.08	0.51±0.07	—
	140 mV	0.11±0.03	0.15±0.05	0.50±0.05	—
	150 mV	0.11±0.01	0.16±0.02	0.49±0.02	—
dA_{40}	130 mV	0.12±0.06	0.15±0.13	0.49±0.02	—
	140 mV	0.12±0.07	0.15±0.11	0.50±0.04	—
	150 mV	0.12±0.03	0.15±0.04	0.47±0.02	—
dA_{20}	120 mV	0.13±0.02	—	0.49±0.25	0.56±0.12
	140 mV	0.14±0.02	—	0.47±0.08	0.51±0.05
	160 mV	0.14±0.02	—	0.45±0.07	0.49±0.10

Table S4. Blockade depth peaks, as a function of voltage, for different lengths of dA interaction with αHL. Peak locations are obtained from an error-weighted fit of the data to a sum of three Gaussians; individual bin errors were assumed to follow a Poisson error distribution. Expanded uncertainty for individual fit parameters (k=2) is reported.

Parameter	ADEPT Small PEG	ADEPT Large PEG	CUSUM+ Small PEG	CUSUM+ Large PEG
Event Identification Threshold (×σ)	2.75	4.0	2.75	4.0
Min. Event Length (data points)	5	5	5	5
Block Size (s)	0.5	0.5	0.5	0.5
StepSize	—	—	3.0	3.0
Min. Threshold	—	—	2.0	2.0
Max Threshold	—	—	100.0	100.0
Max. Event Length (data points)	—	—	—	—
Fit Iterations	50000	50000	50000	50000
Fit Tolerance	1e-7	1e-7	1e-7	1e-7

Table S5. Parameters used in the analysis of PEG with an αHL nanopore

References:

(1) Lathrop, D. K.; Ervin, E. N.; Barrall, G. A.; Keehan, M. G.; Kawano, R.; Krupka, M. A.; White, H. S.; Hibbs, A. H. *J. Am. Chem. Soc.* **2010**, *132* (6), 1878–1885.

(2) Pedone, D.; Firnkes, M.; Rant, U. *Anal. Chem.* **2009**, *81* (23), 9689–9694.

(3) Press, W. H. *Numerical Recipes 3rd Edition*; Cambridge University Press, 2007.

(4) Reiner, J. E.; Kasianowicz, J. J.; Nablo, B. J.; Robertson, J. W. F. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (27), 12080–12085.

(5) Robertson, J. W. F.; Rodrigues, C. G.; Stanford, V. M.; Rubinson, K. A.; Krasilnikov, O. V.; Kasianowicz, J. J. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (20), 8207–8211.

(6) Dimitrov, V.; Mirsaidov, U.; Wang, D.; Sorsch, T.; Mansfield, W.; Miner, J.; Klemens, F.; Cirelli, R.; Yemenicioglu, S.; Timp, G. *Nanotechnology* **2010**, *21* (6), 065502.

(7) Briggs, K.; Kwok, H.; Tabard-Cossa, V. *Small* **2014**, *10* (10), 2077–2086.

(8) Song, L. Z.; Hobaugh, M. R.; Shustak, C.; Cheley, S.; Bayley, H.; Gouaux, J. E. *Science* **1996**, *274* (5294), 1859–1866.

(9) Kasianowicz, J. J.; Brandin, E.; Branton, D.; Deamer, D. W. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93* (24), 13770–13773.

(10) Muzard, J.; Martinho, M.; Mathé, J.; Bockelmann, U.; Viasnoff, V. *Biophysj* **2010**, *98* (10), 2170–2178.

(11) Mathé, J.; Aksimentiev, A.; Nelson, D. R.; Schulten, K.; Meller, A. *Proc. Natl. Acad. Sci. USA* **2005**, *102* (35), 12377–12382.

(12) Meller, A.; Nivon, L.; Brandin, E.; Golovchenko, J.; Branton, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (3), 1079–1084.