Multiple rereads of single proteins at single-amino acid resolution using nanopores

Brinkerhoff, Henry; Kang, Albert S.W.; Liu, Jingqian; Aksimentiev, Aleksei; Dekker, Cees

DOI
10.1126/science.abl4381

Publication date
2021

Document Version
Accepted author manuscript

Published in
Science

Citation (APA)
Brinkerhoff, H., Kang, A. S. W., Liu, J., Aksimentiev, A., & Dekker, C. (2021). Multiple rereads of single proteins at single-amino acid resolution using nanopores. *Science*, *374*(6574), 1509-1513. https://doi.org/10.1126/science.abl4381

Multiple re-reads of single proteins at single-amino-acid resolution using nanopores

Henry Brinkerhoff[1], Albert S. W. Kang[1], Jingqian Liu[2], Aleksei Aksimentiev[2], Cees Dekker[1,*]

[1] Department of Bionanoscience, Kavli Institute of Nanoscience, Delft University of Technology, van der Maasweg 9, 2629 HZ Delft, The Netherlands

[2]Center for Biophysics and Quantitative Biology and Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States

[*] Corresponding author; email c.dekker@tudelft.nl

Abstract

A proteomics tool capable of identifying single proteins would be important for cell biology research and applications. Here, we demonstrate a nanopore-based single-molecule peptide reader sensitive to single-amino-acid substitutions within individual peptides. A DNA-peptide conjugate was pulled through the biological nanopore MspA by the DNA helicase Hel308. Reading the ion current signal through the nanopore enabled discrimination of single-amino-acid substitutions in single reads. Molecular dynamics simulations showed these signals to result from size exclusion and pore binding. We also demonstrate the capability to 'rewind' peptide reads, obtaining numerous independent reads of the same molecule, yielding an error rate $<10^{-6}$ in single amino acid variant identification. These proof-of-concept experiments constitute a promising basis for the development of a single-molecule protein fingerprinting and analysis technology.

One-sentence summary: This study presents proof-of-concept experiments and simulations of a nanopore-based approach for linearly reading individual peptides with sensitivity to single amino acid substitutions.

Main Text

Genetic sequences are a key source of information about protein primary sequence. However, because they do not directly encode information about protein abundance or about post-translational modification and splicing of proteins, neither the DNA genome nor the RNA transcriptome fully describe the protein phenotype. A robust method for directly identifying proteins and detecting post translational modifications at the single-molecule level would greatly benefit proteomics research(*1*), enabling quantification of low-abundance proteins as well as distributions and correlations of post-translational modifications (PTMs), all at a single-cell level. Here, we provide proof-of-concept data for a nanopore-based approach that can discriminate single peptides at single amino acid sensitivity with high fidelity and potential for high throughput. Although it is not presently capable of *de novo* protein sequencing, this nanopore peptide reader provides site-specific information about the peptide's primary sequence that may find applications in single-molecule protein fingerprinting and variant identification.

Recently, biological nanopores have been used as the basis of a single-molecule DNA sequencing technology(*2*) that is capable of long reads and detection of epigenetic markers in a portable platform with minimal cost(*3*). In such experiments, single-stranded DNA is slowly moved step-by-step through a protein nanopore embedded in a thin membrane, partially blocking an electrical current carried by ions through the nanopore. The DNA stepping is accomplished using a DNA-translocating motor enzyme which moves DNA through the pore in discrete steps, yielding a series of steps in the ion current. Each ion current level characterizes the bases residing in the pore at that step, and the sequence of levels can be decoded into the DNA base sequence.

It has been hypothesized that nanopores can also be used for protein fingerprinting or sequencing(*4, 5*). Methods in which small peptide fragments freely translocate through a pore

have shown sensitivity to single amino acids(*6-8*), but we lack a method for determining the order of amino acids and reconstructing the sequence of single proteins. Using a ClpX protein unfoldase to pull a peptide through a nanopore yielded signals that effectively distinguished between different peptides(*9*), but these reads were difficult to interpret, in part due to the irregular stepping behavior of ClpX(*10*). Here, we instead applied the precise stepwise control of a DNA-translocating motor(*11-13*) to pull a peptide through a nanopore, similarly to simultaneous work by Yan et al(*14*) but presenting several key advances: the use of a helicase that pulls the polymer through MspA in smaller, half-nucleotide steps, the ability to identify single amino acid substitutions, and the capability to obtain high-fidelity signals by re-reading the same single molecule multiple times.

We developed a system in which a DNA-peptide conjugate was pulled through a biological nanopore by a helicase that was walking on the DNA section (Fig. 1). The conjugate strand consisted of an 80-nucleotide DNA strand that was covalently linked to a 26-amino acid synthetic peptide by a DBCO click linker on the 5' end of the DNA connecting to an azide modification at the C terminus of the peptide (Materials and Methods 1, Fig. S1). A negatively charged peptide sequence of mostly aspartic acid (D) and glutamic acid (E) residues was chosen so that the electrophoretic force assisted in pulling the peptide into the pore. We used the mutant nanopore M2 MspA(*15*) with a cup-like shape that separates the helicase by ~10 nm from the constriction of the pore where the blockage of ion current occurs (*16*). For the DNA-translocating motor enzyme, we used Hel308 DNA helicase (i) because it pulls single-stranded DNA through MspA in half-nucleotide ~0.33 nm observable steps(*13*), which are close to single-amino acid steps, (ii) because it is a stable and processive helicase that tolerates high salt concentrations(*16*), and (iii) its >50 pN pulling force(*16*) is likely to denature any secondary structure in target peptides.

We found that, similarly to nanopore reads of DNA, ratcheting a peptide through the nanopore generated a distinct step-like pattern in the ion current (Fig. 1C). Durations of ion current steps varied from read to read, but the sequence of levels was highly reproducible (Fig. S2). The progression of ion current steps was accurately identified using custom software (Materials and Methods 2, Fig. S3) and further analysis was performed on the sequence of the median values of ion current for each step (Fig. 1D).

This sequence of ion current levels first closely tracked the sequence expected for the template strand of DNA, which can be predicted using a DNA-sequence-to-ion-current map developed previously(*17, 18*) (Materials and Methods 3). After the end of the DNA crossed to the *cis* side of MspA's constriction, we continued to observe stepping over the linker (a length of ~2 nm, or six Hel308 steps), and subsequently over the peptide. The stepping of the peptide through the MspA constriction produced distinguishable ion current steps, much like those from DNA, but with a higher average ion current. While individual reads might contain a varying number of steps due to helicase backstepping and errors in step segmentation, we identified these features by cross-comparison of several independent reads, producing a "consensus" ion current sequence free of helicase mis-steps or step-segmentation errors (Materials and Methods 4). By counting the steps in these consensus sequence traces, we determined the parts of the traces that corresponded to the linker (the first six steps after the DNA) and the peptide (all steps thereafter) in the MspA constriction. We confirmed this analysis by altering the peptide sequence at a selected site and observing the location of the resulting change in the ion current stepping sequence, as discussed below. We restricted further analysis to reads containing both DNA and peptide sections (Materials and Methods 5, Fig. S4).

Our approach allowed us to discriminate peptide variants that differed by only a single amino acid.

We obtained reads (N=211) of three different DNA-peptides in nineteen different pores, where the peptide sequences consisted of a mixture of negatively D and E residues, with a single variation, i.e., D, glycine (G), or tryptophan (W), placed four amino acids away from the C-terminus that connected to the linker (Table S1 for full sequences). The three variants showed a reproducible difference at the site of the substituted amino acid, which could be seen by comparing the consensus sequences of ion current levels (Figs. 2A and B). As is typical of nanopore experiments, a single-site variation was found to affect several ion current steps, because an "8-mer" of amino acids around the pore constriction of MspA affect the ion current blockage level (*11, 17*) due to the finite constriction height and stochastic displacements of the strand up and down through the nanopore(*19*). The center of the differing region in the ion current sequence was at the expected site: about 10 helicase steps away from the end of the DNA section (6 half-nucleotide steps for the linker and 4 more along the peptide to the variant site). The signals varied by several standard deviations over multiple sequential levels, demonstrating that variations as small as a single amino acid substitution could be resolved. The differences of the ion currents for the W- and G-substituted variants from the D-substituted variant (Fig. 2B) showed an interesting behavior: when G, which has merely a hydrogen atom as a side chain, occupied the nanopore constriction, we saw higher ion current levels, as expected from a smaller amino acid volume. But when the bulky W variant moved through the constriction, the ion current first decreased and then, counterintuitively, increased relative to the medium-sized D variant.

To understand the origin of these patterns, we performed all-atom molecular dynamics simulations measuring the ion current with peptide variants at varying positions within the MspA constriction. In a typical simulation, a polypeptide chain was threaded through a reduced-length model of MspA nanopore that was embedded in a lipid bilayer and surrounded by 0.4M KCl electrolyte (Fig. 2C). Peptides with either one W or G substitution in a mixed D/E sequence were

examined under a +200 mV bias at various locations relative to the MspA constriction (see Materials and Methods 6 and Figs. S5-S8 for details). Patterns of ionic current blockades resulted in Fig. 2E (top panel), matching the counterintuitive blockade current patterns that were experimentally measured for G and W substitutions (cf. Fig. 2A,B). Furthermore, the ion current correlated with the nanopore constriction volume that was available for ion transport near the pore mouth, Fig. 2E (bottom panel), with the latter quantity being more accurately characterized by the all-atom MD method (*19*). In the case of a G residue, its upward motion was accompanied by an increase of the nanopore volume (Fig. 2E, bottom), that subsided as the residue left the nanopore constriction (Fig. 2F), in sync with the blockade current (Fig. 2E, top). A W residue, however, reduced the nanopore constriction volume when it was located below the constriction (Fig. 2E, top), but increased the volume at and above the constriction. The latter counterintuitive effect could be traced back to a binding of the W side chain to the nanopore surface above the constriction (Fig. 2G). Thus, a glycine substitution merely increases the nanopore volume as the residue passes through the constriction, whereas the tryptophan residue decreases the volume when its side chain enters the constriction, and subsequently increases the volume when its side chain binds to the inner nanopore surface (Fig. S9).

To quantitatively assess the distinguishability of peptide variants, we computed a so-called confusion matrix (Fig. 2D). Using a hidden Markov model, we quantified the relative likelihoods of the alignments to the three consensus sequences for 119 reads withheld from the consensus sequence generation, finding that we could identify the correct variant with an average of 87% accuracy (Materials and Methods 7). This high rate of correct single substitution identification compares favorably to early nanopore experiments, that identified single-nucleotide variants with significantly lower accuracy(*17*). Still, the limited single-read accuracy is an ongoing challenge in developing nanopore sequence analysis approaches, requiring the implementation of strategies

to increase sequencing fidelity to acceptable levels (*18, 20*). The largest error modes in nanopore reads are due to random effects as enzymes step stochastically both forwards and backwards, and sometimes step too quickly to be clearly resolved, resulting in incorrect step identifications. In DNA sequencers, this random error is typically addressed by obtaining 20x coverage or more, averaging many independent reads of different molecules. However, for a truly single-molecule technology, single-read accuracy is essential.

The identification fidelity of our nanopore protein reader can be greatly increased by obtaining many independent re-readings of the same individual molecule with a succession of controlling helicases, eliminating the random errors that lead to inaccuracies in nanopore reads. At a very high concentration of helicase, on the order of 1 μM, the DNA in the pore nearly always had a second helicase queued up behind the one controlling its motion (Fig. 3A)(*21*). When the first helicase reached the linker at the end of the DNA section, it could no longer process the molecule and subsequently fell off. The DNA-peptide conjugate was then immediately pulled back into the nanopore such that the queued helicase, which was still bound to the DNA, took control as the new DNA-pulling enzyme. This 'rewound' the system and initiated a new and independent read of the peptide. The numbers of re-reads on the same single peptide can be very large: Fig. 3A shows an example of a raw data trace with 117 re-reads on a single peptide containing the G-substitution. This event was purposefully ended by the reversal of voltage to eject the DNA-peptide conjugate from the pore. We observed a typical rewinding distance of approximately 17 helicase steps, commensurate with a rewinding by a distance of ~17 amino acids, a number that is consistent with the ~9 DNA bases that are bound within the controlling helicase(*16*). Of the 117 re-reads in Fig. 3B, 45 re-reads stepped back far enough to provide a re-read of the variant site.

We observed significant improvement of the read accuracy with an increasing number of re-reads

(Fig. 3C). To quantify the increase in the accuracy of the readings as a function of the number of re-readings, we randomly chose subsets of the 45 measured re-reads and computed the identification accuracy using *N* re-reads as the fraction of subsets containing *N* re-reads that yielded the correct consensus identification (Materials and Methods 8). Even when single reads were limited to as low as ~50% identification accuracy due to only partial coverage of the variant site, the re-reading method allowed single molecules to be identified at high levels of confidence. As the inset to Fig.3C shows, the error rate decreased with the number of re-reads, yielding an undetectably low error rate (< 1 in 10^6) when using more than ~30 re-reads of an individual peptide. Analysis on re-read traces from other variants yielded similar results (Fig. S10).

The method described here provides an approach for reading single proteins with sensitivity to single-amino-acid changes, which is particularly powerful because of the re-reading mode of operation that reduces the stochastic error. Transforming this into a technology capable of *de novo* protein sequencing remains a substantial challenge. With any of 20 amino acids at each position along the protein sequence and a read-head width(*17*) of ~8 amino acids, the number of measurements required to build an ion-current-to-amino-acid map is impractically large. However, many proteomics applications do not require *de novo* sequencing, but instead concern other forms of sequence analysis that rely on a priori knowledge of candidate sequences before decoding. These include identifying or "fingerprinting" proteins even in heterogeneous mixtures, mapping post-translational modifications, and measurements of small samples, which all involve comparing single-molecule measurements to reference signals of known proteins and interesting variants.

Our methodology has several limitations, but these may be addressed experimentally. While the pore is capable of translocating heterogeneously charged peptides with neutral polar, nonpolar,

negative, and positive amino acids (Supplemental Text 1; sample reads shown in Fig. S11), highly positively charged peptides may not efficiently be translocated through the pore. Fortunately, analysis of the human proteome reveals that negatively charged stretches of protein sequence are more common than positively charged stretches(*22*), particularly in alkaline pH conditions like those used in our experiments. If needed, the MspA pore can be engineered to provide stronger electro-osmotic forces, which can exceed electrophoretic forces and translocate analytes regardless of charge(*7, 23*). The read length intrinsic to the technique, approximately 25 amino acids depending on the length of the DNA-peptide linker, does allow application of this method to many biologically relevant short peptides, such as 8-12 amino acid MHC-binding peptides(*24*). Additionally, this finite read length still represents an improvement over the <10 amino acid long peptide fragments used in mass spectrometry(*25*), and protein fragmentation and shotgun sequencing methods similar to those used in traditional protein sequencing can naturally be applied to this new technique. Technical modifications such as using a variable-voltage control scheme(*18*) have been shown to improve the accuracy of DNA sequencing, and the physical principle of this is equally applicable to peptide sequencing (Supplemental Text 2 and Fig. S12).

Reads of DNA-peptide conjugates like those presented here could be measured in high throughput with any existing commercially available nanopore sequencing hardware capable of accommodating MspA (e.g. the commercial MinION system) without requiring any re-engineering of the device, changing only the sample preparation and data analysis. Furthermore, our methodology retains the features that enabled the success of nanopore DNA sequencing: low overhead cost, physical rather than chemical sensitivity to small changes in single molecules, and the flexibility to be re-engineered to target specific applications. Overall, our findings comprise a promising first step towards a low-cost method capable of single-cell proteomics at the ultimate limit of sensitivity to concentration, with a wide range of applications in both fundamental biology

and the clinic.

Acknowledgements:

We thank Prof. Jens Gundlach and his lab members at University of Washington for providing the MspA nanopore and for sharing key pieces of software, and we thank Foteini Mentzou and Xin Shi for assistance with data collection, Eli van der Sluis for Hel308 purification, and Jaco van der Torre for his helpful advice on DNA construct preparation. We also acknowledge supercomputer time on the Blue Waters at UIUC, Expanse at UCSD and Frontera at TACC.

Funding:

Dutch Research Council (NWO) NWO-I680 (SMPS) (CD)

Dutch Research Council (NWO) / Ministry of Education, Culture and Science (OCW) Gravitation programs NanoFront (CD)

European Research Council Advanced Grant 883684 (CD)

European Commission Marie Skłodowska-Curie action Individual Fellowship 897672 (HB)

European Molecular Biology Organization Short-Term Fellowship 8968 (AK)

National Institutes of Health grant R21-HG011741 (AA)

Extreme Science and Engineering Discovery environment allocation MCA05S028 (AA)

Leadership Resource Allocation MCB20012 on Frontera of the Texas Advanced Computing Center (AA)

Author Contributions:

H.B. and C.D. conceived of the protein analysis method. H.B. and A.K. conducted nanopore experiments and analyzed data. H.B. developed additional analysis code. J. L. and A. A. designed and conducted MD simulations. All authors discussed experimental findings and co-wrote the manuscript.

Competing Interests:

TU Delft has filed a patent application (PCT/NL2020/050814) on technologies described herein, with H.B. and C.D. listed as inventors.

Data and Materials Availability:

All data and custom code used in this paper are available for download online(*26*).

Supplementary Materials

Materials and Methods

Figs. S1 to S13

Table S1

References (27-40)

Figure Captions

Fig. 1: Reading peptides with a nanopore. (A) The DNA-peptide conjugate consists of a peptide (pink) attached via a click linker (green) to an ssDNA strand (black). This DNA-peptide conjugate is extended with a typical nanopore adaptor comprised of an extender that acts as a site for helicase loading (blue) and a complementary oligo with a 3' cholesterol modification (gold). (B) The cholesterol associates with the bilayer as shown in (a), increasing the concentration of analyte near the pore. The complementary oligo blocks the helicase, until it is pulled into the pore (b), causing the complementary strand to be sheared off (c), whereupon the helicase starts to step along DNA. (C) As the helicase walks along the DNA, it pulls it up through the pore, resulting in (a) a read of the DNA portion followed by (b) a read of the attached peptide. (D) Typical nanopore read of a DNA-peptide conjugate (black), displaying step-like ion currents (identified in red). The asterisks * indicate a spurious level not observed in most reads and therefore omitted from further analysis. The dagger † indicates a helicase backstep. (E) Consensus sequence of ion current steps (red), which for the DNA section is closely matched by the predicted DNA sequence (blue). The linker and peptide sections are identified by counting half-nucleotide steps over the known structural length of the linker. Error bars in the measured ion current levels are errors in the mean value, often too small to see. Error bars in the prediction are standard deviations of the ion current levels that were used to build the predictive map in previous work(*18*).

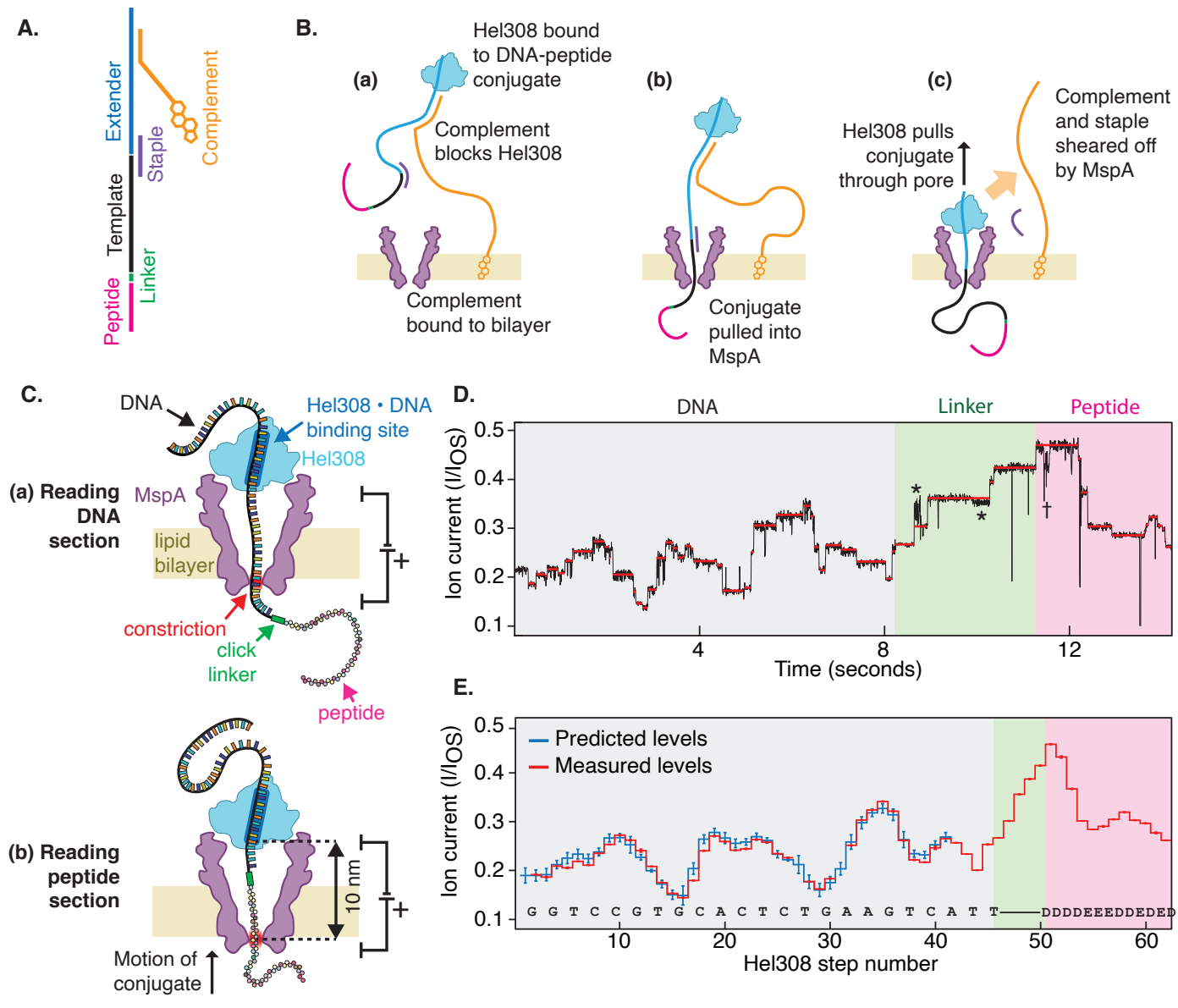
Fig. 2: Detection of single amino acid substitutions in single peptides. (A) Consensus ion current sequences for each of the three measured variants (D, gold; W, red; G, blue), which differ significantly at the site of the amino acid substitution. (B) Difference in ion current between the W (red) and G (blue) variants and the D variant. Error bars are standard deviations. (C) Confusion matrix showing error modes of a blind classifier in identifying variants of reads, demonstrating an 87% single-read accuracy. (D) All-atom model where a reduced-length MspA pore (grey) confines a polypeptide chain (Glu: green, Asp: light blue; Cys: beige). The top end of the peptide is anchored using a harmonic spring potential, representing the action of the helicase at the rim of a full-length MspA. Water and ions are shown as semitransparent surface and spheres, respectively. (E) Top: Ionic current in MspA constriction versus z coordinate of the mutated residue backbone from MD simulations. Bottom: Fraction of nanopore construction volume available for ion transport. Vertical and horizontal error bars denote standard errors and standard deviations, respectively. (F,G) Representative molecular configurations observed in MD simulations of peptide variants. Glycine and tryptophane residues are shown in dark blue and red, respectively. Significant peptide/pore surface interactions are observed.

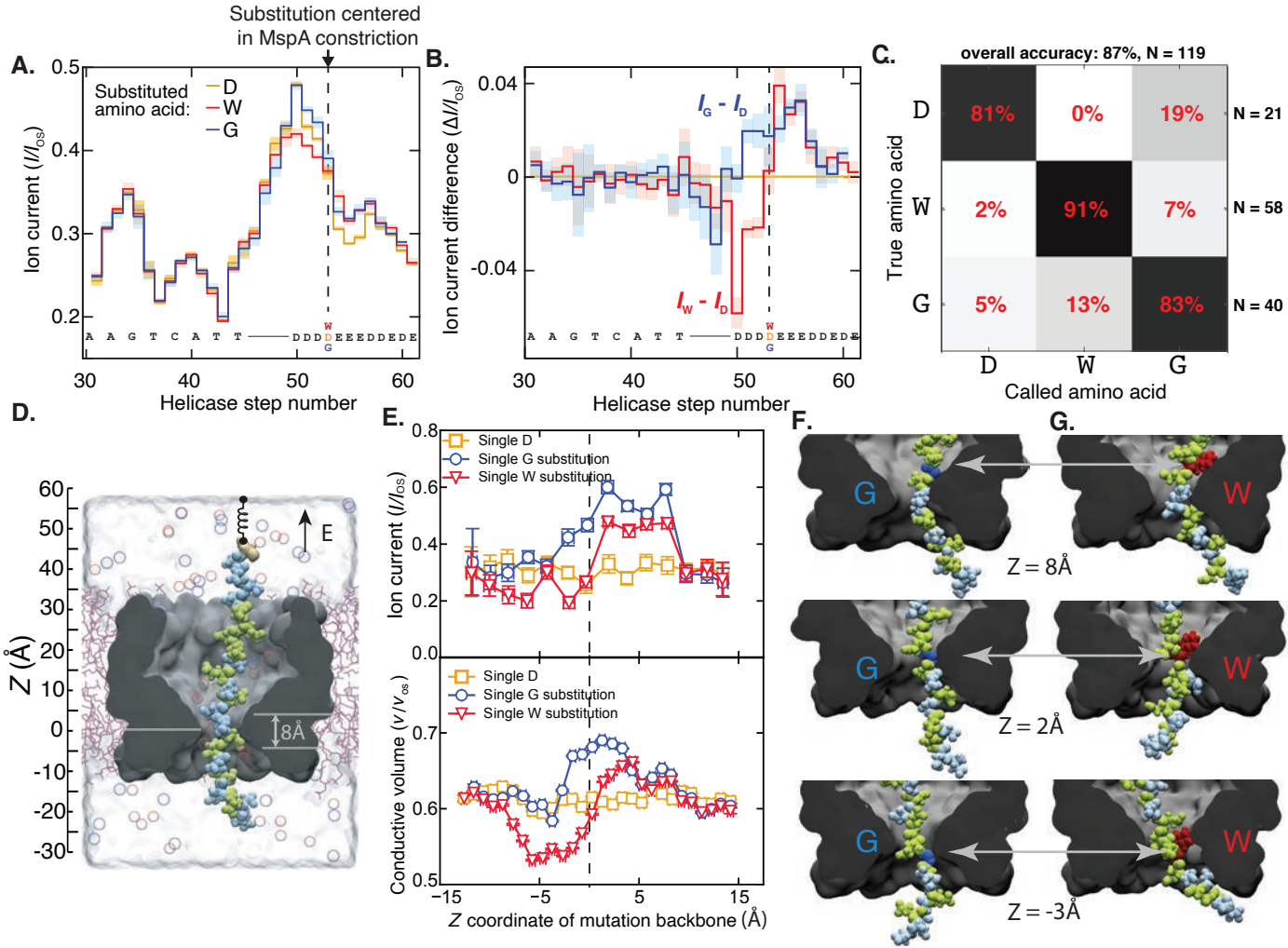
Fig. 3: Re-reading of a single peptide. (A) Highly repetitive ion current signal corresponding to numerous re-reads of the same section of an individual peptide (in this case, the G-substituted variant). The expanded plot below shows a region that contains four rewinding events (red dashed lines), where the trace jumps back to level 52 ± 2 of the consensus displayed in Fig. 2A. (B) Re-reading is facilitated by helicase queueing, where (a) a second helicase binds behind the primary helicase that controls the DNA-peptide conjugate, re-reading starts when (b) the primary helicase dissociates, and (c) the secondary one becomes the primary helicase that drives a new round of reading. (C) By using information from multiple re-reads of the same peptide, the identification accuracy can be raised to very high levels of fidelity. These results indicate that with sufficient numbers of re-reads, random error can be eliminated and single-molecule error rate can be pushed lower than 1 in 10^6 even with poor single-pass accuracy. Inset is a logarithmic plot of the error rate = 1 - accuracy.

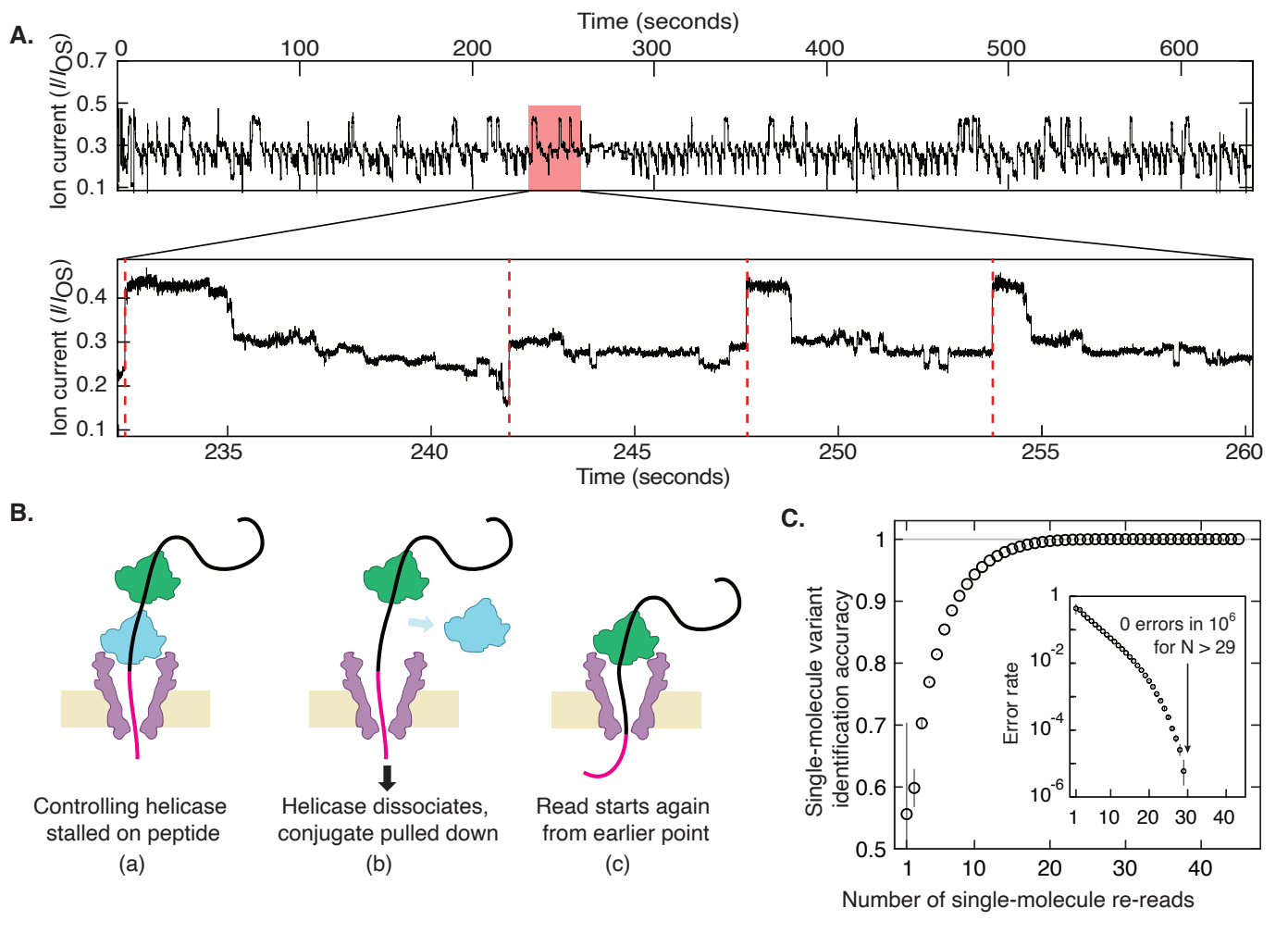
References and Notes

1. J. A. Alfaro *et al.*, The emerging landscape of single-molecule protein sequencing technologies. *Nature Methods* **18**, 604-617 (2021).
2. J. J. Kasianowicz, E. Brandin, D. Branton, D. W. Deamer, Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences* **93**, 13770 (1996).
3. D. Deamer, M. Akeson, D. Branton, Three decades of nanopore sequencing. *Nature Biotechnology* **34**, 518-524 (2016).
4. J. Nivala, D. B. Marks, M. Akeson, Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nature Biotechnology* **31**, 247 (2013).
5. D. Rodriguez-Larrea, H. Bayley, Multistep protein unfolding during nanopore translocation. *Nature Nanotechnology* **8**, 288-295 (2013).
6. F. Piguet *et al.*, Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore. *Nature Communications* **9**, 966 (2018).
7. L. Restrepo-Pérez, C. H. Wong, G. Maglia, C. Dekker, C. Joo, Label-Free Detection of Post-translational Modifications with a Nanopore. *Nano letters* **19**, 7957-7964 (2019).
8. H. Ouldali *et al.*, Electrical recognition of the twenty proteinogenic amino acids using an aerolysin nanopore. *Nature Biotechnology* **38**, 176-181 (2020).
9. J. Nivala, L. Mulroney, G. Li, J. Schreiber, M. Akeson, Discrimination among Protein Variants Using an Unfoldase-Coupled Nanopore. *Acs Nano* **8**, 12365-12375 (2014).
10. Juan C. Cordova *et al.*, Stochastic but Highly Coordinated Protein Unfolding and Translocation by the ClpXP Proteolytic Machine. *Cell* **158**, 647-658 (2014).
11. E. A. Manrao *et al.*, Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature Biotechnology* **30**, 349 (2012).
12. G. M. Cherf *et al.*, Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nature Biotechnology* **30**, 344 (2012).
13. I. M. Derrington *et al.*, Subangstrom single-molecule measurements of motor proteins using a nanopore. *Nature Biotechnology* **33**, 1073 (2015).
14. S. Yan *et al.*, Single Molecule Ratcheting Motion of Peptides in a Mycobacterium smegmatis Porin A (MspA) Nanopore. *Nano Lett* **21**, 6703-6710 (2021).
15. T. Z. Butler, M. Pavlenok, I. M. Derrington, M. Niederweis, J. H. Gundlach, Single-molecule DNA detection with an engineered MspA protein nanopore. *Proceedings of the National Academy of Sciences* **105**, 20647 (2008).
16. J. M. Craig *et al.*, Determining the effects of DNA sequence on Hel308 helicase translocation along single-stranded DNA using nanopore tweezers. *Nucleic Acids Research* **47**, 2506-2513 (2019).
17. A. H. Laszlo *et al.*, Decoding long nanopore sequencing reads of natural DNA. *Nature Biotechnology* **32**, 829 (2014).
18. M. T. Noakes *et al.*, Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage. *Nature Biotechnology* **37**, 651-656 (2019).
19. S. Bhattacharya, J. Yoo, A. Aksimentiev, Water Mediates Recognition of DNA Sequence via Ionic Current Blockade in a Biological Nanopore. *ACS Nano* **10**, 4644-4651 (2016).
20. A. D. Tyler *et al.*, Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Scientific Reports* **8**, 10931 (2018).
21. A. C. Rand, Doctoral Thesis. University of California, Santa Cruz, Santa Cruz, CA, United States (2017).
22. R. D. Requião *et al.*, Protein charge distribution in proteomes and its impact on translation. *PLOS Computational Biology* **13**, e1005549 (2017).
23. D. P. Hoogerheide, P. A. Gurnev, T. K. Rostovtseva, S. M. Bezrukov, Mechanism of α-synuclein translocation through a VDAC nanopore revealed by energy landscape modeling of escape time distributions. *Nanoscale* **9**, 183-192 (2017).

24. M. Wieczorek *et al.*, Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Frontiers in Immunology* **8**, (2017).
25. J. Cox, N. C. Hubner, M. Mann, How Much Peptide Sequence Information Is Contained in Ion Trap Tandem Mass Spectra? *Journal of the American Society for Mass Spectrometry* **19**, 1813-1820 (2008).
26. H. Brinkerhoff, A. S. W. Kang, J. Liu, A. Aksimentiev, C. Dekker. Code and Data for "Multiple re-reads of single proteins at single amino acid resolution using nanopores" [Data set]. Zenodo (2021).
27. P. A. Wiggins, An information-based approach to change-point analysis with applications to biophysics and cell biology. *Biophys J* **109**, 346-354 (2015).
28. J. C. Phillips *et al.*, Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics* **153**, 044130 (2020).
29. K. Hart *et al.*, Optimization of the CHARMM additive force field for DNA: Improved treatment of the BI/BII conformational equilibrium. *J Chem Theory Comput* **8**, 348-362 (2012).
30. J. B. Klauda *et al.*, Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J Phys Chem B* **114**, 7830-7843 (2010).
31. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926-935 (1983).
32. D. Beglov, B. Roux, Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations. *The Journal of Chemical Physics* **100**, 9050-9063 (1994).
33. J. Yoo, A. Aksimentiev, New tricks for old dogs: improving the accuracy of biomolecular force fields by pair-specific corrections to non-bonded interactions. *Physical Chemistry Chemical Physics* **20**, 8432-8449 (2018).
34. H. C. Andersen, Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics* **52**, 24-34 (1983).
35. S. Miyamoto, P. A. Kollman, Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* **13**, 952-962 (1992).
36. T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics* **98**, 10089-10092 (1993).
37. G. J. Martyna, D. J. Tobias, M. L. Klein, Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics* **101**, 4177-4189 (1994).
38. A. T. Brünger, Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472-475 (1992).
39. A. Aksimentiev, K. Schulten, Imaging alpha-hemolysin with molecular dynamics: ionic conductance, osmotic permeability, and the electrostatic potential map. *Biophys J* **88**, 3745-3761 (2005).
40. J. Gumbart, F. Khalili-Araghi, M. Sotomayor, B. Roux, Constant electric field simulations of the membrane potential illustrated with simple systems. *Biochim Biophys Acta* **1818**, 294-302 (2012).


A.
Extender
Staple
Complement
Template
Linker
Peptide
B.
(a)
Hel308 bound to DNA-peptide conjugate
Complement blocks Hel308
Complement bound to bilayer
(b)
Conjugate pulled into MspA
(c)
Hel308 pulls conjugate through pore
Complement and staple sheared off by MspA
C.
DNA
Hel308 • DNA binding site
Hel308
(a) Reading DNA section
MspA
lipid bilayer
constriction
click linker
peptide
(b) Reading peptide section
10 nm
Motion of conjugate
D.
DNA
Linker
Peptide
Ion current (I/IOS)
Time (seconds)
E.
Predicted levels
Measured levels
Ion current (I/IOS)
G G T C C G T G C A C T C T G A A G T C A T T——DDDDEEEDDEDED
Hel308 step number


A.
Substitution centered in MspA constriction
Substituted amino acid:
D
W
G
Ion current (I/I_OS)
Helicase step number
B.
Ion current difference (ΔI/I_OS)
I_G - I_D
I_W - I_D
Helicase step number
C.
overall accuracy: 87%, N = 119
True amino acid
Called amino acid
81%
0%
19%
N = 21
2%
91%
7%
N = 58
5%
13%
83%
N = 40
D.
Z (Å)
E
8Å
E.
Single D
Single G substitution
Single W substitution
Ion current (I/I_OS)
Conductive volume (v/v_os)
Z coordinate of mutation backbone (Å)
F.
G.
G
W
Z = 8Å
Z = 2Å
Z = -3Å


A.
Time (seconds)
0
100
200
300
400
500
600
Ion current (I/IOS)
0.7
0.5
0.3
0.1
Ion current (I/IOS)
0.4
0.3
0.2
0.1
235
240
245
250
255
260
Time (seconds)
B.
Controlling helicase stalled on peptide
(a)
Helicase dissociates, conjugate pulled down
(b)
Read starts again from earlier point
(c)
C.
Single-molecule variant identification accuracy
1
0.9
0.8
0.7
0.6
0.5
1
10
20
30
40
Number of single-molecule re-reads
Error rate
1
10-2
10-4
10-6
0 errors in 10^6 for N > 29
1
10
20
30
40

Science
AAAS

Supplementary Materials for

Multiple re-reads of single proteins at single amino acid resolution using nanopores

Henry Brinkerhoff, Albert S. W. Kang, Jingqian Liu, Aleksei Aksimentiev, Cees Dekker

Correspondence to: c.dekker@tudelft.nl

This PDF file includes:

Materials and Methods
Supplementary Text
Figs. S1 to S13

Other Supplementary Materials for this manuscript include the following:

MDAR Reproducibility Checklist
Table S1

Materials and Methods

Nanopore experiments were carried out as in previous work *(11, 16-18)*, on custom U-tube nanopore experimental devices. Experimental buffer consisted of 400 mM KCl, 10 mM $MgCl_2$, and 10 mM HEPES free acid at pH 8.00 $\pm$ 0.02. To initiate reading, Hel308 was added to a concentration of 150 nM and ATP was added to a concentration of 1000 μM. MspA was a kind gift from the laboratory of Jens Gundlach at the University of Washington, originally expressed by Genentech. Hel308 plasmid was obtained from Genscript (Cat. No. SC1849), and was expressed in-house using standard techniques. DNA-peptide conjugates and DNA oligos were obtained from Biomers. DPhPC lipid suspended in chloroform was obtained from Avanti. All experiments were performed at room temperature (21±1 ºC).

Nanopore ion current was recorded at 50 kHz sampling frequency with an Axopatch 200B patch clamp amplifier, and filtered with a 10 kHz 4-pole Bessel filter. Experiments were controlled through a National Instruments X series DAQ and operated with custom LabVIEW software. Data analysis was performed in Matlab. Preprocessing, data reduction and filtering, alignment and variant identification were performed using custom Matlab software described below and in previous work*(11,16-18)*.

Measured levels (red) in main text Figure 1D,E were identified by hand. Predicted levels (blue) were drawn from a 6-mer map of base sequence to DNA developed in previous work*(18)*. The highlighted linker and peptide sections were identified based on the length of the DBCO linker estimated from its chemical structure, as well as the consensus reads shown in figure 2A, where the variation resulting from the substitution determines the location of the substitution site.

The construction of the consensus reads in main text Figures 2A and B is described in Materials and Methods 4. Main text Figure 2C was generated by choosing the maximum likelihood variant based on a hidden Markov model alignment to each of the three variant consensuses, with the percentage calculated as (number of reads of variant X identified as variant Y)/(total number of reads of variant X), such that each row of the matrix sums to 100%. MD simulation methods used to generate main text Figures 2D-H are described in full in Materials and Methods 6.

To reliably obtain re-reads, helicase concentration was increased to ≈ 1 μM. The segmentation and identification of re-reads used to generate the accuracy values in main text Figure 3C is described fully in Materials and Methods 8.

1. DNA-peptide hybrid construct design and assembly

The DNA-peptide hybrid constructs (main text Figure 1A) used to collect the bulk of the data were constructed of four components:

1. The template (variants “D22-”, “W22-”, “G22-”, and “hetero template” in Table S1),

which consists of a 30-base nucleotide sequence attached at the 5' end to the C-terminus of a 25-amino acid peptide by an azide-DBCO-C5 linker (Figure S1A). This strand is pulled into the nanopore electrophoretically, and is read by the sequencer.

2. The complement ("complement gen2" in Table S1), which is complementary to part of the template strand and serves three functions: (a) a 3' cholesterol allows it to associate with the bilayer, increasing the frequency of DNA-pore interactions; (b) a 5' overhang provides a sticky end to attach the template extender; and (c) on the hybridized construct, it blocks the Hel308 enzyme (which has poor helicase processivity) from processing along the template and using ATP, until the template enters the pore and the complement is sheared off by MspA.

3. The 50-base template extender ("template extender gen2"), which binds to the sticky 5' end of the complement, and extends with its own 10-base 3' sticky end which acts as a binding site for Hel308. The ligation of this extender is necessary to increase the length of the DNA-peptide hybrid, which is only commercially available in lengths too short to be efficiently captured by the pore.

4. The staple ("staple"), a 10-base oligo complementary to both the template and the extender, which enables the efficient ligation of the two oligos. The staple is used to prepare the construct, but once assembled has no functionality in the construct, and like the complement is sheared off upon capture by MspA.

The exact sequences used are provided in Table S1. To assemble the constructs, equal amounts of template, staple, and template extender were mixed and annealed in a thermocycler by heating to 95 ºC for 2 min and letting them cool down slowly to room temperature. The mixture was then incubated for 18h at 16 ºC with 400U of T4 DNA ligase (NEB, M0202T) in 1X of the manufacturer-provided buffer solution. Next, a 1.1X excess of the ligated construct was mixed with the complement at > 1 μM concentration of each and annealed.

Some reads (those labeled "biomers1cholext" as opposed to "biomers1cholext2" or "biomers1cholext2nohp" in the Supplementary Data *(26)*) used an earlier version of the construct, in which the complementary strand itself ("complement gen1" in Table S1) was used to assemble and ligate the template to an extender ("template extender gen1"). This construct is illustrated in Figure S1B. This resulted in both a shorter free length of ssDNA/peptide lowering the rate of template capture by MspA, and the longer cholesterol tether being frequently captured and occupying the pore while excluding the template. These negative impacts on throughput led us to the revised construct described above. The end of the DNA sequence and the peptide sequence read in this earlier generation was identical to that in the later reads, so the signal from the region of interest discussed in the paper was unchanged.

2. Ion current level identification and filtering

To segment the data for consensus refinement through expectation maximization and for blinded variant identification, we used a change point detection algorithm exactly as described in previous work (Noakes 2019 Supplemental Information §11 *(18)*; originally described in Wiggins 2015 *(27)*). A sample trace with automated level finding indicated is shown in Figure S3.

The further filtering steps described in Noakes 2019 *(18)* were also applied to the resultant level sequences. First, a state filter was applied to excise levels that were too short (< 2 ms) or significantly outside the bounds of the consensus currents ($I/I_{\mathrm{OS}} < 0.25$ or $I/I_{\mathrm{OS}} > 0.5$). These filters serve to eliminate a significant number of spurious states resulting from noise spikes or mid-event MspA gating. Next, a backstep recombination filter was applied using the algorithm of Noakes 2019, Supplementary Information §5.2 *(18)*, in order to eliminate the bulk of helicase backsteps. The recombination filter, which relies on comparing levels in an event to other nearby levels in the same event, is more accurate than accounting for a large number of backsteps in the alignment algorithm, because read-to-read error, which can impact the scoring of an alignment to reference, does not affect the matching of observed states in a self-comparison.

3. DNA ion current prediction

Following previous work*(18)*, ion currents for the DNA section in main text Figure 1E were predicted using an empirically derived 6-mer map, converting each 6-base subsequence into "pre-" and "post-" ion current states corresponding to the two substeps of Hel308 per DNA base.

The construction of the 6-mer map from measurements of genomic DNA is described in Noakes 2019, Supplementary Information §7 *(18)*. Briefly, ion current measurements corresponding to each 6-mer pre- and post-state were obtained from kilobase or longer reads of genomic λ phage and Φ X174 viral DNA. The ion currents in the map are the mean of the set of ion currents assigned to each state, and the uncertainty in the ion current is the standard deviation in that set of ion currents.

4. Consensus generation

Consensus reads were generated through a customized Baum-Welch algorithm, a type of expectation maximization (EM) for the hidden Markov model. The EM algorithm, described fully in previous work *(18)* (Noakes 2019, Supplement §7.4) is as follows:

1. Solve the hidden Markov model using a maximum-*a posteriori* likelihood (MAP) algorithm to assign likelihoods that each of the ion current levels in each read were produced by a particular true template position within the constriction (helicase step number).

2. If the change in log likelihood of the HMM solution is greater in magnitude than a threshold (in our case 10^{-3}), continue. Otherwise, reject the latest consensus sequence and terminate.

3. Compute a new mean and uncertainty in ion current value for each HMM state using an average of the measured values weighted by the probability that each value was assigned to that state.

The EM algorithm requires an initial guess at an HMM in order to begin. To seed the EM algorithm with an initial set of HMM observation probabilities, a selection of typical reads of each construct were cross-compared by eye to identify the unique ion current states and put them in the correct order, while eliminating single-read errors like spurious states, missed states, or enzyme backsteps. The result was a set of aligned sequences of levels, where each level is characterized by a mean, standard deviation and number of measurements included.

Different nanopore reads may vary by an overall scale in ion current due to variations in buffer salt concentration caused by evaporation and due to day-to-day variations in temperature*(17)*. Therefore, reads must always be calibrated by applying an appropriate scale m to all ion currents they contain. To find the maximum likelihood estimators for m for all N aligned sets of reads of length L, we want to minimize the total error between reads

$$\hat{\boldsymbol{m}} = \arg\min_{\boldsymbol{m}} \sum_{k=1}^{L} \sum_{i,j=1}^{N} \begin{cases} \frac{(m_i x_{ik} - m_j x_{jk})^2}{\delta x_{ik}^2 + \delta x_{jk}^2} & \text{if } x_{ik} \text{ and } x_{jk} \text{ both exist,} \\ 0 & \text{otherwise.} \end{cases}, \tag{1}$$

where x_{ik} and δx_{ik} are the mean and uncertainty in the mean of state k in read i, and we have made the approximation that the uncertainties do not scale with the ion currents when the near-unity calibration is applied. However, this optimization still leaves one degree of freedom: the sum is invariant if every read is subject to the same overall scale. To eliminate these ambiguities, we choose by convention to also include the requirement that $\frac{1}{N}\sum_i m_i = 1$. We end up with a full rank linear system of equations, which can be easily solved for $\hat{\boldsymbol{m}}$. The scales are applied to the reads, and a consensus mean, standard deviation, and uncertainty are then computed as

$$\bar{x}_k = \frac{1}{N} \sum_{i=1}^{N} x_{ik}, \tag{2}$$

$$\bar{\sigma}_k = \frac{1}{N} \sum_{i=1}^{N} \sigma_{ik}, \tag{3}$$

$$\delta \bar{x}_k = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \delta x_{ik}^2 + \frac{1}{N} \sum_{i=1}^{N} (x_{ik} - \bar{x}_k)^2} \tag{4}$$

where x_{ik} now refers to an element of a calibrated read.

Next, to ensure cross-construct calibration consistency, the same procedure is carried out to find an optimal scale and offset for each of the three handmade consensuses using only the DNA section of the reads. Since the DNA section is known to be identical across different reads, we replace the DNA section in each consensus with its mean across all three variants. The means and standard deviations of the three calibrated consensuses are used as initial guesses for the EM algorithm.

Calibration scales also need to be found for every read, including those not used in the handmade consensus generation. These reads were calibrated straightforwardly by choosing a scale such that the ion current of level 43 (see Figure S13, asterisk *) matched the value of that state in the average of the three consensuses. This level was chosen because it was easily identifiable, relatively low in noise, and was present in every analyzed measurement with both DNA and peptide sections due to its position between the two regions.

With a set of properly calibrated reads and an initial guess for the consensus, we updated the peptide section of each consensus by running the EM algorithm to convergence using a randomly chosen subset of the peptide section of 20 of each variant's reads, and thus arrived at the three consensuses used to classify the reads to produce Figures 2C and 3C in the main text.

5. Event selection

Candidate events were identified with a simple thresholding algorithm and filtered by duration, keeping only those blockages longer than 1 second. The candidate reads were then inspected by eye, and only reads matching the DNA prediction in their first part, and containing further enzyme stepping behavior after the end of the DNA were retained. Reads containing significant amounts of MspA gating or spurious noise, or reads with fewer than 12 observed levels were also rejected. These criteria are illustrated in Figure S4. As visible in figure S4 as well as figure S2, the ion current sequence of reads is highly reproducible, but subject to the usual random error intrinsic to nanopore reads, much of which may be systematically removed. Because some analyses rely on separate analysis of the peptide section, we identified the peptide section as beginning at consensus level 49.

6. Molecular dynamics simulations

All simulations were performed using the classical MD package NAMD *(28)*, periodic boundary conditions, and a 2 fs integration time step. The CHARMM36 force field *(29)* was used to describe proteins, dioctadecatrienoylphosphatidylcholine (DPhPC) phospholipids*(30)*, TIP3P *(31)* water, and ions *(32)* along with the CUFIX corrections applied to improve description of charge-charge interactions *(33)*. RATTLE *(34)* and SETTLE *(35)* algorithms were applied to covalent bonds that involved hydrogen atoms in protein and water molecules,

respectively. The particle mesh Ewald (PME) *(36)* algorithm was adopted to evaluate the long-range electrostatic interaction over a 1 Å-spaced grid. Van der Waals interactions were evaluated using a smooth 10–12 Å cutoff. Langevin dynamics were used to maintain the temperature at 295 K. Multiple time stepping was used to calculate local interactions every time step and full electrostatics every two time steps. The Nose-Hoover Langevin piston pressure control *(37)* was used to maintain the pressure of the system at 1 atm by adjusting the system's dimension. Langevin thermostat *(37)* was applied to all the heavy atoms of the lipids with a damping coefficient of 1 ps^{-1} to maintain the system temperature at 295 K.

An all-atom model of reduced-length MspA was constructed as described previously *(19)* to include residues 75–120 of the full-length protein, merged with an 8 $\times$ 8 nm^2 patch of DPhPC bilayer and solvated with 0.4 M KCl electrolyte, a system of approximately 39,500 atoms. Thirty-two aspartate residues were replaced by asparagine or arginine to create the D90N/D91N/D93N/D118R mutant used in experiment. Eleven additional systems were constructed to have a 23-amino acid polypeptide strand placed inside the nanopore to span through the nanopore constriction, differing by the amino acid sequence and the location of the single amino acid substitution relative to the constrictions, see Figures S5–S7 for details. The peptides were built to have a stretched conformation characterized by the end-to-end distance of approximately 65Å.

Following assembly, each peptide system was minimized in 2,000 steps using the conjugate gradient method and then equilibrated for 45 ns at a constant number of atoms, pressure, and temperature (NPT) ensemble performed while keeping the ratio of the systems size along the place of the bilayer constant. During the equilibration and in all subsequent simulations, a harmonic restraint (k_{SPRING} = 10 kcal mol^{-1} $Å^{-2}$) was applied to the C_α atom of top (C-terminal) residue of the peptide. Additionally, each C_α atom of the MspA protein was harmonically restrained (k_{SPRING} = 1 kcal mol^{-1} $Å^{-2}$) to its X-ray coordinate *(38)*. The systems were then simulated for 50 ns in a constant number of particles, volume and temperature (NVT) ensemble under a constant electric field $E = -V/L_z$ applied along the z-axis (normal to the membrane) to produce a transmembrane bias V; where L_z is the dimension of the simulated system in the direction of the applied electric field *(39,40)*. For the NVT simulations, the systems dimensions were set to the average dimensions observed within the last 5 ns of the restrained NPT equilibration.

To obtain a representative ensemble of peptide conformations within the nanopore, each of the seven peptide systems were simulated under a transmembrane bias of either 200 mV (G and D systems) or 600 mV (W systems) while moving the top residues of the peptide strand by 6 Å away from the constriction and back, four times over the course of 400 ns, Figures S5–S7. Sixty four instantaneous configurations were chosen from the four simulations of the G and W systems producing an ensemble of conformations differing by the location of the amino acid substitution relative to the MspA constriction; twenty six configurations were chosen from the

four simulations of the D systems. Each system was then simulated for 200 ns under 200 mV having the top residue of each peptide stationary restrained to its coordinate in the chosen instantaneous configuration. During these 200 ns simulations, the amino acid substitutions maintained their z coordinate within -15 to $+15$Å from the nanopore constriction, Figure S9. The blockade current analysis was done on these 90 trajectories. The open pore system was minimized (2,000 steps) and equilibrated (45 ns in NPT) similar to the peptide systems. The open pore current was obtained from a 450 ns NVT simulation under a 200 mV bias.

Instantaneous ionic current was calculated as *(39)*

$$I(t) = \frac{1}{\Delta t l_z} \sum_{j=1}^{N} q_j (z_j(t + \Delta t) - z_i(t)), \tag{5}$$

where $z_j(t + \Delta t) - z_i(t)$ is the displacement of ion j along the z direction during the time interval Δt = 20 ps and q_{j} is the charge of ion j. To minimize the effect of thermal noise, the current was calculated within an l_z = 20 Å thickness slab centered at the nanopore constriction (the slab spanned the entire simulation system in the x-y plane). The instantaneous values of the ionic current were recorded simultaneously with the center of mass z coordinate of the backbone of the amino acid substitution. For each amino acid substitution, the data from all trajectories were sorted according to the z coordinate of the substitution in ascending order. The average value of the current and its standard error were computed using 2 Å bins along the z coordinate.

To calculate the fraction of the nanopore volume available to conduct ionic current, we first computed the average number of bulk-like water molecules confined within the nanopore constriction for the open pore simulation. Bulk-like water molecules were defined as those located more than 2.5 Å away from any protein atoms. Previously, we found the number of bulk-like water molecules in the MspA constriction to determine the ionic current through MspA blockade by a DNA strand *(19)*. Following that, we calculated the instantaneous number of bulk-like water molecules in the nanopore constriction for every second frame of the 90 MD trajectories that we used for the ionic current blockade analysis (31 for G, 33 for W and 26 for D systems, Figure S9). The nanopore constriction was defined as the inner volume of the nanopore located within 4Å along the z axis from the center of mass of residues 90 and 91.

The fraction of nanopore constriction volume occupied to conduct ionic current was obtained by dividing the number of bulk-like water molecules in the nanopore constriction blocked by the peptide by the number of bulk-like water molecules in the open pore constriction. For each amino acid substitution, the data from all trajectories were sorted according to the z coordinate of the substitution in ascending order. The average value of the volume fraction and the standard error were computed using 1 Åbins along the z coordinate.

7. Variant identification

When identifying variants, we used those calibrated reads randomly reserved from inclusion in the consensus generation. Using a Viterbi algorithm accommodating both forwards and backwards steps, unobserved steps, spurious ion current levels, and over-segmented levels (described in previous work: Laszlo 2014*(17)*, supplementary note 2), the peptide section consisting of levels 49 (see Figure S13, dagger †) to either termination or a rewinding event, of each read was aligned and assigned likelihood scores, and the highest scoring read was determined to be the variant for that read.

8. Re-read analysis

Re-reads were reliably obtained by increasing helicase concentration to an excess of 1 μM, at which nearly all events had at least 1 re-read. At the lower 100 nM helicase concentrations used to collect the bulk of the experimental data, at least 1 helicase slip-back event was seen in approximately 16% of full peptide reads.

For the re-read analysis in main text Figure 3, a particularly long read of the G22 template containing approximately 117 rewinding events was parsed by hand to separate each separate re-read (Figure S10A). They were then processed using the level segmentation, filtering and backstep removal described in Materials and Methods 2. Only re-reads with at least one ion current level greater than 0.35 I/I_{OS} were included in the analysis shown in main text Figure 3C. This restriction was applied in order to ensure that the read re-wound at least to consensus level 53, covering at least half of the ion current level sequence affected by the substitution, and omit attempts to identify the variant from reads of a section not containing the substitution. 45 re-reads were ultimately included in the final analysis. Each re-read was assigned a likelihood of being drawn from each of the three variants, as described in Materials and Methods 7, using a consensus trained on all 216 single-read events (available in the Supplementary Data *(26)*).

The accuracies in main text Figure 3C were computed as follows: For each integer value of N = 1 to 45, 10^6 randomly selected subsets of N reads were generated. For N = 1, 2, 43, 44, and 45, the number of possible subsets is smaller than 10^6, so the random sampling closely reproduces the results of analyzing all possible subsets of these sizes.

For each subset, the variant likelihoods of each re-read in the subset were multiplied and then normalized to sum to 1. The maximum-likelihood variant was then chosen. The N-re-read accuracy was defined as the proportion of the 10^6 subsets of size N whose maximum-combined-likelihood variant was the true (G) variant for the read. For example, to estimate the accuracy from 10 re-reads, we randomly selected 10 of the re-reads in the event. We calculate the likelihood that each of the 10 reads was of each variant, and multiply together the 10 values for each variant. The variant with the largest product is the identification. This is repeated 10^6 times, and the number of 'G' identifications was divided by 10^6 to obtain the fractional

accuracy. Single-pass accuracy in the re-read data (the N=1 data point in main text Figure 3C) is lower than that reported in main text Figure 2C because re-reads often only partially cover the variant site, resulting in some loss of sequencing information.

This approach was carried out with further reads, including many shorter than the one shown in the main text, and the results are summarized in figure S10B.

In future work, we expect that even better results can be obtained by first generating a consensus sequence of ion currents, and using that signal to classify variants or sequence, rather than simply combining match likelihoods, because this in principle allows for convergence on an easily recognizable consensus sequence much faster, typically with fewer than 10 re-reads.

Supplemental Text

1. Heterogeneously charged peptide reads

Our paper presents the principle of our approach using model peptides that are homogeneously charged. Preliminary data indicate that the method will also work with more heterogeneous template constructs ("hetero template" in Table S1), which consist of a mixture of positive and negative charges as well as polar and non-polar neutral side chains. Indeed, first reads yielded clear and reproducible stepping ion currents (Figure S11). While further experimentation and analysis will be required to fully understand the ion current signals observed, these reproducible traces are evidence for the sufficiency of the electro-osmotic force to trap peptides in MspA for cases where a strong electrophoretic force is absent.

2. Variable voltage reads

In previous work*(18)*, we showed that by driving a nanopore experiment with a time-varying voltage, and thereby measuring a conductance-voltage curve at each enzyme step instead of only a mean ion current, we obtained significantly more detailed information about the DNA being sequenced and single-read sequencing accuracy increased dramatically. The principle of this method is applicable to any polymer analyte whose mean position in the pore depends on the pulling force applied by the voltage, including electrophoretically or electro-osmotically trapped peptides.

To assess the feasibility of future experiments using high-fidelity variable voltage sequencing, we obtained and processed variable voltage reads exactly following the methods of Noakes 2019. These reads display similar qualitative characteristics to variable voltage experiments with DNA: namely, the ability of each conductance-voltage curve to be fit well with a second-order polynomial, and rough continuity of the curves enabling backsteps and unobserved enzyme steps to be more readily identified (Figure S12). This suggests that in future work, using the variable-voltage sequencing method will allow for comparable improvements to peptide

sequencing fidelity. In the current paper, we focused instead on the constant-voltage current stepping signals for clarity of communication.

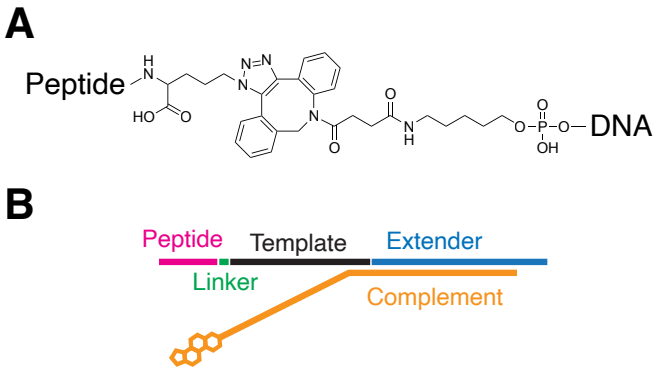


Fig. S1
Details of sequencing constructs. (A) The chemical structure of the DBCO-Azide linker used in the DNA-peptide hybrids.(B) First-generation sequencing construct that was used in early data acquisition, which ultimately was replaced by the one shown in main text Figure 1A.

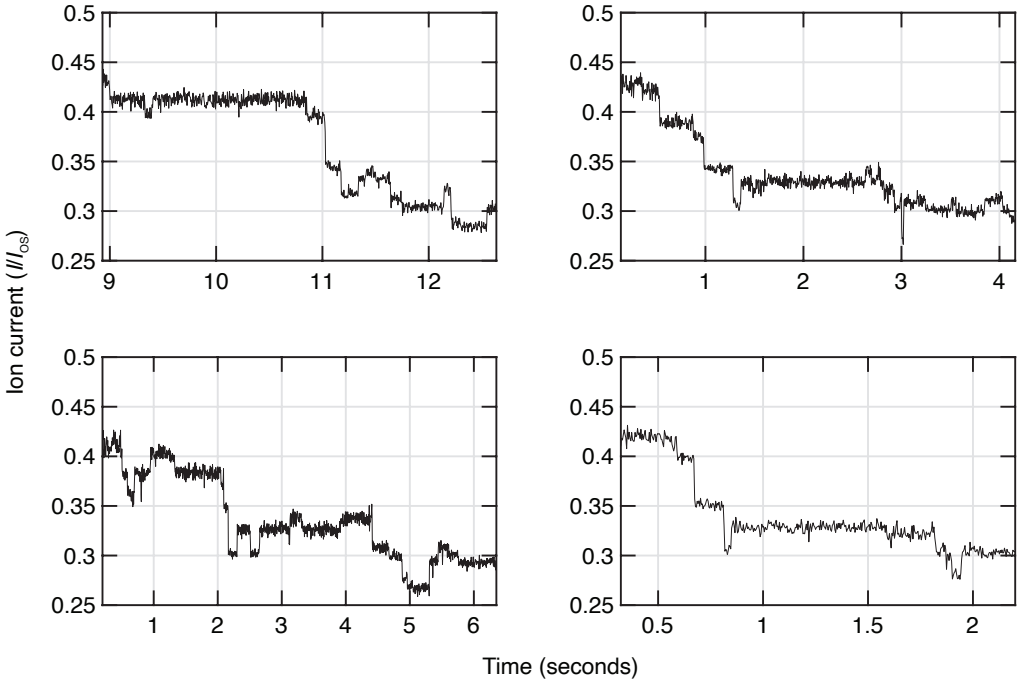


Fig. S2

Reproducibility of ion current sequences. A random selection of the peptide section of W-variant reads used in the paper, demonstrating the consistency of the observed ion current levels. Different reads contain apparent variation: they differ considerably in terms of the durations of individual helicase steps, noise sometimes makes it difficult to segment steps, very short states are sometimes apparently missing from reads, and numerous helicase backsteps are apparent. However, the underlying sequence of ion current means is highly reproducible, and the random errors may be accounted for in the process of backstep removal (Materials and Methods 4) and read alignment to consensus (Materials and Methods 7), or eliminated through re-read analysis (main text Figure 3).

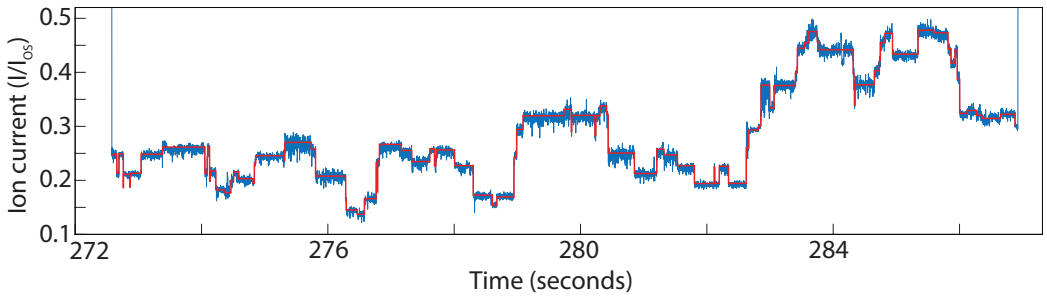


Fig. S3
Typical performance of information-based level finder on nanopore data.

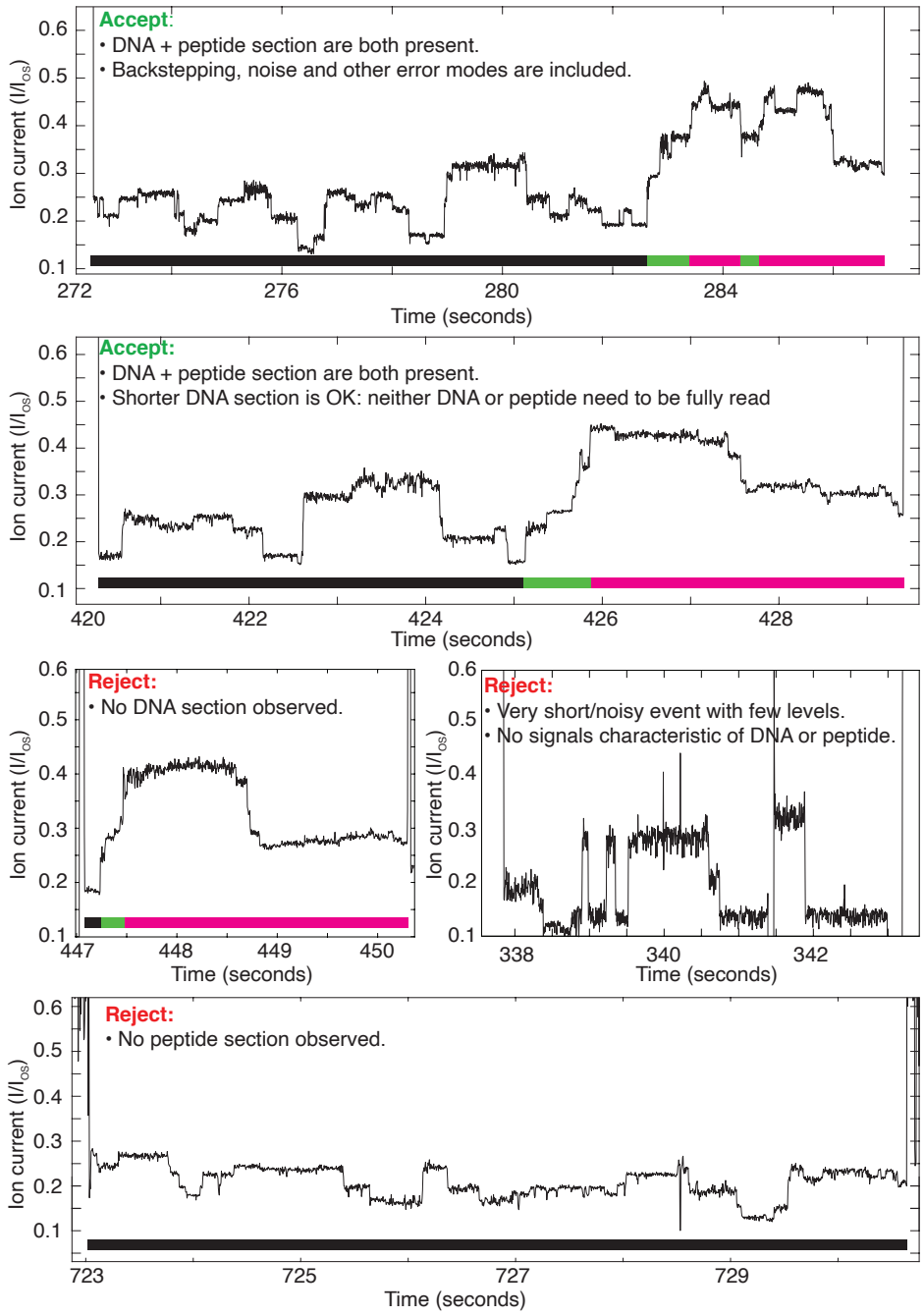


Fig. S4

Examples of accepted and rejected events. Colored bar at the bottom indicates which portion of the conjugate polymer is in the constriction of MspA: black = DNA, green = linker, magenta = peptide.

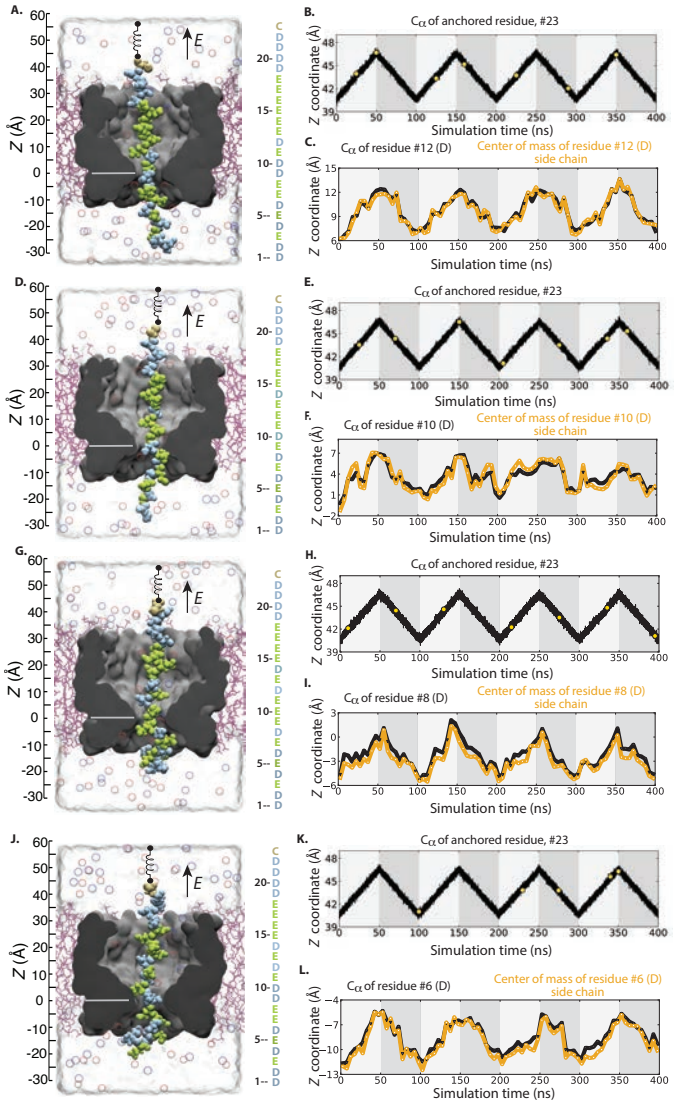


Fig. S5

Preparation of initial configurations for the production simulations of the D system. Starting from the pre-equilibrated configuration shown in panel A, the system was simulated using the all-atom MD method under a 200 mV bias while the C_α atom of the top residue (#23) was moved up and down by approximately 6 Å four times in 400 ns, panel B. Panel C shows the corresponding displacements of the D residues at position #12 separately for the backbone C_α atom and for the side chain center of mass. The yellow circles in the trace in panel B show the initial configurations chosen for the production simulations, Figure S9B. (D–F, G–I and J–L) Same, but for different placement of the D residue (at residue #10, #8 and #6) relative to the pore constriction.

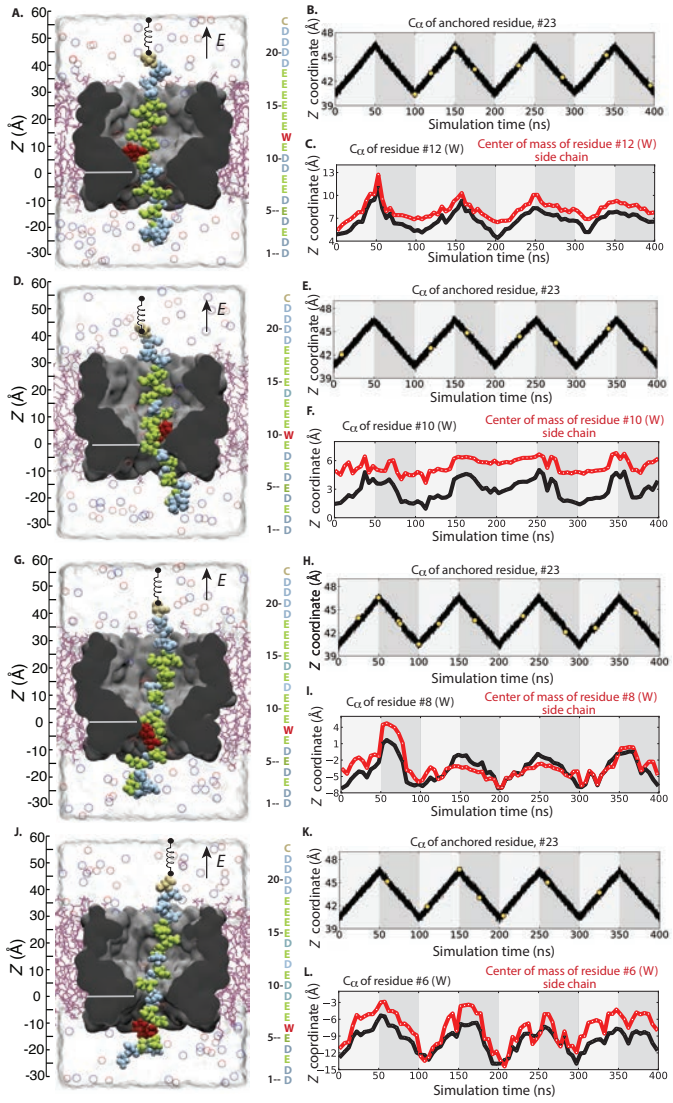


Fig. S6

Preparation of initial configurations for the production simulations of the W system. (A–C) Starting from the pre-equilibrated configuration shown in panel A, the system was simulated using the all-atom MD method under a 600 mV bias while the C_α atom of the top residue (#23) was moved up and down by approximately 6 Å four times in 400 ns, panel B. Panel C shows the corresponding displacements of the W residues at position #12 separately for the backbone C_α atom and for the side chain center of mass. The yellow circles in the trace in panel B show the initial configurations chosen for the production simulations, Figure S9B. (D–F, G–I and J–L) Same, but for different placement of the W mutation (at residue #10, #8 and #6) relative to the pore constriction.

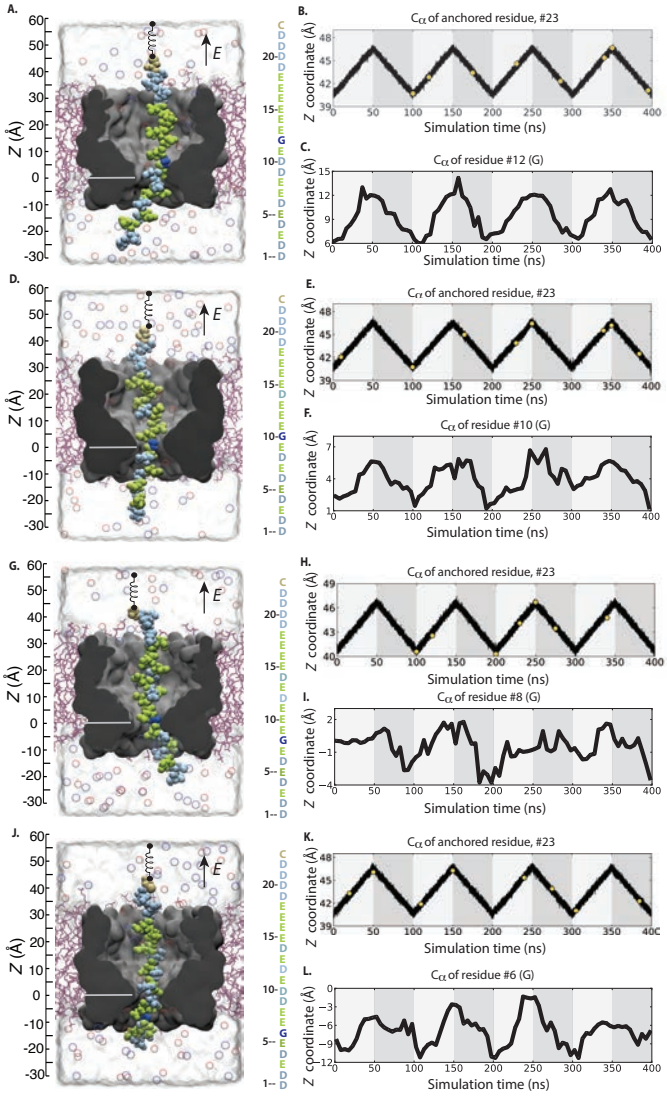


Fig. S7

Preparation of initial configurations for the production simulations of the G system. (A–C) Starting from the pre-equilibrated configuration shown in panel A, the system was simulated using the all-atom MD method under a 200 mV bias while the C_α atom of the top residue (#23) was moved up and down by approximately 6 Å four times in 400 ns, panel B. Panel C shows the corresponding displacements of the G residue at position #12 by plotting to coordinate of the backbone C_α atom. The yellow circles in the trace in panel B show the initial configurations chosen for the production simulations, Figure S9B. (D–F, G–I and J–L) Same, but for different placement of the G mutation (at residue #10, #8 and #6) relative to the pore constriction.

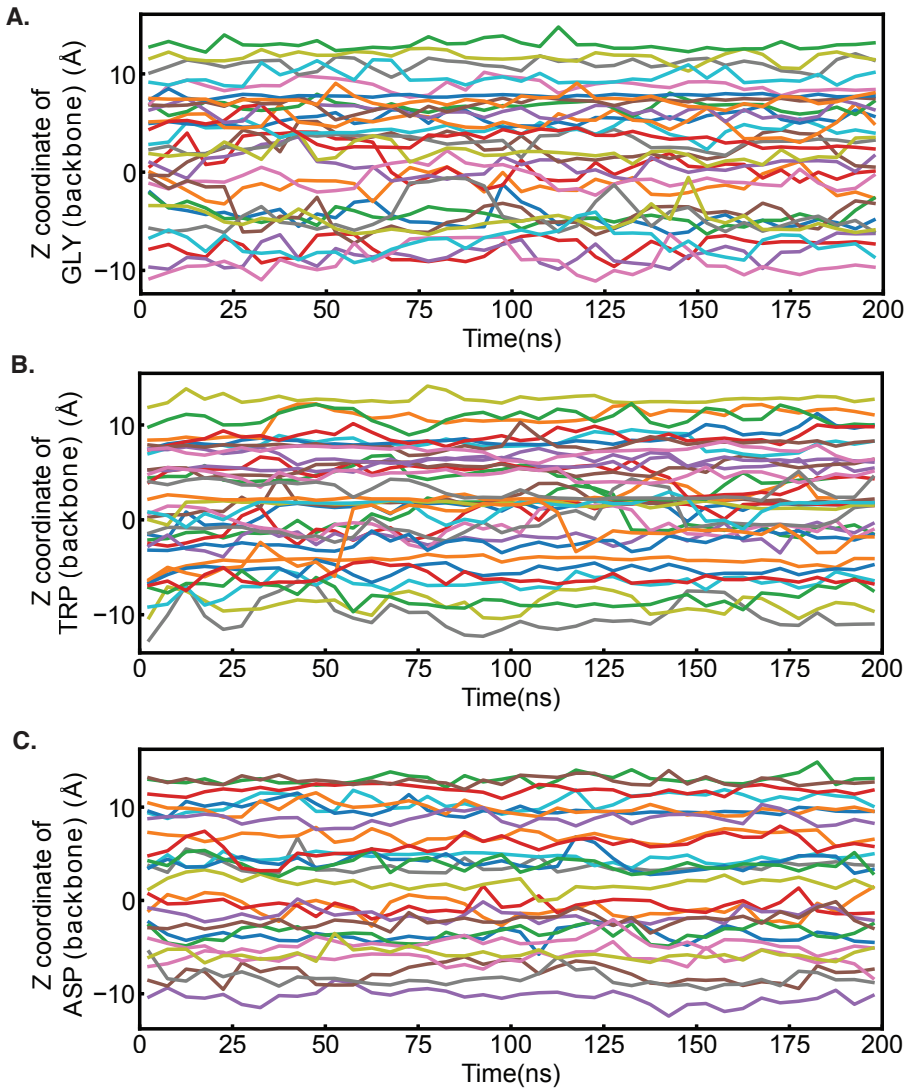


Fig. S8

Production simulations of MspA–peptide systems. (A–C) Center-of-mass z coordinate of a single amino acid backbone versus simulation time for 90 independent MD simulations carried out under a 200 mV bias. Data in panels A, B and C correspond to 31 G, 33 W and 26 D simulations. The initial states for the MD simulations were chosen from the periodic displacement simulations featured in Figures S7–S5.

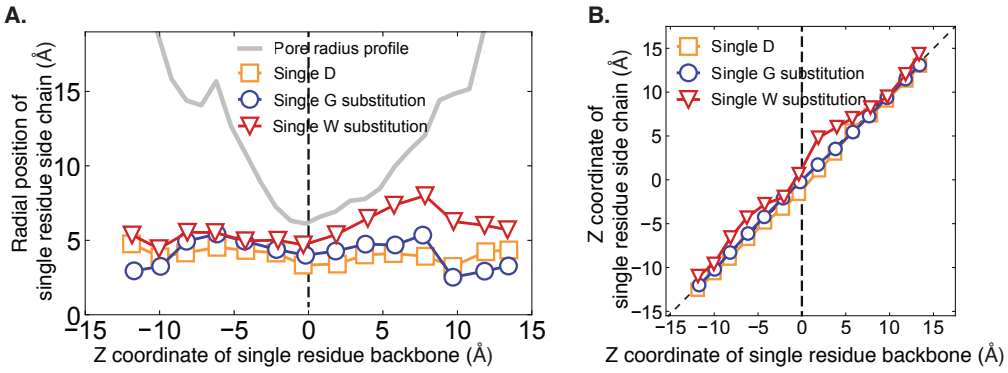


Fig. S9

Side chain conformation in MspA constriction. (A) Radial center-of-mass coordinate of a single residue side chain versus the center-of-mass coordinate of the residue's backbone. The radial coordinate was computed relative to the symmetry axis of the MspA nanopore. The vertical dashed line illustrates the location of the MspA constriction whereas the grey line shows the local radius of the MspA nanopore. Data for D and G/W substitutions were obtained by averaging over all production simulations, Figures S9A and B, respectively. As coordinates of the glycine side chain, we used the coordinates of its C_α atom. (B) Center-of-mass z coordinate of the single residues side chain versus center-of-mass z coordinate of that residue's backbone. Dashed diagonal line corresponds to a situation where the side chain and the backbone have the same z coordinate within the nanopore.

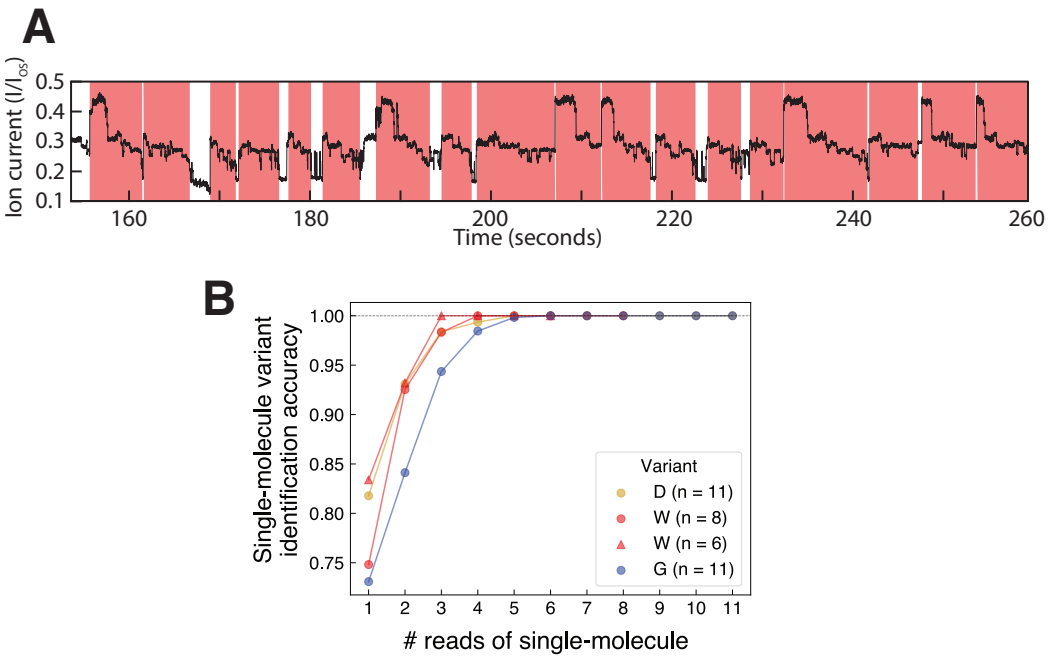


Fig. S10

Re-read analysis. (A) Re-read segmentation. Each identified independent re-read is bounded in a red highlighted region. Re-read ends were marked at approximately consensus level 60, or at the end of the re-read if rewinding occurred before level 60. Re-read beginnings were marked when the current returned to a previously visited level more than 3 steps back, in order to avoid representing normal helicase backsteps as very short re-reads. (B) Plots similar to main text Figure 3C for several events with various numbers n of total re-reads. Color indicates true variant: gold = D-variant, blue = G-variant, red = W-variant. Two multi-read events, one G-variant with n = 4 rereads, and one D-variant with n = 9, yielded 100% accuracy for all re-reads in the event and are omitted from the plot for clarity.

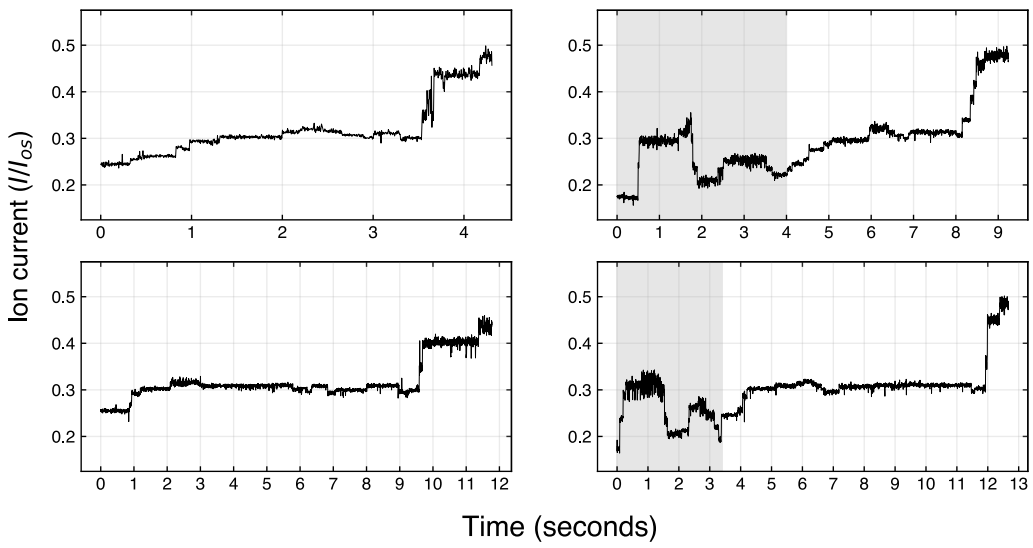


Fig. S11
A selection of reads of heterogeneously charged peptides ("hetero template" in Table S1). Clear stepping can be seen, with a similar number of steps as observed in the negatively charged peptide experiments. The gray shaded regions in the right column reads correspond to the DNA portion of the read; compare to main text Figure 1E levels 1 through 44.

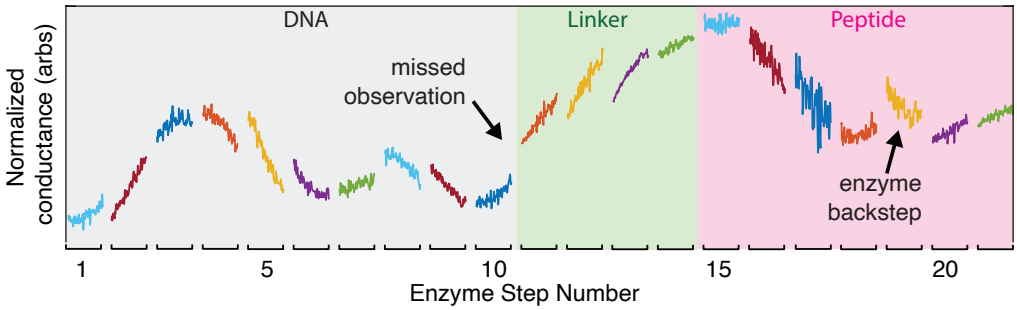


Fig. S12
Sample variable-voltage read. This variable-voltage read covers both the DNA and D-variant peptide parts of the sequence. Compare the trends in conductance to the trends in ion current seen in main text Figures 1D, 1E, and 2A. Major qualitative features yielding improved sequencing fidelity in DNA reads are retained in reads of the peptide, including the smooth character of the conductance-voltage curve at each enzyme step, and the rough continuity of a forward-stepping read allowing us to infer where backsteps or missed observations occur in a read. Colors are only to aid the eye in discerning separate ion conductance curves.

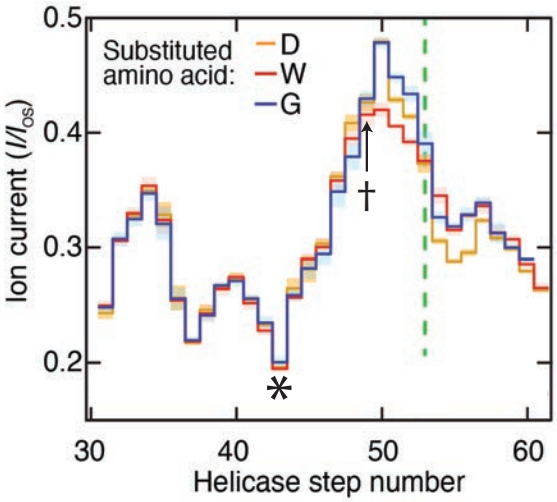


Fig. S13

Important consensus levels. Level 43, marked with an asterisk *, was used to calibrate all reads with an overall scaling of ion current. This level was identified in raw data as the locally minimal ion current level preceding the maximum ion current level in the trace. Level 49, marked with a dagger †, marked the beginning of the peptide section used for analysis. This level was identified in raw data as the level immediately preceding the maximum ion current level in the trace.

Oligo name	Sequence
D22 template	[N-term] DEDEDEDEDEEDDEDEDDEEEDDDD [C-term] (linker) 5' TTACTGAAGTCTCACGTGCCTGGTATATTA 3'
W22 template	[N-term] DEDEDEDEDEEDDEDEDDEEEWDDD [C-term] (linker) 5' TTACTGAAGTCTCACGTGCCTGGTATATTA 3'
G22 template	[N-term] DEDEDEDEDEEDDEDEDDEEEGDDD [C-term] (linker) 5' TTACTGAAGTCTCACGTGCCTGGTATATTA 3'
hetero template	[N-term] DDDDDDDDDDDDDDYAVEGRDLTLS [C-term] (linker) 5' TTACTGAAGTCTCACGTGCCTGGTATATTA 3'
complement gen1	5' TGATCAATTCACTGTGGATGTAATATACTTTTTTTTTTTTTTTTTTTTTTTTTTTTT 3' (cholesterol)
complement gen2	5' GATGTAGAATTTTTTTTTTTTTTTTTTTT 3' (cholesterol)
template extender gen1	(phosphate) 5' CATCCACAGTGAATTGATCAGGTCGTAGCC 3'
template extender gen2	(phosphate) 5' CATCCACAGTGAATTGATCATTATGACGTTATTCTACATCGGTCGTAGCC 3'
staple	5' GGATGTAATAGC 3'

Table S1
Sequences of DNA oligos and DNA-peptide hybrid oligos.