Developing Multiplexed Molecular Assays for Synthetic Biology and DNA Data Storage with Nanopore Sensing Technology

Nicolas Cardozo

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Jeffrey Nivala, Chair

Luis Ceze

Georg Seelig

Eric Klavins

Program Authorized to Offer Degree:

Molecular Engineering

University of Washington

Abstract

Developing Multiplexed Molecular Assays for Synthetic Biology and DNA Data Storage with Nanopore Sensing Technology

Nicolas Cardozo

Chair of the Supervisory Committee:

Jeffrey Nivala

Department of Computer Science and Engineering

Multiplexed molecular tools are a powerful means of interrogating biomolecular systems. Multiplexed assays offer significant advantages over singleplex assays, including time, reagent costs, sample requirements, and the amount of data that can be generated. In molecular biology, genetically encoded reporter proteins have expanded the toolbox of researchers to track biological phenomena. While they are widely used to measure many biological activities, the current number of uniquely addressable reporters that can be used together for one-pot multiplexed tracking is small due to overlapping detection channels such as fluorescence. Similarly, in DNA data storage, the ability to selectively target data files (i.e “randomly access”) in a multiplexable manner would lessen decoding latency and cost and enable deployment of practical DNA data storage architectures. The primary focus of this dissertation is to develop new multiplexable molecular assays for synthetic biology circuits and DNA data storage random

access readout using nanopore sensor arrays. To overcome genetically encoded reporter protein multiplexing limitations, we built an expanded library of orthogonally-barcoded Nanopore-addressable protein Tags Engineered as Reporters (NanoporeTERs), which can be read and demuxed by nanopore sensors at the single-molecule level. Subsequently, to improve upon previous random access architectures, we demonstrate a new random access approach in which files can be selected in multiplex using a CRISPR-Cas9 target address and then decoded using a nanopore sequencer. This work presents a new class of reporter proteins that permit multiplexed, real-time tracking of gene expression along with a new random access DNA data storage strategy that increases one-pot multiplexing and decreased time-to-decoding.

TABLE OF CONTENTS

List of Figures.....(vi)

Acknowledgments.....(vii)

Chapter 1: Introduction.....(1)
1.1 Nanopore sensors.....(1)
1.1.1 DNA Sequencing.....(2)
1.1.2 Protein analysis.....(3)
1.1.3 Fluorescent protein reporters.....(4)
1.1.4 DNA data storage.....(5)
1.2 Overview of thesis……(5)

Chapter 2: Multiplexed direct detection of barcoded protein reporters on a nanopore array .....(7)
2.1 Introduction…….....(7)
2.2 Results…….....(8)
2.3 Discussion……...(16)
2.4 Methods……...(18)

Chapter 3: Cas9-mediated random access in DNA data storage…….....(49)
3.1 Introduction…….....(49)
3.2 Results…….....(51)
3.3 Discussion……...(53)
3.4 Methods……..…(53)

Chapter 4: Towards single-molecule protein sequencing on a high-throughput nanopore sensor array.....(61)
4.1 Introduction…….....(61)
4.2 Discussion……...(64)
4.3 Conclusion……..(67)

References**……(68)**

LIST OF FIGURES AND TABLES

Figure Number Page

Figure 1.1 Nanopore sensor illustrating molecule sequencing and trace generated……2

Figure 2.1: Nanopore-addressable protein Tags Engineered as Reporters (NanoporeTERs)…...24

Figure 2.2: NanoporeTER sequences……..25

Figure 2.3 Raw nanopore ionic traces……27

Figure 2.4 Applied voltage on NTER capture……...36

Figure 2.5 NTER tail length on nanopore capture……37

Figure 2.6: Mapping the NanoporeTER sequence and nanopore signal space on a MinION…...38

Figure 2.7. Violin plots……..40

Figure 2.8: Violin plots part two……41

Figure 2.9: Classification and multiplexed detection of NanoporTER expression levels with a MinION……..42

Figure 2.10: Immediate NTER re-capture analysis……...44

Figure 2.11 MinION flow cell lifetime……..45

Figure 2.12 95% confidence intervals……...46

Figure 2.13 Relationship between NTER concentration and MinION run time……...47

Figure 2.14 NTER barcode signal space……48

Figure 3.1: Schematic of random access pipeline……..57

Figure 3.2: Single-file enrichment of file 10…….58

Figure 3.3: Cas9 gRNA sequences for file 10 AND 13 in library…….59

Figure 3.4: The mean number of payload copies (concatemers) per read…….60

Figure 4.1 Challenges in protein translocation through a nanopore……..63

ACKNOWLEDGMENTS

I will begin by thanking my advisor, Prof. Jeff Nivala. From day one, Jeff gave me the freedom to pursue the projects that interested me most. He actively brainstormed new directions to take experiments and projects, continuously gave me support, and provided me with the tools that I needed to succeed. He has been an incredible mentor and friend and taught me skills and lessons about being a scientist that I never could have learned in a classroom or at the lab bench.

I would like to thank my graduate committee members, prof. Georg Seelig, prof. Luis Ceze, and prof. Eric Klavins. They were there for each milestone in my graduate career and offered ideas and asked the critical questions required to move the projects forward. Their support and encouragement consistently inspired me to develop as a scientist.

I am grateful to the entire Molecular Information Systems Lab group and the Nanopore group. I would especially like to thank prof. Luis Ceze and prof. Karin Strauss for providing such a welcoming and supportive environment that allows for collaboration and creative thinking. Keisuke Motone for being a dear friend and colleague and bringing such a positive attitude to the lab. And all my other lab members and the many other members of MISL.

Finally, I would not be where I am today without the unconditional support and love of my mom, dad, brother, and grandfather. I am so fortunate to have such a supportive family and I hope to continue to make them proud. This could not have been possible without you!

Thank you!

Chapter 1

INTRODUCTION

Molecular assays are employed in numerous disciplines, from clinical medicine to agriculture and biotechnology. These assays are used to study the interaction of molecules in biological and synthetic systems. Output readout can be measured through several means, including visual outputs such in immunochromatographic assays[1-3], RT-qPCR assays[4-6], and more recently through molecular sensors such as nanopore sensors[7-9]. These nanopore sensors have traditionally been used for nucleic acid sequencing and measure ionic current disruptions across nanometer-sized pores as molecules flow through or block the pore (Figure 1.1). As individual molecules translocate through the pore, they cause an observable signal change that provides structural insights into the molecules[10].

The goal of this dissertation is to explore novel multiplexed molecular assays that utilize commercially available nanopore sensors for output readout. First, I will introduce a new class of

multiplexed barcoded protein reporters. Later, I will present a new multiplexed random access architecture for DNA data storage.

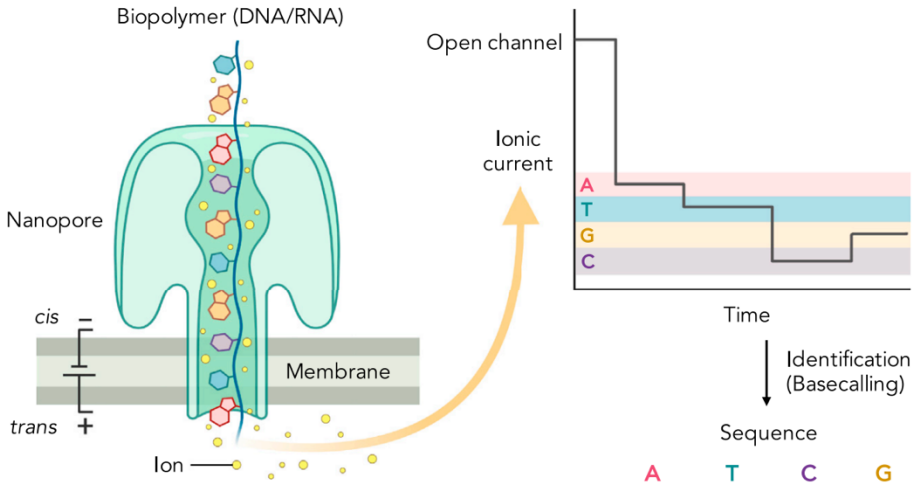


Figure 1.1 Nanopore sensor illustrating molecule sequencing and trace generated. In nanopore sequencing, a nanometer-sized protein pore is embedded in an insulating membrane that separates two electrolyte-filled wells. Voltage is applied between the wells, causing ionic current flow through the pore. As single biopolymer molecules translocate through the channel, they generate sequence-specific ionic current signals that are diagnostic of the polymer sequence. These sensors are traditionally used for nucleic acid sequencing. (*Figure obtained from a previously published manuscript in iScience [79].*

1.1 Nanopore sensors

1.1.1 DNA sequencing

Nanopore sensors have had a long and rich history as powerful single-molecule sensors[11,12]. More recently, through the commercialization of nanopore technology by Oxford Nanopore Technologies (ONT), has real-time sequencing of DNA[11] and RNA[13] been explored. ONT's MinION device is a portable, high-throughput nanopore sensor array that connects to a laptop. It holds several key advantages over 2nd-generation sequencing platforms (sequencing-by-synthesis-Illumina sequencing), including direct DNA and RNA sequencing, long-read lengths,

portability, and single-molecule resolution (a single nucleic acid disrupts the ionic current at a time).

Nanopore DNA sequencing is possible due to DNA being inherently negatively charged. To control translocation speed, sequencing adapters are ligated to the ends of the DNA molecule of interest. The adapters, once bound, unwind the DNA and facilitate strand capture and translocation of the molecule across the pore. As each nucleotide crosses the pore, distinct ionic current traces are measured. The MinION contains 512 individually addressable pores that measure ionic current in each sequencing run. The ionic current data of reads collected throughout the sequencing run are stored in FASTA5 files. These reads are then deconvoluted into the individual nucleotides through a process termed basecalling.

1.1.2 Protein Analysis

Nanopore-based analyzers (i.e MinION) focused on protein structure and function have recently gained momentum in research institutions world-wide. These sensors provide low-cost and high-throughput platforms with adaptability towards native protein/peptide studies[14-16]. Nanopore sensors have been used in the field of proteomics to study several phenomena. These include protein unfolding kinetics, protein structure stability, and free-energy profiles of nucleic acid–protein binding[17]. Proteins have been unfolded during forced translocation through the nanopores, revealing distinct steps in the unfolding process[18,19]. In addition, the technology has also been used to demonstrate single-molecule, site-specific detection of protein phosphorylation[20]. Thus, nanopore technology has unique potential in expanding the field of

single-molecule proteomics. Specifically, ONT's high-throughput nanopore sensor, the MinION, could serve as a general platform to enable high-throughput nanopore-based proteomic analysis.

1.1.3 Fluorescent protein reporters

The monitoring of transcriptional events via linkage with reporter gene expression has been used extensively to investigate various biological processes[21]. Reporter genes and proteins, such as β-galactosidase, bacterial luciferase, and green fluorescent protein (GFP), have been utilized as a simple and rapid means of detecting and quantifying molecular and genetic events. The most widely used fluorescent protein is GFP, which was originally isolated from the jelly-fish Aequoria victoria[22,23]. These reporter systems function in a simple manner: a gene is attached to a regulatory sequence, which when introduced into a biological system, provides an easily measurable signal output upon modulation of its expression[21]. A key limitation of these reporter schemes is that the typical number of uniquely addressable reporters that can be used together while sharing a common readout is small[24-26]. This is primarily due to the optical nature of traditional reporters, such as fluorescent protein variants, which have overlapping spectral properties that make simultaneous measurement of unique genetic elements difficult[24]. Therefore, the ability to increase the multiplexability of genetically-encoded protein reporters would enable more comprehensive and scalable monitoring of biological systems. This would have a significant importance in the field of synthetic biology for more complex circuit design and readout.

1.1.4 DNA data storage

Synthetic DNA has been researched as a storage medium for long-term data storage[27-34], making it a possible alternative to electronic data storage. DNA is a robust information storage molecule, featuring ultrahigh information density, and long-term chemical stability, capable of retaining its integrity for thousands to millions of years. As DNA synthesis and sequencing costs continue to decrease, DNA will likely only become an increasingly attractive medium for digital data storage[35-38]. Recent efforts have established end-to-end workflows of DNA data storage, which encompass encoding, synthesis, retrieval, sequencing, and decoding of data stored in DNA[28,32,34]. However, to make practical DNA data storage architectures, scaling up DNA data storage will require methods to selectively retrieve and read pieces of data (random access). Thus, it is important to develop multiplexed assays that allow the targeting of more files in a single-pot reaction (greater multiplexability) while optimizing energy and time efficiencies.

1.2 Overview of thesis

The work presented in this dissertation demonstrates that a commercially available nanopore sensor (ONT's MinION) can be utilized for applications outside of its traditional use case to address some of the challenges discussed above. First, I present a new class of barcoded protein reporters (Chapter 2). Later, I show that the MinION sensor can be coupled with the CRISPR-Cas9 bacterial immune system to target and randomly access files of interest in a large pool of DNA, ultimately improving upon previous architectures of random access (Chapter 3). Finally, I discuss the potential of nanopore technology to enable de novo protein identification through

single-molecule sequencing as well as the barriers that will have to be overcome to achieve this goal (Chapter 4). Together, these efforts aim to increase the power we have in processing information in molecular form for new applications in synthetic biology and data storage. The works also demonstrate the ability to adapt a high-throughput nanopore sensor array, which is typically used for nucleic acid sequencing (ONT's MinION), for novel multiplexed assays and its potential implementation for single-molecule protein/peptide detection and sequencing methods.

Chapter 2

MULTIPLEXED DIRECT DETECTION OF BARCODED PROTEIN REPORTERS ON A NANOPORE ARRAY

Portions of this chapter were previously published as a manuscript in Nature Biotechnology [39].

2.1 Introduction

Reporter proteins have been used to track biological activities such as genetic regulation[40]. Although several different reporter strategies have been developed over this period, the typical number of uniquely addressable reporters that can be used together while sharing a common readout is small[24–26, 41]. An ability to increase the multiplexability of genetically-encoded protein

reporters would enable more comprehensive and scalable monitoring of biological systems, enabling, for instance, high-dimensional phenotyping[42]. This is particularly important for synthetic biology, in which scalable reporter systems are needed to keep pace with the complexity of engineered biological systems in applications such as whole-cell biosensing[43] and genetic circuit design[44].

While biomolecular sensing with nanopore sensors has been explored[45], only recently have high-throughput nanopore sensor platforms emerged for real-time sequencing of DNA[46] and RNA[47]. The commercial emergence and popularization of these technologies creates an opportunity to build an accessible general nanopore-based platform for direct sensing of engineered reporter proteins. In this context, we present here a new class of genetically-encoded protein reporters, which we term Nanopore-addressable protein Tags Engineered as Reporters (NanoporeTERs, or NTERs), that use commercially-available nanopore sensors (Oxford Nanopore Technologies' MinION device) [46] for multiplexed direct protein reporter detection without the need for any other specialized equipment nor laborious sample preparation prior to analysis (Figures 2.1).

2.2 Results

NanoporeTER design

We started by engineering a protein that is easily detectable by a nanopore sensor and can be expressed and secreted by *E. coli*. We based our initial NTER design on the synthetic protein construct 'S1', which we previously developed for unfoldase-mediated nanopore analysis[48,49]. S1 contains a small, folded domain (Smt3) along with a flexible, negatively-charged 67 amino acid C-terminal 'tail' composed of glycine, serine, and acidic amino acid residues, in addition to an

11 amino acid ssrA tag[50]. The tail's lack of structure and net negative charge promotes capture of the protein in a nanopore sensor under an applied voltage. The ssrA tag allows for ClpX-mediated unfolding and translocation of the Smt3 domain, which otherwise inhibits translocation of S1 through the nanopore. For use as a reporter protein in *E. coli*, we modified protein S1 in two ways (Figure 2.1 and Figure 2.2). First, we replaced the ssrA tag with additional glycine/serine/acidic residues to preserve its nanopore threading activity but prevent targeting of the protein for degradation by ClpXP in-vivo. Second, we added an N-terminal OsmY domain[51]. In *E. coli*, OsmY-tagged proteins are secreted into the extracellular medium[52]. We reasoned secretion would facilitate NTER nanopore analysis by avoiding the need to lyse cells, thereby simultaneously reducing both experimental labor and signal noise that could be generated by non-specific interaction of intracellular molecular species (e.g. DNA, RNA, and other proteins) with the nanopores during analysis. Experiments in BL21 (DE3) *E. coli* showed that expression of this modified version of S1, which we term here 'NTERY00', resulted in secretion of the protein into the medium, as detected by SDS-PAGE analysis.

NTER barcode experiment

We next purified the secreted NTERY00 by immobilized metal affinity chromatography (IMAC) and then determined if the NTER could be detected on a MinION. To do this, we used an unmodified R9.4.1 flow cell (which uses a variant of the CsgG pore protein[53]) and a custom MinION run script (Methods). The script applies a constant voltage of -180 mV to all the active pores on the flow cell and statically flips the voltage in the reverse direction in 15 second cycles (ie. 10 seconds 'ON' at -180 mV and 5 seconds 'OFF' or in 'Reverse', Figure 1e). The typical R9.4.1 open pore current level at -180 mV and 500 mM KCl is ~220 pA. As expected, when

NTERY00 was introduced into the flow cell at a concentration of 0.5 uM in these conditions, the current level during each -180 mV portion of the voltage cycle typically underwent a stepwise drop from the open pore value to a consistent lower ionic current state (Figure 2.1 and Figure 2.3), signaling a putative capture of an NTER within the pore. This current drop was reversible (back to open pore) following reversal of the voltage. We further found that the average time spent in the open pore state before transitioning to the lower ionic current state was dependent on both NTER concentration and the applied voltage (Figure 2.1 and Figure 2.4). We also explored if shorter tail lengths could similarly promote capture in the nanopore and found that a tail length truncated by 20 amino acids captured at a rate similar to the full-length design, while reduction by 40 amino acids substantially reduced capture rates (Figure 2.5). Overall, these observations are consistent with a model in which the negatively-charged NTER polyGSD tail is electrophoretically captured in the pore under the applied voltage, and can be ejected from the pore by reversal of the electric field.

Mapping MinION pore with engineered protein NTERs

We proposed that the ionic current characteristics of the NTERY00 capture state should be dependent upon the amino acid sequence of the NTER residues residing within the pore's sensitive limiting constriction. To test this, we made a series of NTER mutants (NTERY01-15) in which a sliding three residue region of the polyGSD sequence was mutated to tyrosines (Figure 2.6). Tyrosines were chosen because their larger side chain structure was predicted to decrease the ionic current flow through the pore relative to the glycines and serines of NTERY00 when captured within the pore. Following purification and MinION analysis of NTERs 01-15, we found the capture state to be NTER mutant-dependent up to NTERY08, after which we

observed NTER mutants 09-15 to have signal characteristics indistinguishable from NTERY00 (Figures 2.6). These results support a model in which the first ~17 amino acids of the polyGSD tail reside with the CsgG nanopore's sensitive region and contribute to its ionic current signature during a capture event.

NTER sequence-to-signal characteristics

After determining the number of amino acids that contribute to the NTER nanopore signal (the NTER sequence space), we next sought to determine how different amino acid types modulate the ionic current through the pore. These results help define the possible future NTER signal space. To investigate this, we constructed NTER variants in which positions 3-11 within the polyGSD region were mutated to all the 20 possible standard amino acid homopolymers. Figure 2.6 and Figure 2.8 show the signal features of the ionic current levels for 12 out of the 20 NTER homopolymer mutants (the homopolymers C, F, I, K, L, V, W, and Y, most of which have significant hydrophobic character, did not express sufficient soluble protein). To see how the different amino acid physical properties contribute to the NTER ionic current, we investigated whether certain properties correlate with different signal features. While no strong correlations were found across all the 12 amino acid types, we did find that the median current level moderately correlated with both amino acid volume and helical propensity within the uncharged amino acid types (R = ~0.75 for each, Figure 2.6.

Next, to probe the potential of this method to resolve between amino acid barcodes with subtler sequence differences (for example, point mutations or post-translational modifications), we cloned and tested two additional NTER barcodes based on the protein kinase A (PKA) phosphorylation motif[54]. The first PKA-based barcode contained a canonical PKA motif

(RRGSY), while the second had a single amino acid difference (RRGEY) that mimics the PKA motif's phosphorylated serine state in structure and charge (commonly referred to as a 'phosphomimetic', Figure 2.6). Following purification and MinION analysis of these two NTERs, we found that the phosphomimetic barcode could be distinguished from the canonical PKA motif barcode, as the two barcodes typically had substantially different nanopore ionic current state medians (Figure 2.6). These results suggest the potential of using NanoporeTERs to report on the activity of enzymes that regulate specific post-translational modifications, such as phosphorylation and methylation.

NTER multiplexing experiments

Having explored the potential NTER barcode sequence space, signal space, and sensitivity to single residue modifications, we sought to demonstrate proof-of-principle NTER applications for multiplexed tracking of gene expression. To do this, we first used supervised machine learning to train classifiers that could accurately discriminate amongst combinations of the NTER barcodes explored above. Using either a set of engineered signal features as input to a Random Forest (RF) classifier or the raw ionic current signal directly into a Convolutional Neural Network (CNN) (Figure 2.9), we used our purified NTER datasets for model training and validation. Both models achieved similar accuracies that ranged from ~80-90% depending on the model hyperparameters and barcode set (Figure 2.9, Methods). Next, we used the CNN classifier to assess if NTERs were being immediately re-captured in the nanopore at a non-negligible frequency following their initial capture and ejection, which could lead to molecules being double counted, affecting relative NTER quantification. To investigate this, we analyzed the frequency with which successive captures of the same or different barcode occurred in a mixed

barcode experiment with 5 different NTER barcodes mixed at varying concentrations (Y00: 0.05uM, Y02: 0.1uM, Y05: 0.05uM, Y07: 0.2uM, and Y08: 0.1uM). We found that successive NTER captures of the same barcode were not disproportionately represented, suggesting that immediate re-capture of the same NTER molecule was not occurring at a high frequency (Figure 2.10).

We then used the best performing CNN that was trained on NTERs Y00-08 (in addition to a background/noise class, Methods) to determine the relative NTER expression levels within bacterial cultures composed of mixed populations of strains engineered with different NTER-tagged plasmid-based circuits. To do this, we grew independent mono-barcoded cultures overnight with NTER expression either induced or inhibited (by autoinduction media containing lactose or LB supplemented with glucose, respectively). In the morning, just prior to nanopore readout, the cultures were mixed into a single solution and diluted into MinION running buffer and loaded directly into a flow cell for analysis. Importantly, these cell cultures underwent no processing or purification prior to analysis, in contrast to our previous experiments. Results from these experiments showed higher classification counts for the NTER barcodes for which expression was induced (NTERs 02 and 06), and lower counts for strains that were inhibited (glucose: NTERs Y00, Y04 and Y08) or not present at all in the mixed population (NTERs Y01, Y03, Y05, and Y07) for all replicates (Figure 2.9) over a ten-minute MinION runtime. We then conducted a time course experiment in which we tracked expression of two different NTERs over multiple hours, one of which was induced with IPTG (NTERY06), and the other of which NTER expression was inhibited with glucose (NTERY02). Again, cultures were grown independently, but then mixed just prior to nanopore readout. Figure 2.9 shows the results of this time course (and replicates) during ten minutes of MinION analysis at 2, 4, 6, and 21-hour

timepoints following induction (NTERY06) or inhibition (NTERY02) of the NanoporeTER circuit. Again, the rate of NTER classification (reads/pore-minute) was substantially higher for the induced NTERY06 circuits, compared to the uninduced NTERY02 circuits. Importantly, leaky expression of NTERY02 was still detectable over the background false-positive classification rates for the NTER barcodes that were not present at all in the experiment (Y00, Y01, Y03, Y04, Y05, Y07 and Y08). These results demonstrate that NanoporeTERs can be used as reliable reporters of relative protein expression levels in bacterial cell culture.

Mammal NanoporeTERs (mNTERs)

To show that NanoporeTERs can also be used for detecting gene expression in alternative cell types, such as mammalian cell culture, we modified the *E. coli* NTER design to function in HEK293 cells (Figure 2.2). We did this by making two key changes: 1) we replaced the OsmY bacterial secretion domain with a human secretion tag (IFNα2) [55], and 2) made two mutations to the Smt3 domain that make it more resistant to intracellular degradation in mammalian cells (SUMOstar)[56]. We then cloned several different barcoded versions of this mammalian-optimized NanoporeTER design (mNTER) into vector XYZ under the control of a constitutive CMV promoter, and performed experiments in which varying combinations of these vectors were transfected into HEK293 suspension cultures. Results from these experiments are shown in Figure 2.9. Specifically, we could detect mNTERs from media supernatant collected XX days after the cultures were transfected with either with one (Y00), two (Y00,Y02), or three (Y00,Y03,Y07) different mNTER barcodes. Importantly, the number of mNTER counts for each barcode class was reflective of the barcode combinations that were introduced into each of the cultures, as shown by the mNTER classification counts being substantially higher for the

transfected barcode classes relative to classes included in the classifier but absent from the experiment. We note that while mNTERs could be detected over background classification levels directly from the raw media supernatant with no further processing (Figure 2.9), superior classification results were obtained from media samples post-IMAC enrichment (Figure 2.9), suggesting media contaminants in cell culture led to higher levels of mNTER misclassification events.

Finally, we assessed the experimental throughput of these types of protein measurements on a MinION, as they go beyond the platform's normal mode of operation. We found that experiments performed on new R9.4.1 flow cells typically started with >450 functional pores (Figure 2.11). Pores became non-functional (e.g. permanently clogged or ruptured bilayer) at a rate of ~1 pore/minute over the first two hours of flow cell lifetime, which was typically the longest amount of time we used a single flow cell. The usual MinION runtime for each of our experimental analyses was ~10 minutes per sample, along with an additional 5-minute wash step between samples. Thus, close to 10 experiments could be performed per flow cell while maintaining a high number of functional pores (>300). We note, however, that the required runtime for each experiment is dependent on NTER concentrations, as lower concentrations require longer runtimes in order to collect enough observations to determine the average capture rate confidently. For example, to determine the approximate runtimes required to confidently arrive at the true mean NTER time between captures at varying concentrations, we computed the 95% Confidence Interval (CI) with respect to the number of observed captures (Figure 2.12). Overall, we found that a few hundred NTER reads was sufficient for CI convergence over the two orders of magnitude concentration range we tested (1 uM to 0.01 uM). For high concentrations, this number of reads can be collected in ~1 minute. While, at lower

concentrations, it would take on the order of tens of minutes based on our extrapolations (Figure 2.13). Precisely how these numbers relate to absolute expression levels remains to be elucidated as more work will be needed to determine how NTER measurements correlate to quantitative expression measurements such as those that can be obtained with RNA and Ribo-Seq methods[57-59].

2.3 Discussion

In conclusion, we have laid the foundations for a new class of multiplexable protein reporters (NanoporeTERs) that can be analyzed using a commercially available nanopore sensor array, the ONT MinION. Although our current NanoporeTER classification pipeline is currently done post-runtime, real-time results will be possible with the addition of software that runs concurrently (or within) the MinION operating scripts[60]. While we have characterized here a set ~20 orthogonal NanoporeTERs, examination of the sparsity of the current NTER signal space suggests that finding additional orthogonal barcodes is achievable (Figure 2.14). Although it is difficult to confidently estimate an upper bound to this signal-orthogonal barcode space at this juncture, in comparing the current NTER signal space to the DNA signal space of an R9.4 MinION flow cell (which has a sequence space of 4^6 unique kmers), we see that our current set of NTERs occupy a similar or even larger total area of the ionic current level vs std feature space (Figure 2.14). The potential to push this space further is reasonable considering that the homopolymer mutants only spanned 9 (residues 3-11) out of the 17 total positions that contribute to the barcode's ionic current signature. We are also able to consider more signal features than the DNA kmer model (owing to the long dwell times we can achieve by stalling the NTER barcode in the pore), such as the signal minimum and maximum. These additional signal features are useful for

classification, as shown by the Random Forest classifier's feature importances (Figure 2.14). Signal features beyond this limited set can be manually extracted or learned by deep learning models that operate on the raw signal, such as our CNN approach, which is perhaps more scalable as the number of barcodes increases. This is somewhat analogous to the advances in nanopore sequencing of DNA achieved with deep learning-based models[61]. Ultimately, barcode space could be expanded through several different strategies, including: 1) high-throughput methods to empirically characterize more barcode sequences for classifier training, 2) engineering NanoporeTERs to contain multiple barcode regions that can be consecutively read out with the aid of processive motor proteins[48,49] or voltage-mediated translocation[62], which would allow the number of orthogonal NTERs to scale exponentially with the number of individually characterized barcodes, and 3) semi-supervised machine learning models trained to accurately predict the sequence of empirically uncharacterized NTER barcodes given only their nanopore signal[63]. NanoporeTERs should also not be limited to use in *E. coli* and *HEK293* cells, as their modular design requires that only the secretion domain be modified, of which many different N-terminal secretion domains have been characterized in a range of diverse organisms[64-66].

We foresee many potential NanoporeTER applications, including simultaneously reading the protein-level outputs of many genetically engineered circuit components in one-pot, enabling more efficient debugging and tuning than current analysis methods. For instance, in comparison to traditional sets of fluorescent protein reporters, NanoporeTERs have a (potentially much) larger sequence and signal space that allows for the simultaneous analysis of a greater number of unique genetic elements in a single experiment (multiplexing). And while RNA-seq is another strategy that can be used to measure the transcriptional output of many circuits in parallel with

high-throughput DNA sequencing technology[67], our method has the advantages of 1) little to no sample preparation, which makes it more amenable to automation[68-70] and reduces both time to analysis (latency) and cost, and 2) direct detection of outputs at the protein level. The latter advantage opens new opportunities to custom engineer reporters with NTER barcodes that can report on both protein expression and specific post-translational modifications simultaneously. We anticipate this capability will be especially useful as the nascent field of synthetic protein-level circuit engineering advances[71].

2.4 Methods

NanoporeTER construction, cloning, expression, and purification

The initial NanoporeTER protein was constructed with a gBlock (Integrated DNA Technologies) composed of the Smt3 and tail sequence and cloned into plasmid pCDB180 downstream of the OsmY domain. The Q5 site-directed mutagenesis method (New England Biolabs) was used to generate the different NTER barcode mutants. All cloning was performed using the 5-alpha competent E. coli strain following NEB's cloning protocol (New England Biolabs). Sequence verification was obtained through Genewiz Inc. Expression of the NanoporeTER protein was done in BL21 (DE3) E. coli strain using Overnight Express instant TB medium (Novagen).

Proteins were purified via immobilized metal affinity chromatography (IMAC) using TALON metal affinity cobalt resin (Takara). The purification used the associated buffer set from Takara, following their specified protocol. Proteins were concentrated using Amicon Ultra 0.5 mL centrifugal filters with Ultracel 30K (Amicon). The final concentration of proteins averaged ~7 mg/ml from 5 mL overnight cultures. The purified proteins were stored for long-term storage at -80C in 10 uL aliquots, as well as for short-term storage at 4C.

E.coli raw culture mixing experiments

Cultures were picked from single colonies on plates and used to inoculate 3mL LB supplemented with 0.5 mM IPTG and kanamycin (induced), or 3mL LB supplemented with 0.2% glucose and kanamycin (inhibited). After overnight incubation at 37C with shaking, cultures were equally mixed together in a total volume of 45uL, 50uL 4X C17 buffer (2 M KCl, 100 mM HEPES, pH 8), and 105 uL water (total volume 200uL). This solution was then immediately loaded into a MinION flow cell for analysis.

E. coli expression time course

Time course experiments were performed by diluting 30uL of overnight cultures (LB) into 3mL fresh LB supplemented with 0.5 mM IPTG and kanamycin (induced), or 3mL fresh LB supplemented with 0.2% glucose and kanamycin (inhibited). The cultures were placed in a shaker/incubator at 37C to allow for culture growth. Samples were then collected at 2, 4, 6, and 21-hour time-points. At each time-point, cultures were equally mixed together in a total volume of 10 uL, 50uL 4X C17 buffer, and 140 uL water (total volume 200uL). This solution was then immediately loaded into a MinION flow cell for analysis.

HEK293 transfection

To clone mammalian NTERs, we used a mammalian expression vector consisting of a CMV enhancer and CMV promoter driving expression of mCherry and a N-terminal nuclear localization signal (NLS). First, we replaced the NLS by inverse PCR of the vector at the edges of the NLS and assembled via Gibson assembly (NEB) with a gBlock (IDT) comprising the sequence for an IFNalpha2 secretion tag and 10x His tag. To add the mNTER to the C terminus

we used inverse PCR at the mCherry stop codon and assembled via Gibson assembly (NEB) with a gBlock (IDT) synthesized with the NTERY00 sequence. To generate mNTER variants, we used inverse PCR at the variable site using primers with overlapping extensions containing the new NTER barcode and compatible overhangs. All mNTERs were transformed and cultured in DH5a-electrocompetent cells (NEB) and verified with Sanger sequencing through Genewiz Inc.

Cells used for transfection experiments were FreeStyle 293-F cells (Gibco, ThermoFisher, no. R79007) and were grown in FreeStyle 293 expression medium (Gibco) with no added antibiotic. The day before transfection, cells were seeded at a density of 500,000 ml^{-1} to reach a density of 1 million ml^{-1} the following day. On the day of transfection, cells were transfected with 1 µg of DNA per 1 million cells using a lipid-based method of transfection. Cells were then left to express for 3 days on a shaker platform, shaking at 135 r.p.m. at 37 ºC and supplemented with 8% CO_2, before collection of medium supernatant for subsequent nanopore analysis or IMAC purification.

Nanopore analysis of HEK293 mNTER expression was conducted by mixing 5–10 µl of raw supernatant, 50 µl of 4x C17 buffer and 140–145 µl of water (total volume, 200 µl). This solution was then immediately loaded into a MinION flow cell for analysis. For IMAC-purified samples, protein was diluted to a final concentration of 0.02 uM total protein in 1x C17 before loading into a MinION flow cell for analysis.

MinION experiments

All experiments were performed with unmodified R9.4.1 MinION flow cells (Oxford Nanopore

Technologies) by diluting analyte solution into C17 buffer for a final concentration of 0.5M KCl and 25mM HEPES (pH 8), into the flow cell priming port. Flow cells were run on the MinION at a temperature of 30°C and a run voltage of -180mV with a 10khz sampling frequency and 15 second static flip frequency. Use of a modifiable MinKNOW script (available from ONT) enabled voltage flipping cycle parameters to be set as well as collection of raw current data across the entire run. Individual flow cells could be reused for different analytes after flushing them with 1mL C17 buffer three times between experiments. Flow cells were stored at 4°C in C18 buffer (150mM potassium ferrocyanide, 150mM potassium ferricyanide, 25mM potassium phosphate, pH8) when not in use.

Nanopore Signal Analysis, Quantification, and Classification

The analysis pipeline for a NanoporeTER sequencing run begins with extracting the segments of the raw nanopore signal that contain capture events. A capture is defined as a region where the signal current falls below 70% of the open pore current for a duration of at least one millisecond. The fractional current values (as compared to open pore current) computed from the segmentation process, as well as the start and end times of each capture, are saved in separate data files. This information is then passed through a general filter that separates putative NanoporeTER captures from noise captures based on features of the normalized raw current (mean, standard deviation, minimum, maximum, median) as well as the duration of the capture. Captures that pass this initial filter are then fed into a classifier and classified as a specific NTER barcode or a background/noise blockade. The metadata for the captures within each NTER class are subsequently fed to a quantifier which calculates the average time elapsed between those captures and (optionally) converts this time to the predicted NTER concentration using a

standard curve. An alternative method of quantification is to calculate the number of reads per class per active pore per minute (reads/pore-minute or RPMs). In addition to the NTER data sets, a background/noise class data set was also used in training the models to recognize data generated from non-NTER-specific pore blockages that made it through the filtering step. This data was collected from experiments in which only running buffer, LB media, or NTER-free E. coli cultures were loaded into the flow cell.

We explored two different classifiers for NTER barcode discrimination. The first, a Random Forest model, was implemented in scikit-learn (sklearn.ensemble.RandomForestClassifier). The second classifier was a CNN implemented in PyTorch. An 80/20 train/test split was used to generate the classification accuracy estimates and confusion matrix results. For both models, only the first two seconds of each capture were considered for analysis. The Random Forest was trained on an array composed of the mean, standard deviation, minimum, maximum, and median of that two second window. Default Random Forest hyperparameters were modified to: n_estimators=300 and max_depth=100. The CNN used the two seconds of raw signal directly as input following reshaping of the 1D signal into a 2D structure. The neural network was composed of four 2D convolutional layers each with ReLU activation and max pooling. These were followed by a fully connected layer which had a log-sigmoid activation function, and then a final output layer of the same size as the number of NTER classes (plus noise class) considered in the experiment. This output layer can be interpreted as the confidence scores associated with each class, which can also be applied as a confidence threshold filter (e.g. only assigning labels for events with 95% confidence in a single class). Full model details and code can be found at https://github.com/uwmisl/NanoporeTERs.

While both classifiers (Random Forest and CNN) achieved similar accuracies on the train/test data, the main difference of the CNN is that it does not require any manual feature extraction. While, for example, the Random Forest used specific signal features (current blockade mean, median, max, min, and std) as input. Using the raw signal directly as input, as the CNN does, is then potentially a more generalizable classification approach as the number of barcode sequences (and thus potential differences in signal features) increases.

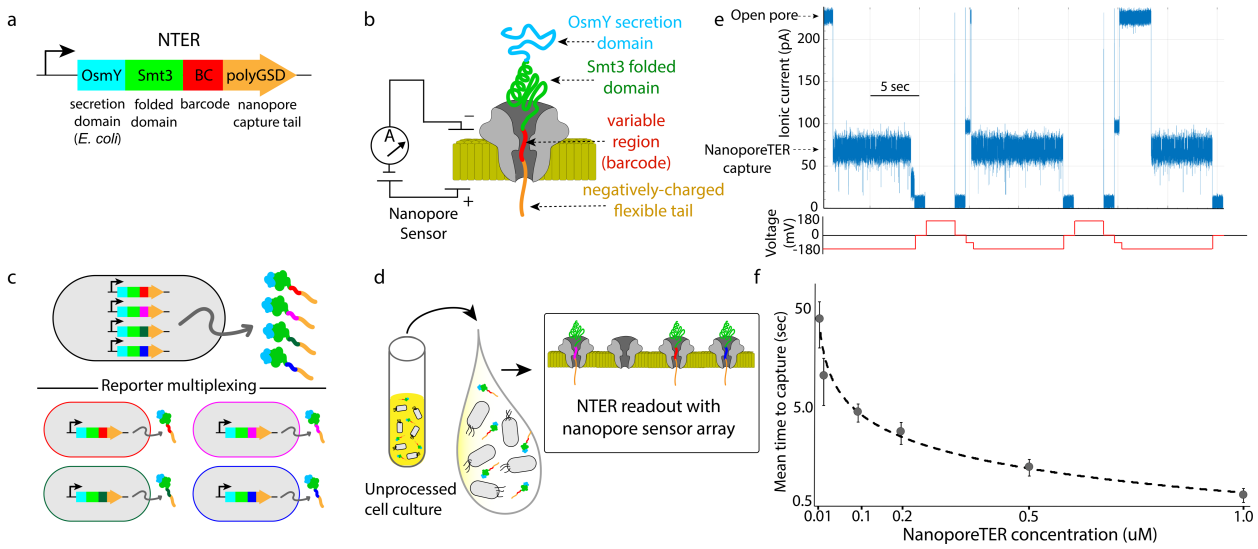


Figure 2.1: Nanopore-addressable protein Tags Engineered as Reporters (NanoporeTERs). **a**, Gene schematic of NanoporeTER (NTER) design. Cyan: OsmY; promotes extracellular secretion of the reporter protein in *E. coli*. Green: Smt3; folded domain stalls translocation of the protein through the pore and facilitates a "static read" of the NTER barcode within the nanopore sensor. Red: barcode; region of the protein that is held within the sensitive region of the nanopore lumen upon which the changes to the barcode sequence manifest changes to the nanopore ionic current signal. Orange: polyGSD tail; long, flexible, negatively charged C-terminal domain promotes electrophoretic capture of the NTER into the nanopore under an applied voltage. **b**, Schematic of a NanoporeTER captured within a nanopore. **c**, NanoporeTERs are designed to enable multiplexed readout of protein expression, with the potential to report on multiple outputs within a single strain (top), or report on expression across multiple strain types in a one-pot mix (bottom). **d**, Secretion of the NanoporeTERs into the extracellular medium eliminates the need for any sample preparation prior to loading into the nanopore sensor array flow cell. **e**, Example of raw nanopore data generated from a single nanopore showing repeated captures and ejections of NTERY00. **f**, Concentration standard curve showing the relationship between NanoporeTER concentration within a flow cell versus the average time between captures or "reads." Error bars represent standard deviation

NTER amino acid sequence (*E. coli*):

MTMTRLKISKTLLAVMLTSAVATGSAYAENNAQTTNESAGQKVDSSMNKVGNFMDDSAITAKVKAALVDHDNIKSTDISVKTDQKVVT
LSGFVESQAQAEEAVKVAKGVEGVTSVSDKLHVRDAKEGSVKGYAGDTATTSEIKAKLLADDIVPSRHVKVETTDGVVQLSGTVDSQ
AQSDRAESIAKAVDGVKSVKNDLKTK MGHHHHHHHHHHHGS LQDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRR
LMEAFAKRQGKEMDSLRFLYDGIRIQADQAPEDLDMEDNDIIEAHREQI XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX DGGSSGG
SGGDGSSGDGGSDGDSDGSDGDGDSDGDDGGDDEDDGSDD*

mNTER amino acid sequence (*HEK293*):

MALTFALLVALLVLSCKSSCSV GGHHHHHHHHHH VSKGEEDNMAIIKEFMRFKVHMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKV
TKGGPLPFAWDILSPQFMYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIYKVKLRGTNFPSDGPV
MQKKTMGWEASSERMYPEDGALKGEIKQRLKLKDGGHYDAEVKTTYKAKKPVQLPGAYNVNIKLDITSHNEDYTIVEQYERAEGRHS
TGGMDELYKSGS LQDSEVNQEAKPEVKPEVKPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLTFLYDGIEIQADQ
APEDLDMEDNDIIEAHREQI XXXXXXXXXXXXXXXXXXXXXXXXXXXXXX DGGSSGGSGGDGSSGDGGSDGDSDGSDGDGDSD
GDDGGDDEDDGSDD*

```
NTERY00 X = GGGGSSGGSGGSGSSGDGGSSGGSGGSGSSG
```

YYY mapping mutants:

```
NTERY01 X = YYYGSSGGSGGSGSSGDGGSSGGSGGSGSSG
NTERY02 X = GGYYYSGGSGGSGSSGDGGSSGGSGGSGSSG
NTERY03 X = GGGGYYYGSGGSGSSGDGGSSGGSGGSGSSG
NTERY04 X = GGGGSSYYYGGSGSSGDGGSSGGSGGSGSSG
NTERY05 X = GGGGSSGGYYYSGSSGDGGSSGGSGGSGSSG
NTERY06 X = GGGGSSGGSGYYYSSGDGGSSGGSGGSGSSG
NTERY07 X = GGGGSSGGSGGSYYYGDGGSSGGSGGSGSSG
NTERY08 X = GGGGSSGGSGGSGSYYDGGSSGGSGGSGSSG
NTERY09 X = GGGGSSGGSGGSGSSGDYYSSGGSGGSGSSG
NTERY10 X = GGGGSSGGSGGSGSSGDGYYYGGSGGSGSSG
NTERY11 X = GGGGSSGGSGGSGSSGDGGSYYYSGGSGSSG
NTERY12 X = GGGGSSGGSGGSGSSGDGGSSGYYYGSGSSG
NTERY13 X = GGGGSSGGSGGSGSSGDGGSSGGSYYYGSSG
NTERY14 X = GGGGSSGGSGGSGSSGDGGSSGGSGGYYYSG
NTERY15 X = GGGGSSGGSGGSGSSGDGGSSGGSGGSGYYY
```

Homopolymer mutants:

```
NTER A X = GGAAAAAAAAASGSSGDGGSSGGSGGSGSSG
NTER D X = GGDDDDDDDDDSGSSGDGGSSGGSGGSGSSG
NTER E X = GGEEEEEEEEESGSSGDGGSSGGSGGSGSSG
NTER G X = GGGGGGGGGGGSGSSGDGGSSGGSGGSGSSG
NTER H X = GGHHHHHHHHHSGSSGDGGSSGGSGGSGSSG
NTER M X = GGMMMMMMMMSGSSGDGGSSGGSGGSGSSG
NTER N X = GGNNNNNNNNNSGSSGDGGSSGGSGGSGSSG
NTER P X = GGPPPPPPPPPSGSSGDGGSSGGSGGSGSSG
NTER Q X = GGQQQQQQQQQSGSSGDGGSSGGSGGSGSSG
NTER R X = GGRRRRRRRRRSGSSGDGGSSGGSGGSGSSG
NTER S X = GGSSSSSSSSSSGSSGDGGSSGGSGGSGSSG
NTER T X = GGTTTTTTTTTSGSSGDGGSSGGSGGSGSSG
```

PKA motif mutants:

```
NTER PKA                X = GGRRGSYYSGGSGSSGDGGSSGGSGGSGSSG
NTER PKA_phosphomimetic X = GGRRGEYYSGGSGSSGDGGSSGGSGGSGSSG
```

Tail mutants:

```
NTER truncated 40 tail = GDGSSGDGGSDGDSDGSDGDGDSDGDDGGDDEDDGSDD* (-18 charge)
NTER truncated 20 tail = GGGGSSGGSGGSGSSGDGGSSDGSSGDGGSDGDSDGSDGDGDSDGDDGGDDEDDGSDD*
(-19 charge)
NTERY00 standard tail =
GGGGSSGGSGGSGSSGDGGSSGGSGGSGSSGDGGSSGGSGGDGSSGDGGSDGDSDGSDGDGDSDGDDGGDDEDDGSDD* (-20
charge)
```

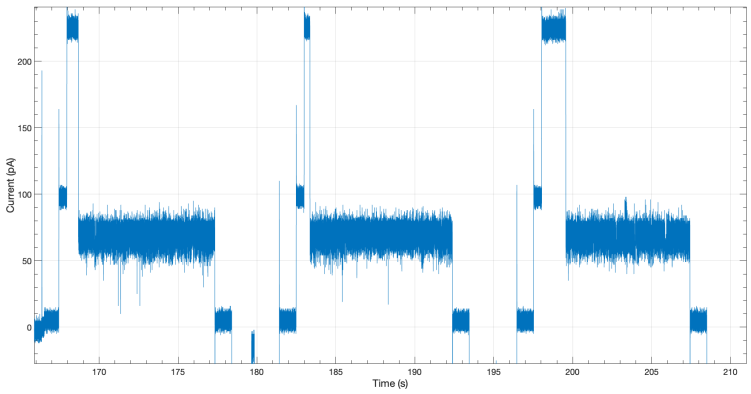
Figure 2.2: NanoporeTER sequences. Blue: Secretion domain. Black: His-tag. Green: Smt3 domain (Smt3* in mNTER). Red: Variable region (contains barcode). Orange: PolyGSD tail. Purple: mCherry domain (only present in mNTER for convenience).

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTT
CTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTTTACGGCACCTCGACCCCAAAAAACT
TGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAAC
AACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAA
CAAAATATTAACGTTTACAATTTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAATTA
ATTCTTAGAAAAACTCATCGAGCATCAAATGAAACTGCAATTTATTCATATCAGGATTATCAATACCATATTTTTGAAAAAGCCGTTTCTGTAATGAAGGAGAAAACTCAC
CGAGGCAGTTCCATAGGATGGCAAGATCCTGGTATCGGTCTGCGATTCCGACTCGTCCAACATCAATACAACCTATTAATTTCCCCTCGTCAAAAATAAGGTTATCAAGTG
AGAAATCACCATGAGTGACGACTGAATCCGGTGAGAATGGCAAAAGTTTATGCATTTCTTTCCAGACTTGTTCAACAGGCCAGCCATTACGCTCGTCATCAAAATCACTCG
CATCAACCAAACCGTTATTCATTCGTGATTGCGCCTGAGCGAGACGAAATACGCGATCGCTGTTAAAAGGACAATTACAAACAGGAATCGAATGCAACCGGCGCAGGAACA
CTGCCAGCGCATCAACAATATTTTCACCTGAATCAGGATATTCTTCTAATACCTGGAATGCTGTTTTCCCGGGGATCGCAGTGGTGAGTAACCATGCATCATCAGGAGTAC
GGATAAAATGCTTGATGGTCGGAAGAGGCATAAATTCCGTCAGCCAGTTTAGTCTGACCATCTCATCTGTAACATCATTGGCAACGCTACCTTTGCCATGTTTCAGAAACA
ACTCTGGCGCATCGGGCTTCCCATACAATCGATAGATTGTCGCACCTGATTGCCCGACATTATCGCGAGCCCATTTATACCCATATAAATCAGCATCCATGTTGGAATTTA
ATCGCGGCCTAGAGCAAGACGTTTCCCGTTGAATATGGCTCATAACACCCCTTGTATTACTGTTTATGTAAGCAGACAGTTTTATTGTTCATGACCAAAATCCCTTAACGT
GAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTA
CCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTA
GGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCA
AGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAG
CTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGG
TATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTA
CGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGC
AGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCACACCGCATATATGGTGCA
CTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTGCGCCCCGACACCCGCCAACACCCGCTGACGC
GCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCA
GCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCTGTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGAT
AAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCACTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCT
CACGATACGGGTTACTGATGATGAACATGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATG
CCAGCGCTTCGTTAATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGACT
TTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGATTCATTCTGCTAACC
AGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGATAATGGCCTGCTTCTCGCCGAAACGTTT
GGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCCGATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAAT
GACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTCATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTT
GAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAATGAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCA
GCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCG
CCTGGCCCTGAGAGAGTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTAT
CGTCGTATCCCACTACCGAGATGTCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCATCGCAGTGGGAA
CGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTTGATTGCGAGTGAGATATTTATGCC
AGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAATGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTT
CATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCCGGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCG
GATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGTGCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCA
GTTGATCGGCGCGAGATTTAATCGCCGCGACAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTG
CCACGCGGTTGGGAATGTAATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAG
AGACACCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTTTTGCGCC
ATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCGCCGCCGCAAGGAATGGTGCA
TGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGAGCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTG
ATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCGTCCGGCGTAGAGGATCGAGATCGATCTCGATCCCGCGAAAT**TAATACGA**
CTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTTAACTTTAAGAAGGAGATATACATATGACTATGACAAGACTGAAGATTTCGAAAA
CTCTGCTGGCTGTAATGTTGACCTCTGCCGTCGCGACCGGCTCTGCCTACGCGGAAAACAACGCGCAGACTACCAATGAAAGCGCAGGGCAAAAAGTCGATAGCTCTATGA
ATAAAGTCGGTAATTTCATGGATGACAGCGCCATCACCGCGAAAGTGAAGGCGGCCCTGGTGGATCATGACAACATCAAGAGCACCGATATCTCTGTAAAAACCGATCAAA
AAGTCGTGACCCTGAGCGGTTTCGTTGAAAGCCAGGCCCAGGCCGAAGAGGCAGTGAAAGTGGCGAAAGGCGTTGAAGGGGTGACCTCTGTCAGCGACAAACTGCACGTTC
GCGACGCTAAAGAAGGCTCGGTGAAGGGCTACGCGGGTGACACCGCCACCACCAGTGAAATCAAAGCCAAACTGCTGGCGGACGATATCGTCCCTTCCCGTCATGTGAAAG
TTGAAACCACCGACGGCGTGGTTCAGCTCTCCGGTACCGTCGATTCTCAGGCACAAAGTGACCGTGCTGAAAGTATCGCCAAAGCGGTAGATGGTGTGAAAAGCGTTAAAA
ATGATCTGAAAACTAAGATGGGTCACCACCACCACCACCACCACCACCACCACCACGGTAGCCTGCAAGATTCGGAAGTCAACCAAGAAGCAAAGCCGGAAGTCAAGCCGGAAG
TGAAACCGGAAACCCATATTAACCTGAAAGTTAGTGACGGCAGCTCTGAAATTTTCTTTAAGATCAAAAAGACCACGCCGCTGCGTCGCCTGATGGAAGCGTTTGCCAAAC
GTCAGGGCAAGGAAATGGATAGCCTGCGTTTCCTGTATGACGGTATTCGCATCCAGGCAGATCAAGCGCCGGAAGACCTGGACATGGAAGACAACGACATCATTGAAGCCC
ACCGTGAACAGATTGGTGGTGGTGGTTCGAGCGGCGGGAGCGGAGGGTCTGGGTCTAGCGGGGATGGTGGGTCGTCTGGGGGGAGCGGCGGTTCGGGTTCATCTGGTGACG
GCGGTTCCTCGGGTGGTTCCGGGGGAGATGGTTCTTCCGGAGATGGCGGGAGTGATGGGGACTCTGATGGTTCGGATGGTGACGGGGACAGTGATGGTGATGACGGAGGGG
ACGACGAGGATGACGGTTCAGACGATTAATGAGGCGGCTCGAGGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATA*ACTAGC*
*ATAACCCCTTGGGGCCTCT**AAACGGGTCTTGAGGGGTTTTTTG***CTGAAAGGAGGAACTATATCCGGAT

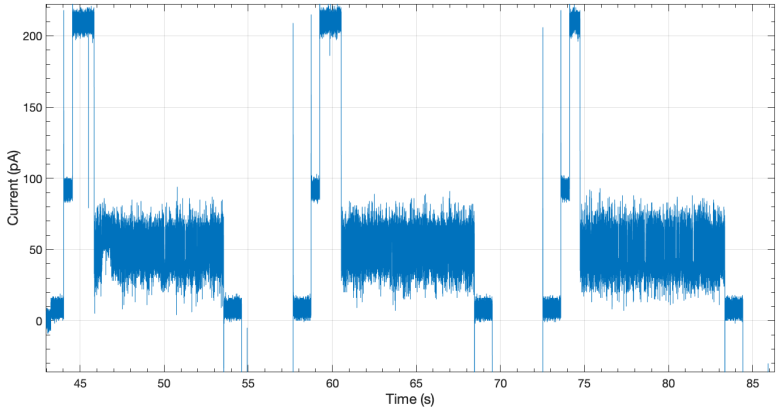
Figure 2.2 (cont.): NanoporeTER (Y00) expression vector DNA sequence (based on the pCDB180 plasmid backbone). Annotations: T7 promoter and lacO (bold), RBS (underlined), NanoporeTER ORF (box), and T7 terminator (italic).

Figure 2.3 Raw nanopore ionic traces: Representative MinION nanopore ionic traces of NTER barcodes analyzed in this work. Three consecutive "reads" are included in each trace.

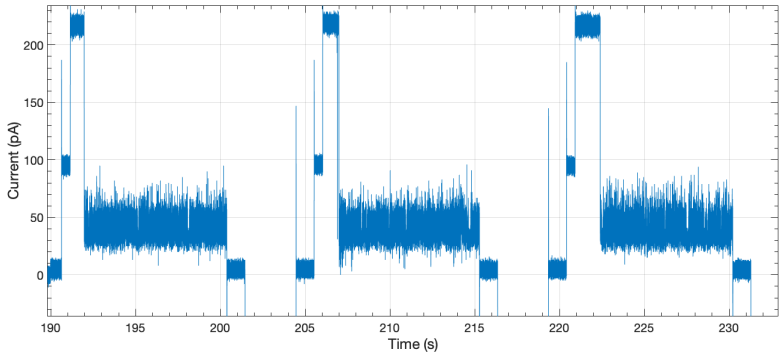
NTER00



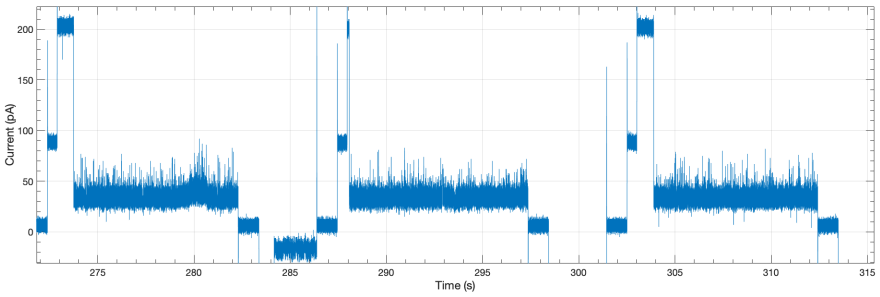
NTER01



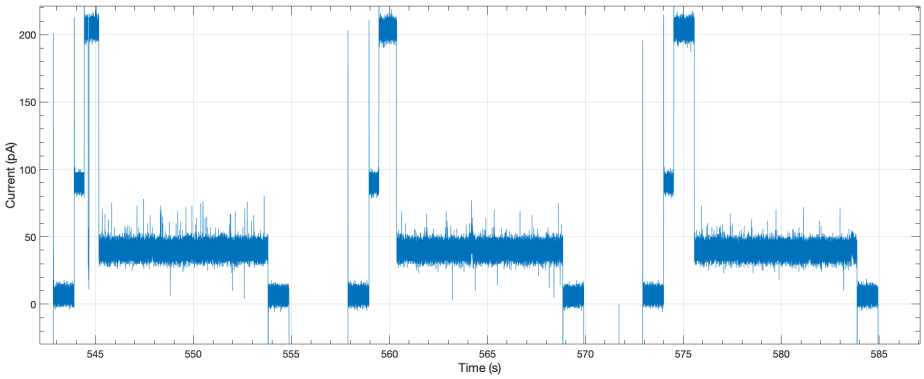
NTER02



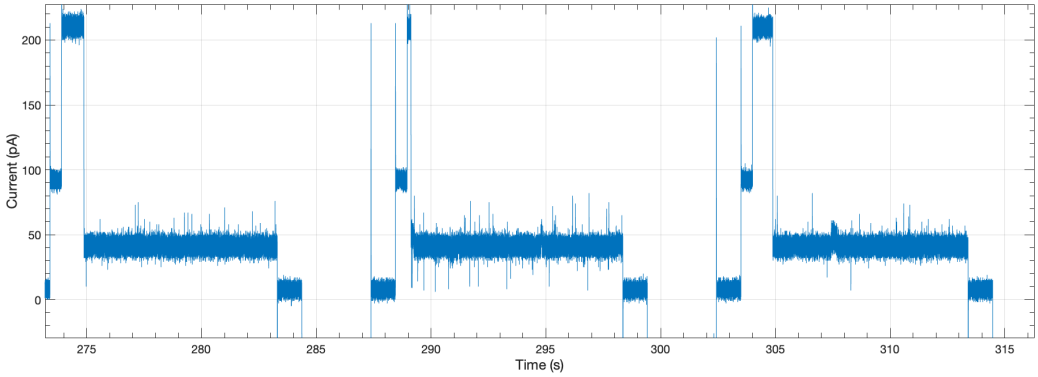
NTER03



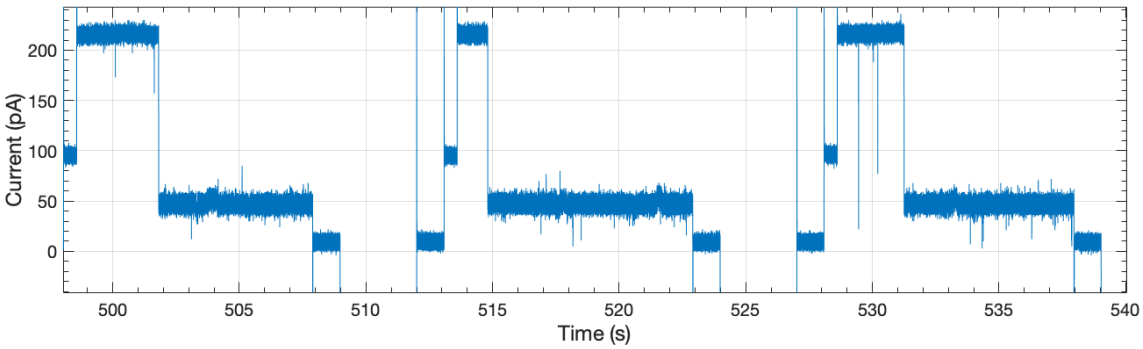
NTER04



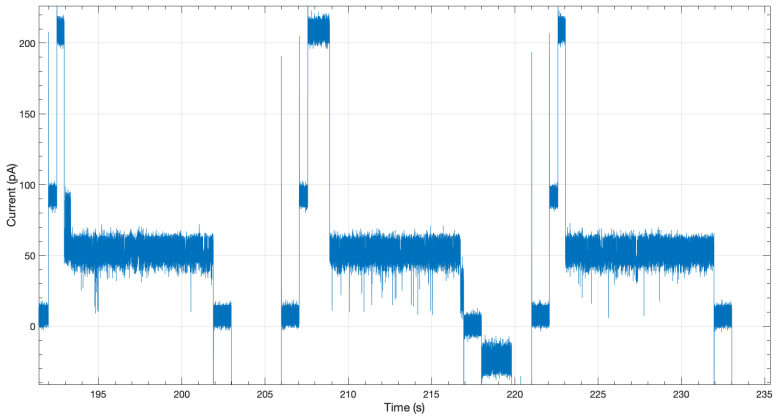
NTER05



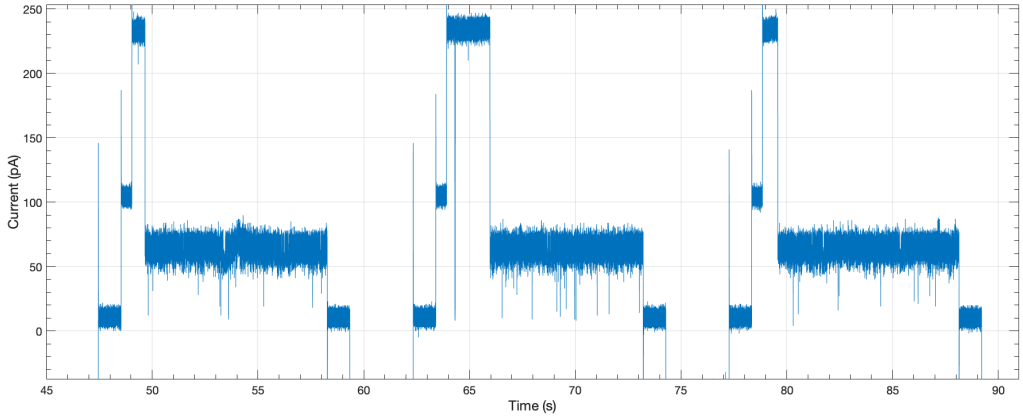
NTER06



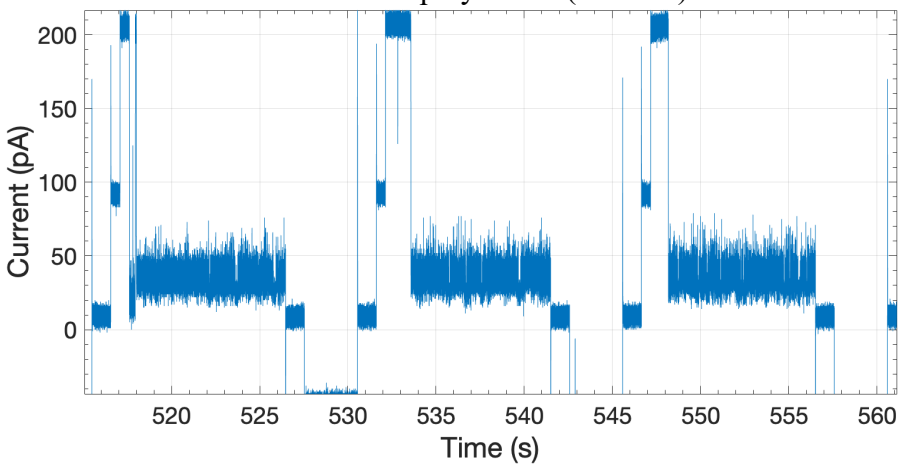
NTER07



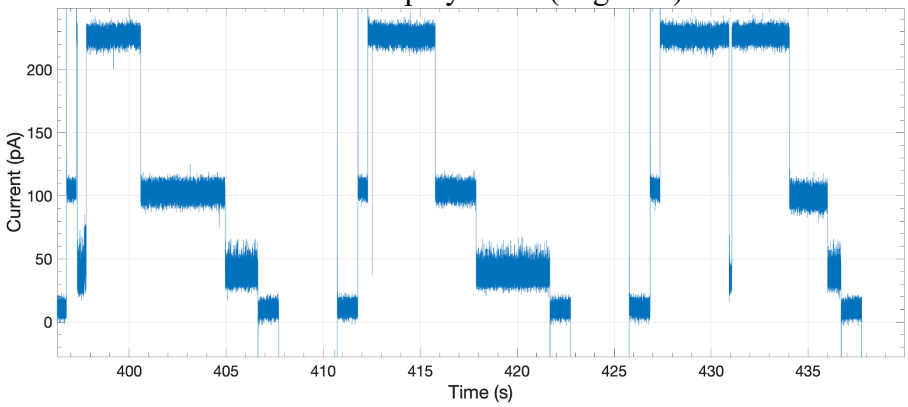
NTER08



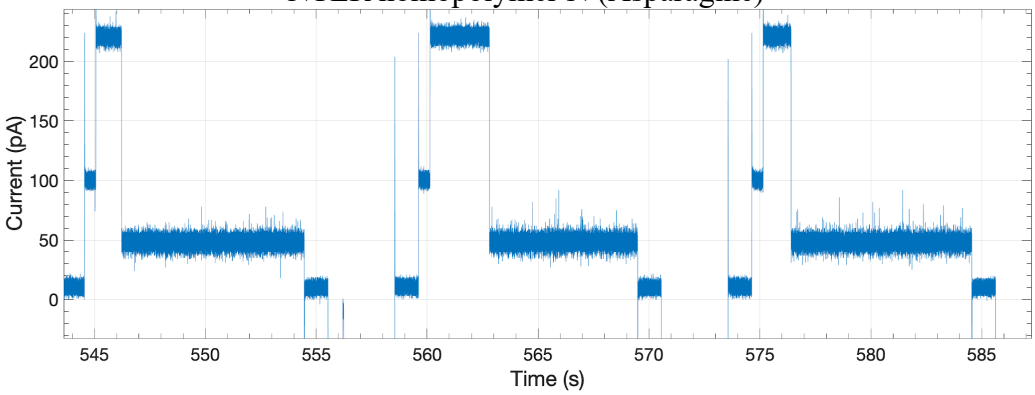
NTER homopolymer A (Alanine)



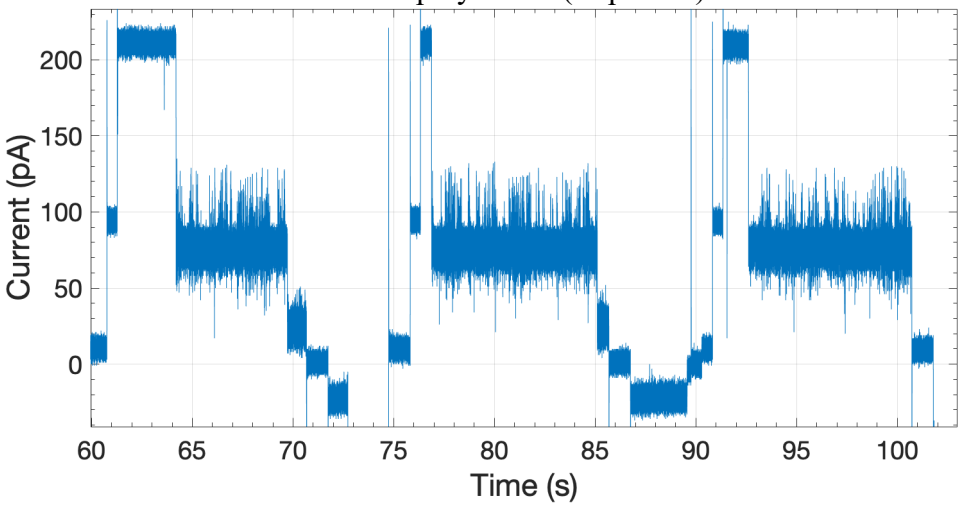
NTER homopolymer R (Arginine)



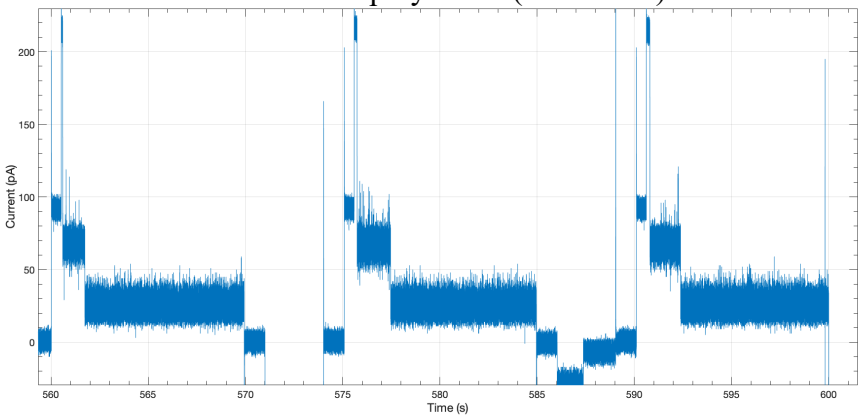
NTER homopolymer N (Asparagine)



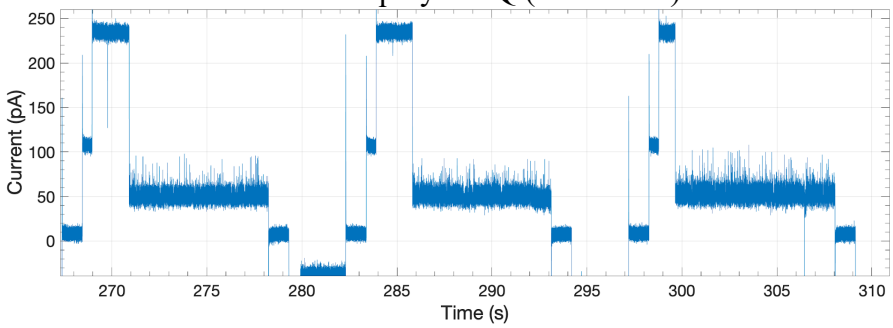
NTER homopolymer D (Aspartate)



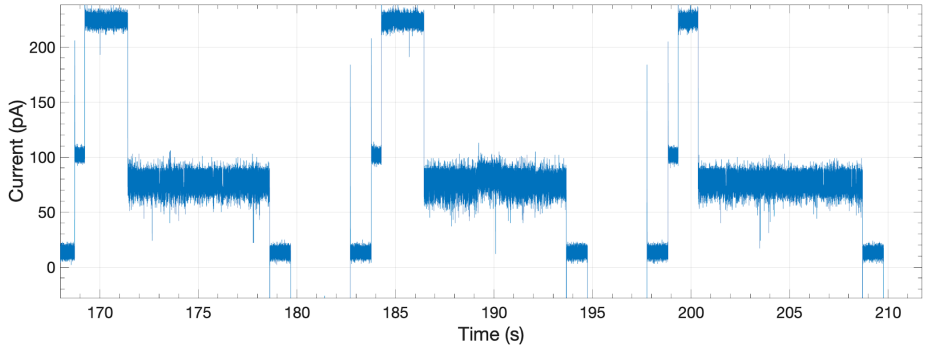
NTER homopolymer E (Glutamate)



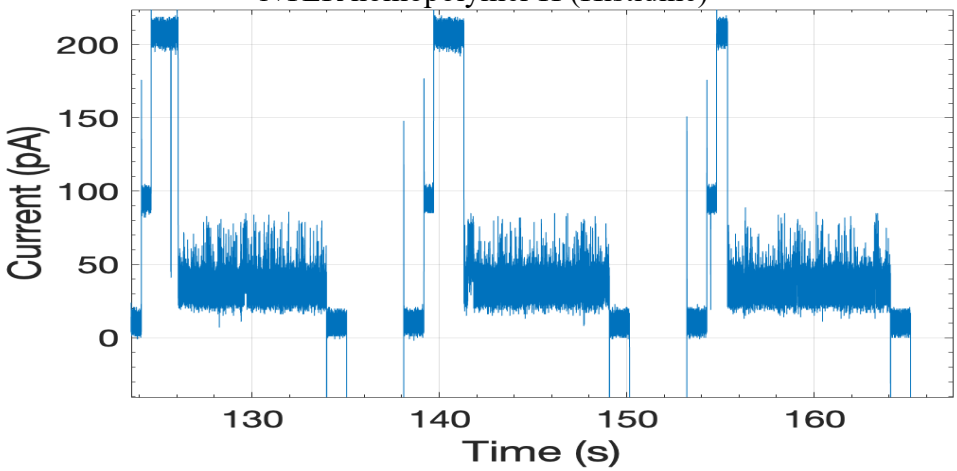
NTER homopolymer Q (Glutamine)



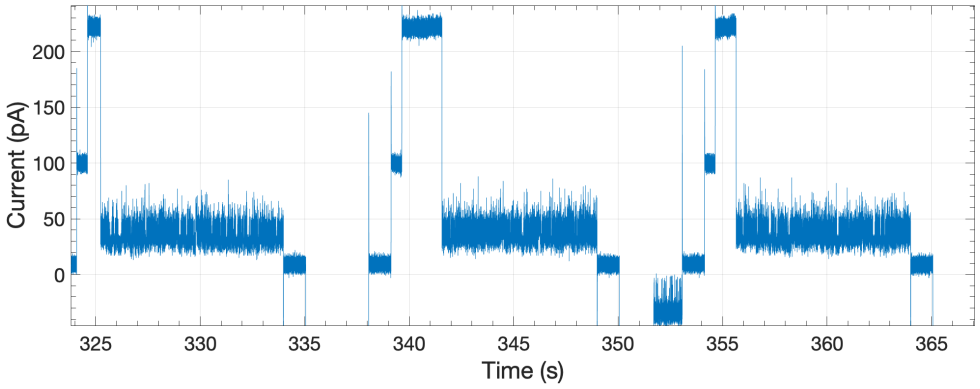
NTER homopolymer G (Glycine)



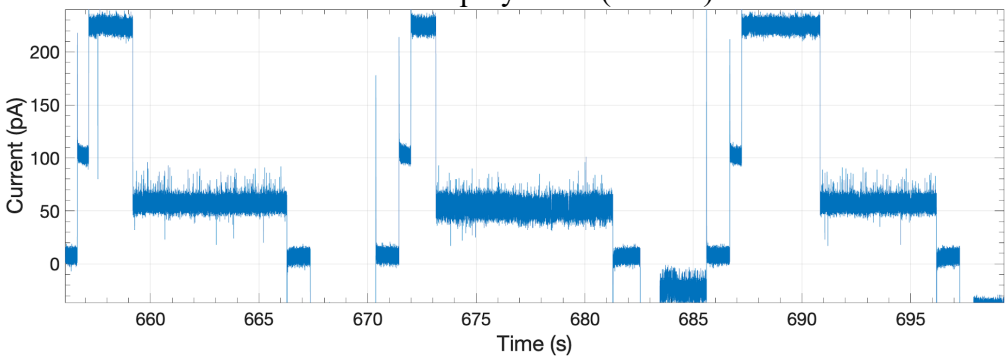
NTER homopolymer H (Histidine)



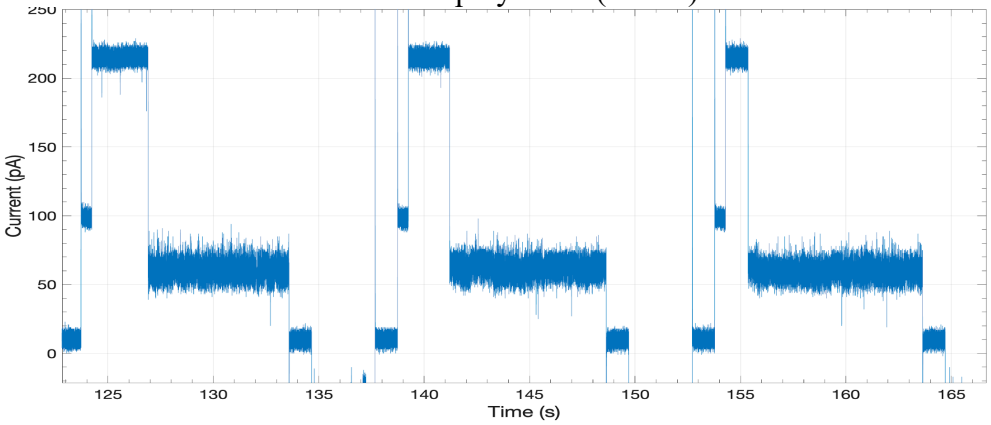
NTER homopolymer M (Methionine)



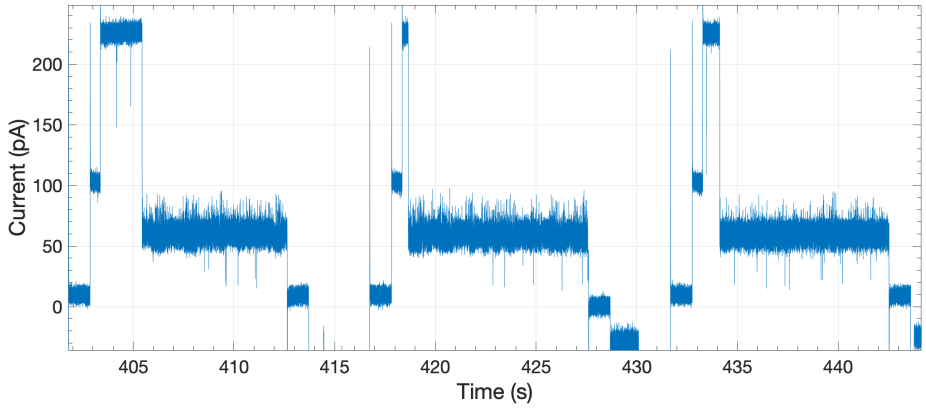
NTER homopolymer P (Proline)



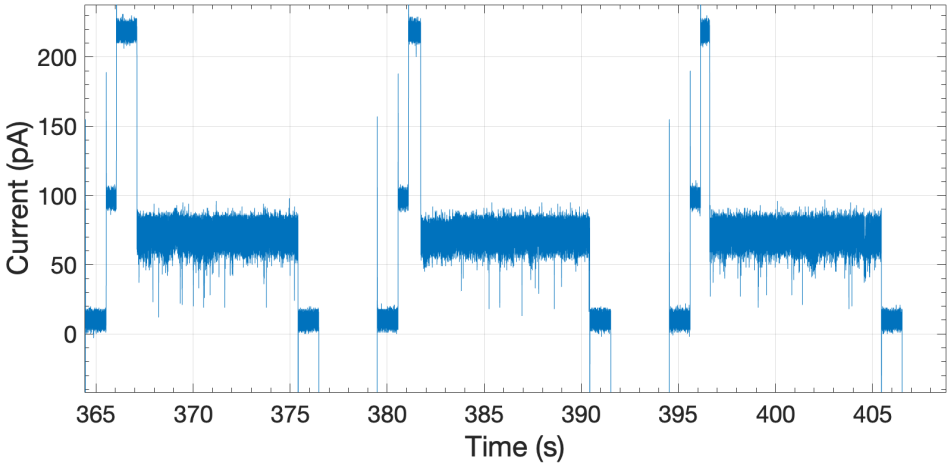
NTER homopolymer S (Serine)



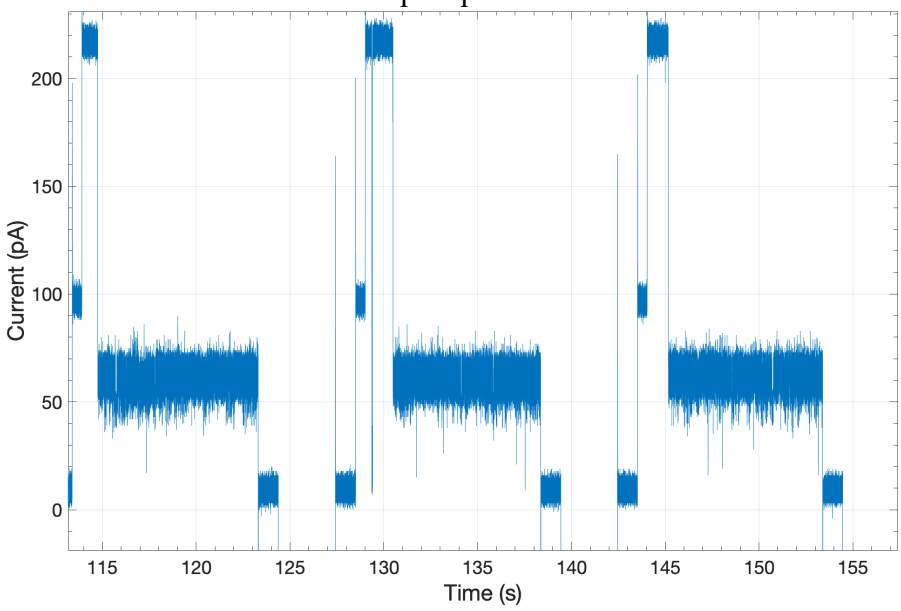
NTER homopolymer T (Threonine)



NTER PKA motif



NTER PKA phosphomimetic motif



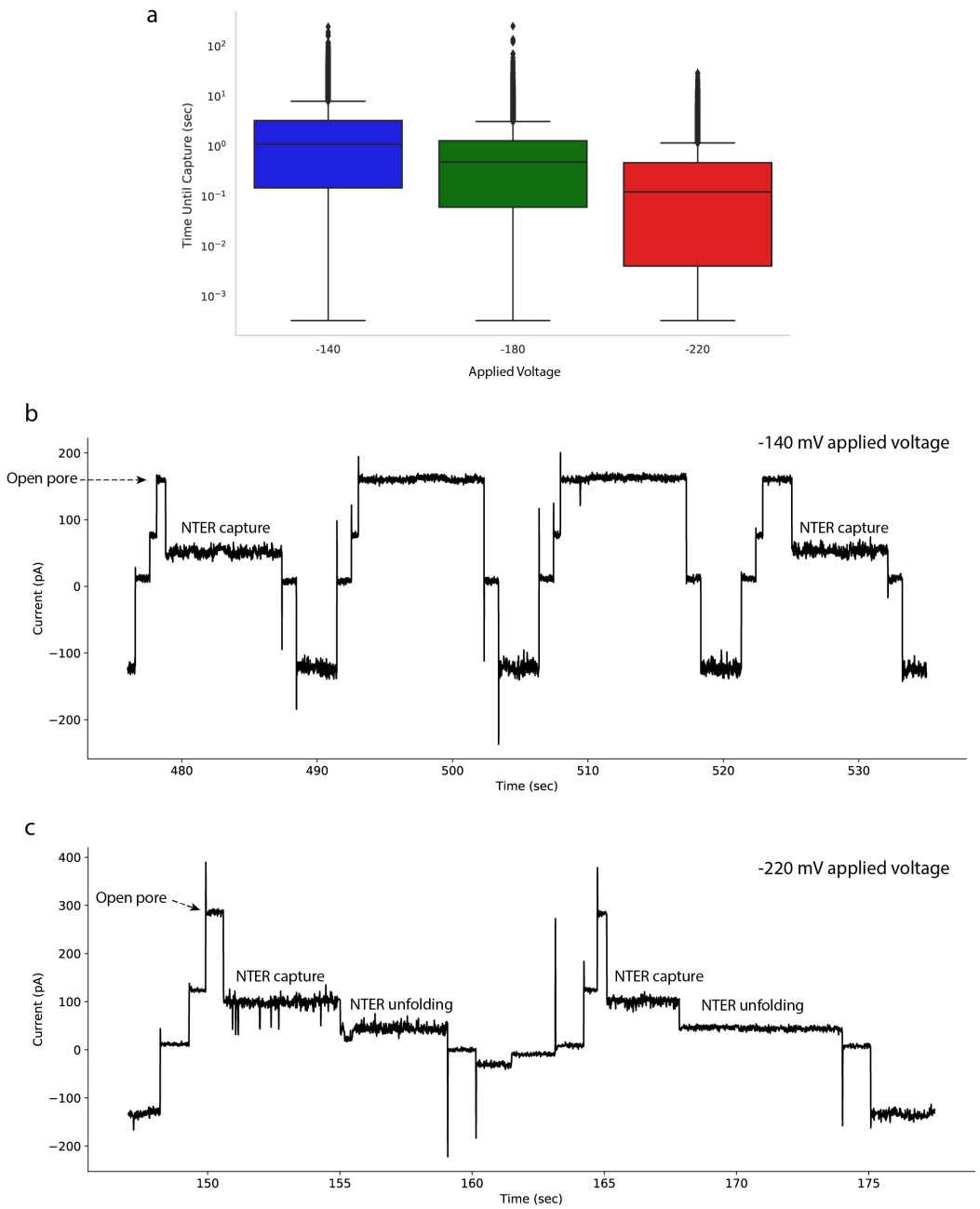


Figure 2.4 Applied voltage on NTER capture The effect of the applied voltage on NTER capture rates. A, The "time between captures" data are presented as a box-and-whisker plot (center line: median, box: 1st and 3rd quartiles, whiskers: min and max, dot: outliers). Results represent the average times collected from experiments conducted with NTERs Y00-08 at 0.5uM. Example NTER captures at **b**, -140 mV, and **c**, -220 mV applied voltage. Increasing capture rates with higher applied voltages could be advantageous for this reporter system in the future, however, we do note that blockade events collected at -220 had a higher tendency to transition into a secondary blockage state following the initial capture state. We attribute this secondary state to unfolding of the Smt3 domain within the pore under the higher applied voltage force. Unfolding is not ideal, as it obscures analysis of the NTER barcode region in the pore.

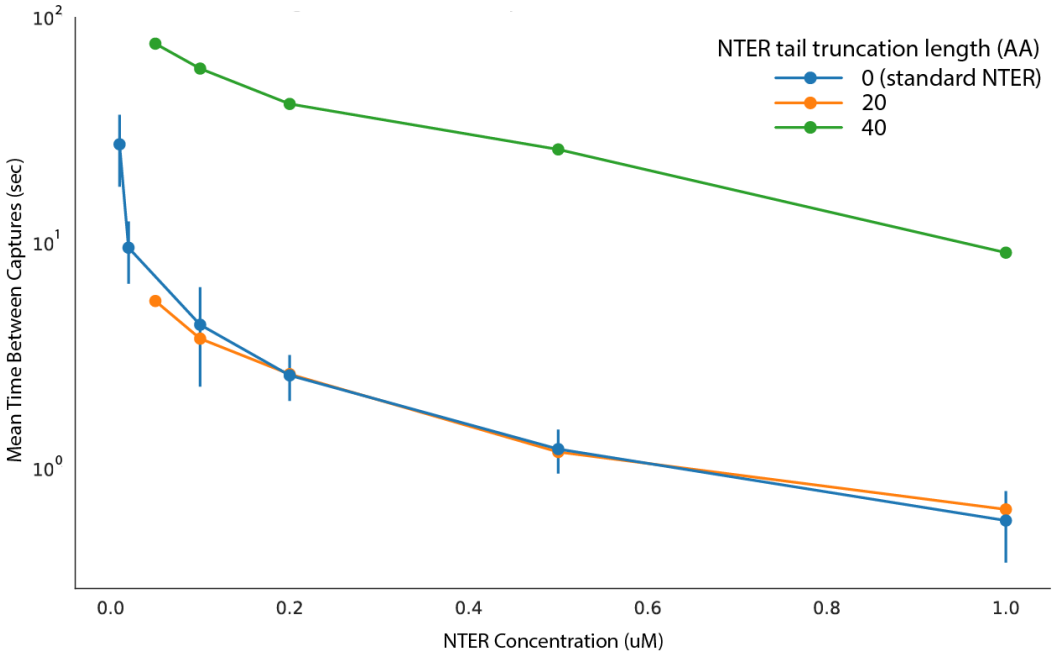


Figure 2.5 NTER tail length on nanopore capture The effect of NTER tail length on nanopore capture rates across different concentrations. Standard NanoporeTER tail sequence, 0 (blue). Tail sequences that have been truncated by either 20 (orange) or 40 (green) amino acids. Error bars represent standard deviation.

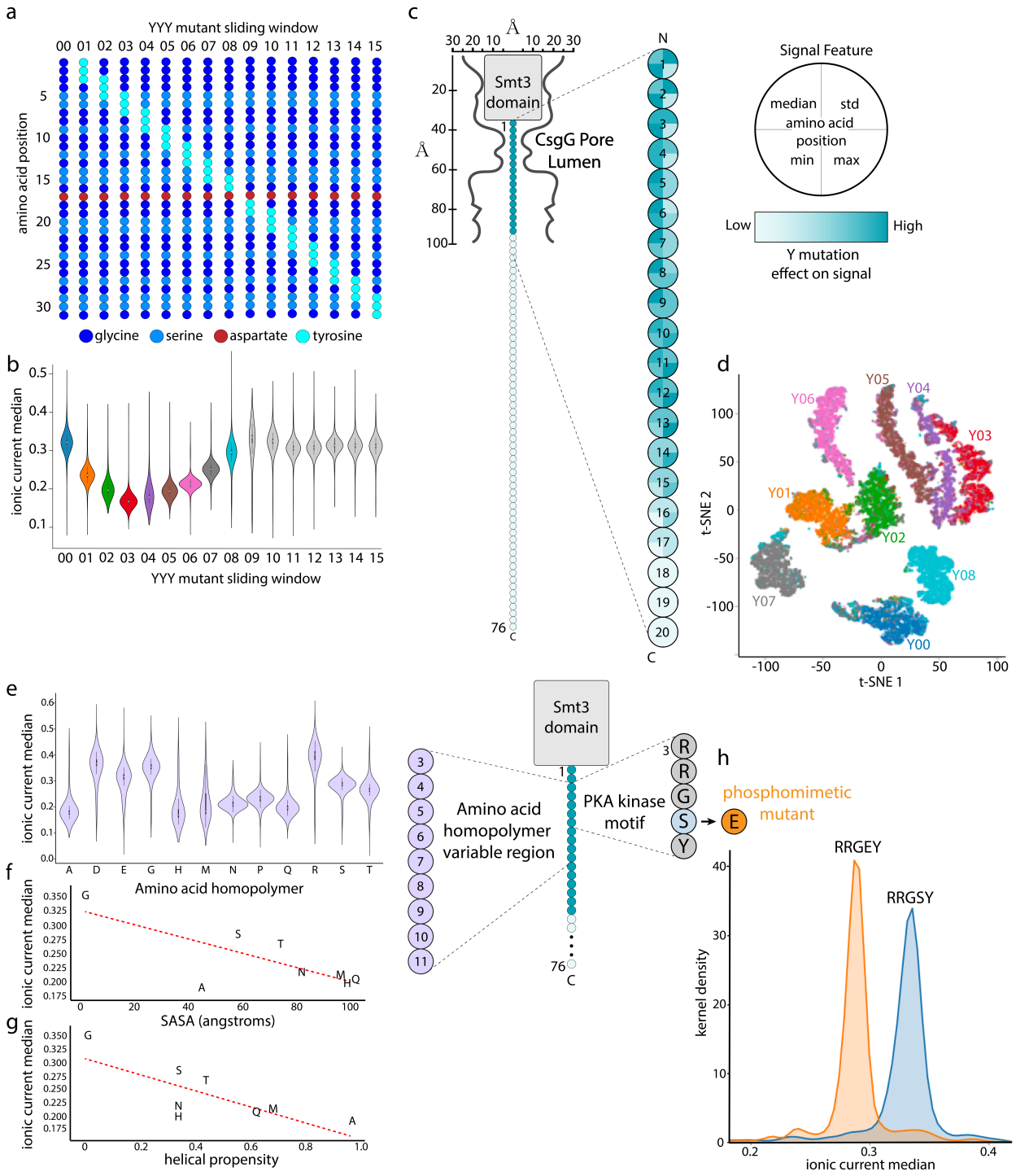


Figure 2.6: Mapping the NanoporeTER sequence and nanopore signal space on a MinION. **a**, Schematic of the NTERs Y00-15 mutant sequences in which a sliding block of three tyrosine mutations was introduced along the NanoporeTER polyGSD barcode and tail region to map the NTER's nanopore-sensitive region and define the potential barcode sequence space. **b**, Violin plot showing the median ionic current level (normalized to the open pore level) of the nanopore capture state for NTERs Y00-15. The introduction of the three tyrosine block (YYY), reduces the ionic current level in a position-dependent manner for positions 01-08. The median current level returns

to the baseline (NTER Y00) level starting at position 9 and through position 15, supporting a model in which the first 17 amino acids of the polyGSD tail contribute to the observed NTER ionic current signature, and defining the NTER barcode region. Each NTER distribution is composed of several thousand single-molecule measurements. **c**, Simple structural model of the NTER position within the nanopore during a read (capture event). A heat map displaying the relative change to specific signal features (median, standard deviation, minimum, and maximum) is projected onto the NTER tail residue positions (1-20) that were mutated in NTERs Y00-15, showing the relative magnitude of effect tyrosine mutations at each residue have on the NTER's nanopore ionic current signal. **d**, t-SNE plot clustering NTER reads (each read is represented as a single point) based on ionic current signal features (mean, std, min, max, median), and colored by the NTER's barcode identity (Y00-08). n = ~4000 events per barcode class. **e**, Violin plot showing the median ionic current level (normalized to the open pore level) of the nanopore capture state for amino acid homopolymer NTERs alanine (A), aspartate (D), glutamate (E), glycine (G), histidine (H), methionine (M), asparagine (N), proline (P), glutamine (Q), arginine (R), serine (S), and threonine (T). Each NTER distribution is composed of ~1500 single-molecule measurements. **f**, Scatter plot showing the relationship between amino acid solvent accessible surface area (SASA) versus the respective amino acid homopolymer NTER mutant's median ionic current level (normalized to the open pore level). **g**, Scatter plot showing the relationship between amino acid helical propensity versus the respective amino acid homopolymer NTER mutant's median ionic current level (normalized to the open pore level). **h**, Kernel density plot comparing the ionic current median (normalized to the open pore level) of reads generated by an NTER containing a PKA phosphorylation motif (RRGSY) within its barcode region to those with a phosphomimetic mutation (RRGEY). Each NTER distribution is composed of several thousand single-molecule measurements.

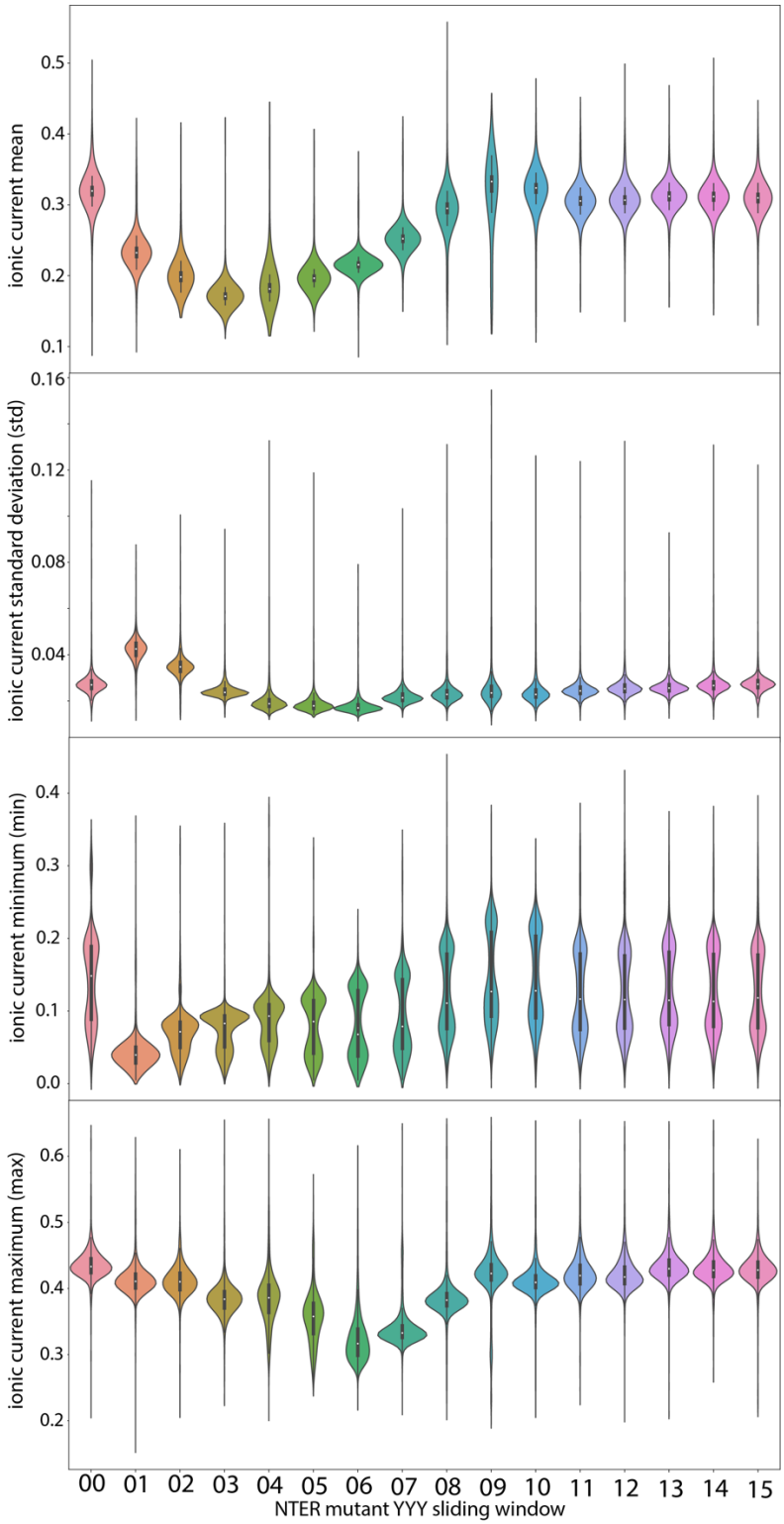


Figure 2.7. Violin plots: showing the ionic current level signal characteristics (mean, std, min, and max. All normalized to the open pore level) of the nanopore capture state for NTERs 00-15. Each NTER distribution is composed of >1000 single-molecule measurements.

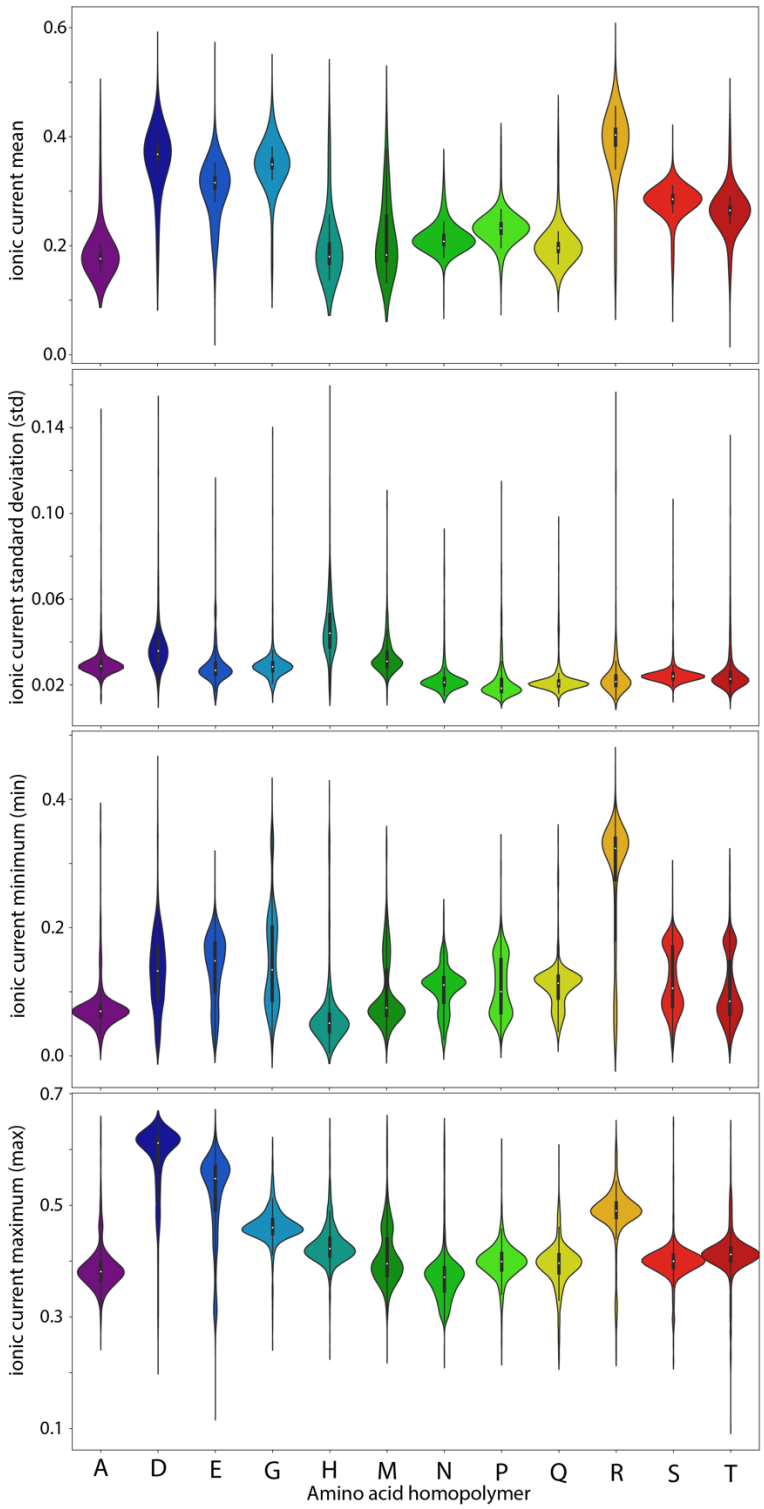


Figure 2.8: Violin plots part two: showing the ionic current level signal characteristics (mean, std, min, and max. All normalized to the open pore level) of the nanopore capture state for the amino acid homopolymer mutants. Each NTER distribution is composed of >1000 single-molecule measurements.

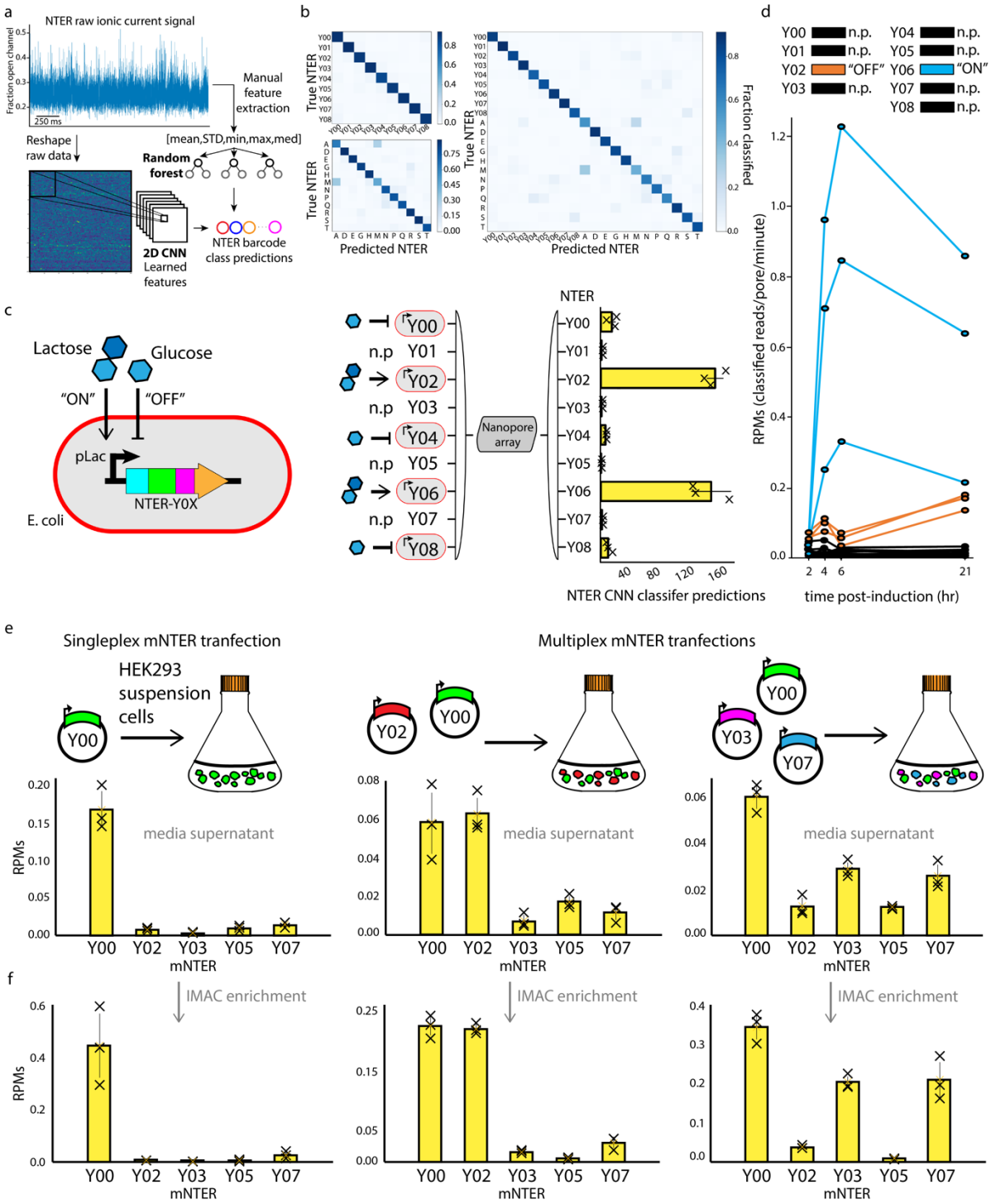


Figure 2.9: Classification and multiplexed detection of NanoporeTER expression levels with a MinION. a, Raw ionic current data was classified using either a set of engineered features (mean, std, min, max, and median) or the unprocessed signal directly, and input into either a Random Forest or Convolutional Neural Network classifier, respectively. **b**, Confusion matrices showing the Random Forest test set classification accuracies on models using different combination of NTER barcodes. Top left: NTERs Y00-08. Bottom left: amino acid homopolymer mutants A, D, E, G, H, M, N, P, Q, R, S, and T. Right: Both the NTERs Y00-08 and amino acid homopolymer

mutants. **c**, Schematic showing the gene construct used for controllable NTER expression. Lactose is used to induce NTER expression ("ON"), while glucose inhibits expression ("OFF"). The diagram and bar plot on the right shows the results of a mixed culture experiments in which NTER expression was induced for NTERs Y02 and Y04, and inhibited for NTERs Y00, Y02, and Y08. NTERs Y01, Y03, Y05, and Y07 were held out of the experiment as negative controls. Plot shows the total number of reads classified as each NTER barcode during MinION analysis. **d**, Line plot showing a time course of NTER expression levels as determined by the rate of classified reads (RPMs: reads/pore/min) for each NTER barcode. NTER Y06 was induced, while NTER Y02 was inhibited. The other NTERs were held out as negative controls and show false-positive classification rates. Three replicates for each condition are plotted. **e**, Bar plots show the results of singleplex and multiplexed HEK293 transfection experiments. For each experiment, a culture of HEK293 suspension cells was transfected with a different barcode combination of vectors containing mNTER proteins (Y0, Y0+Y02, or Y0+Y03+Y07) under the control of a constitutive CMV promoter. Bars show the average rate of classified reads (RPMs) for each barcode during MinION analysis. Three technical replicates for each experiment are plotted. **f**, Same as in **e**, but with the addition of an IMAC purification step prior to MinION analysis.

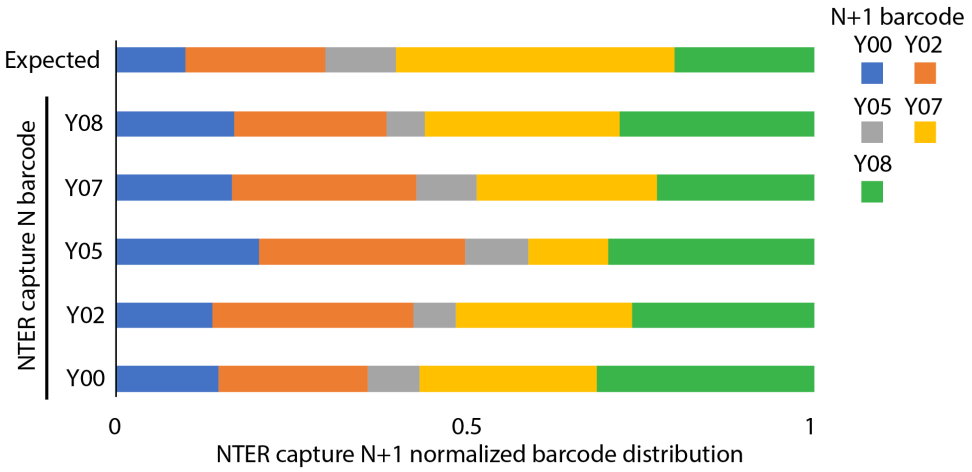


Figure 2.10: Immediate NTER re-capture analysis: The bar plot shows the fraction of each barcode capture in a pore, in comparison to the expected fractions given 5 different NTER barcodes were mixed at varying concentrations. Y00: 0.05uM, Y02: 0.1uM, Y05: 0.05uM, Y07: 0.2uM, and Y08: 0.1uM. For each NTER barcode capture (N), we then determined distribution of the following captured NTER's barcode (N+1).

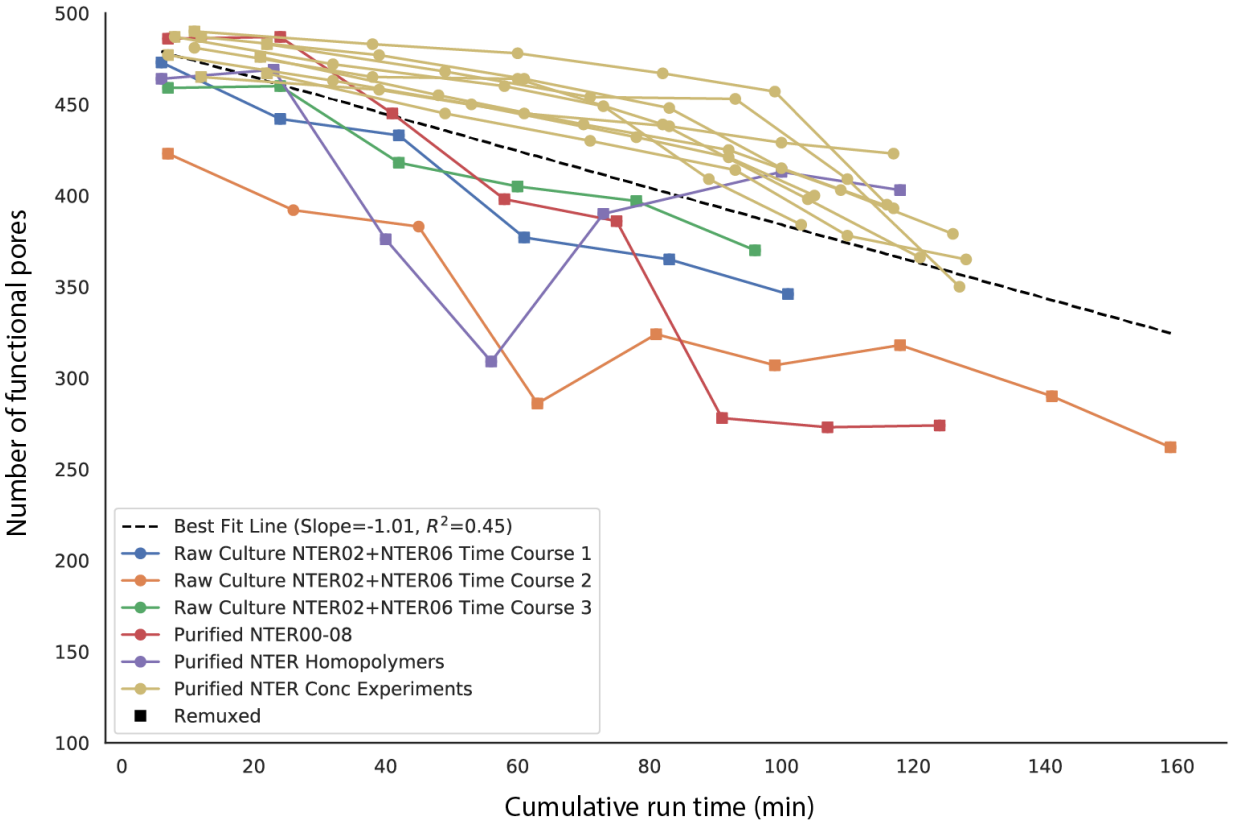


Figure 2.11 MinION flow cell lifetime: The number of nanopores determined to be functional at the start of each experiment vs. cumulative flow cell runtime. Each line is a unique flow cell. Each point represents the start of a new experiment. Square points denote if the flow cell was re-muxed prior to the start of the experiment. Colors denote different sample types and experimental conditions.

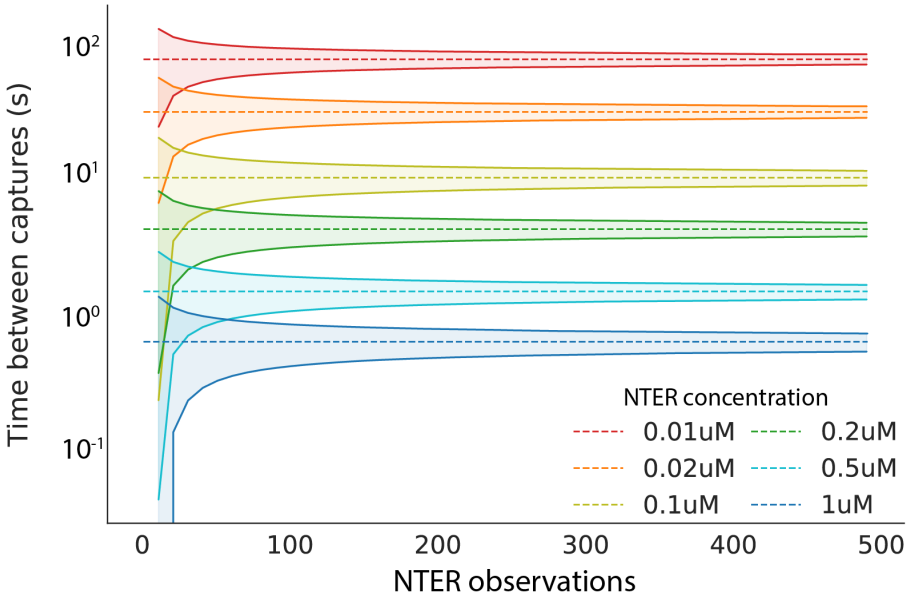


Figure 2.12 95% confidence intervals: the mean time between peptide captures at multiple concentrations, varying the number of observed captures. NTERs Y00-Y08 were run at concentrations of 0.01 μM, 0.02 μM, 0.1 μM, 0.2 μM, 0.5 μM, 1 μM, then classified. We then calculated the time between peptide captures for each run, and pooled these values to combine all runs for the same concentration. The dashed line represents the mean time between captures for the runs pooled by concentration, and the shaded region represents the 95% confidence interval based on the observed mean and standard deviation, for varying numbers of observed NTERs. Note: this plot counts sequential NTER captures, meaning you have already observed at least one capture prior to these calculations.

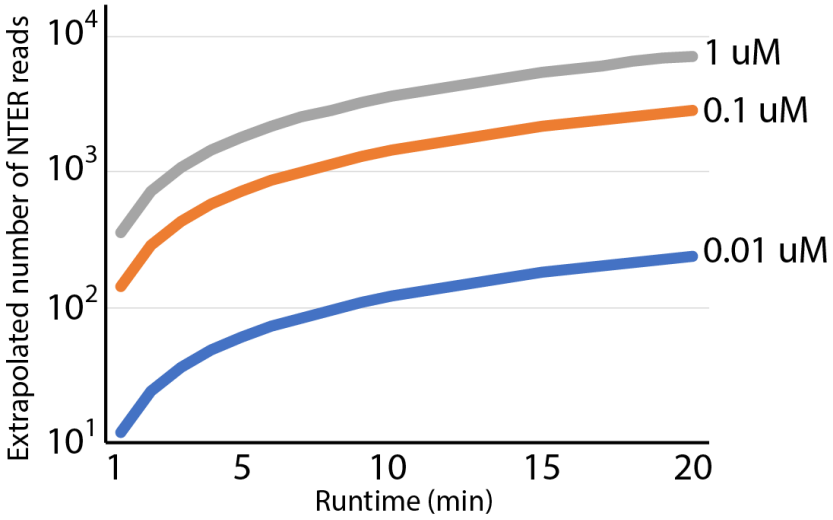


Figure 2.13 Relationship between NTER concentration and MinION run time: various NTER concentrations. These values are extrapolated from concentration experiments conducted on purified NTERY00, which had mean read rates of 361, 144, and 12 reads/min at NTER concentrations of 1 uM, 0.1 uM, and 0.01 uM (respectively), and an average of ~436 functional pores over the course of the experiments.

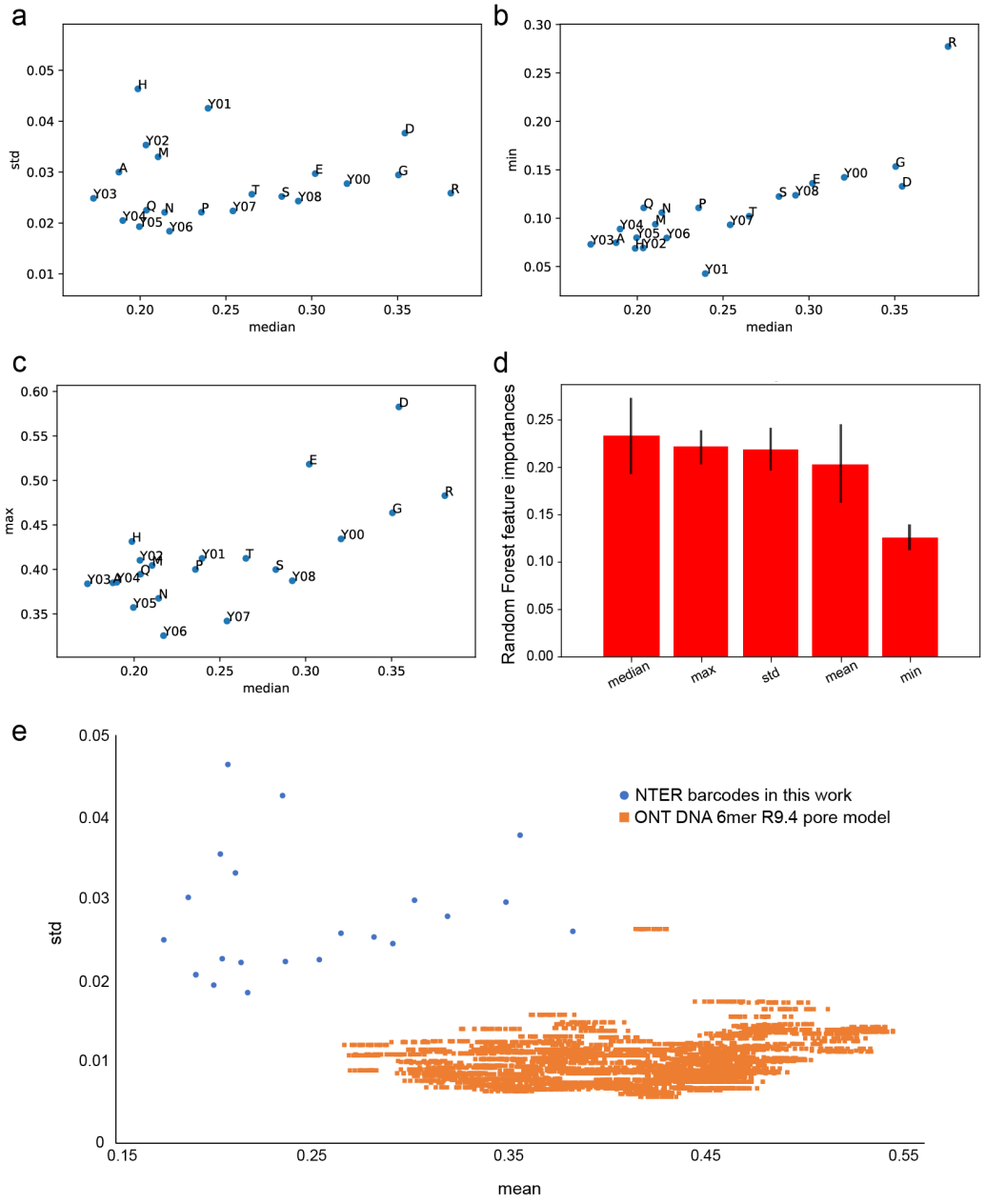


Figure 2.14 NTER barcode signal space: a, Scatter plot showing the mean of the ionic current level signal characteristics of NTER barcodes Y00-08 and homopolymer mutants median vs std, **b**, median vs min, and **c**, median vs max. Each NTER point is the mean of >1000 single-molecule measurements. **d**, NTER Y00-08 and homopolymer mutant barcode classification feature importance determined using a Random Forest model trained on the five extracted signal features (median, mean, std, min, and max). **e**, NTER Y00-08 and homopolymer mutant barcodes mean vs std (blue) compared to Oxford Nanopore Technology's DNA 6mer R9.4 pore model values (level_mean vs level_std) for every possible DNA 6mer (orange). (https://github.com/nanoporetech/kmer_models). The ONT pore model raw current values were converted into fractional current by normalizing them by a typical R9.4 open channel ionic current level of 220 pA.

Chapter 3

Cas9-mediated random access in DNA data storage

3.1 Introduction

Synthetic DNA is being explored as a storage medium for long-term data storage[27-34] because it is an attractive alternative to traditional data storage substrates (eg. magnetic tape) for archival applications. As nature's most robust information storage molecule, DNA features ultrahigh information density (10^{18} bytes per mm^3) and long-term chemical stability, capable of retaining its integrity for thousands to millions of years[35-38]. Recently, end-to-end workflows have been shown, going from encoding, synthesis, retrieval, sequencing, and decoding of data stored in DNA[27,32,34]. To reduce costs and increase performance, scaling up DNA data storage requires methods to selectively retrieve and read pieces of data (random access). Therefore, emphasis has been placed on the retrieval of specific DNA strands in the end-to-end pipeline.

In early DNA data storage architectures, accessing specific files required sequencing the entire DNA library[27,29-31,33]. Large-scale random access is needed to minimize the amount of resources required to access the stored information. Previous work from our group has shown a PCR-based random access approach, in which primer-addressable files are selectively amplified using PCR[32]. This approach requires careful design and validation of primers to avoid crosstalk[72]. Here, we present a CRISPR-Cas9-based random access approach for DNA data storage together with nanopore sequencing. This approach incorporates the sequence-specific targeting capability of the CRISPR-Cas system, which evolved in bacteria as an adaptive immunity mechanism that recognizes and cleaves invading viral nucleic acids[73,74]. We use this feature to target DNA data files and show that our Cas9-based random access architecture can be used to extract files in multiplex from a large DNA pool quickly.

Previous work in genomics has incorporated a Cas9-based targeting system to selectively enrich regions of chromosomal DNA for nanopore sequencing[75]. In this work, a Cas9/gRNA ribonucleoprotein complex (RNP) is used to cleave sites flanking 5' dephosphorylated genomic DNA regions of interest. Nanopore sequencing adapters are then ligated to the cleaved DNA ends that have exposed phosphates. We adapted this technique for DNA data storage using a DNA pool encoding for 25 files. Payload strands were addressed with file specific Cas9 target sites. We then combined this addressing approach with the Rolling Circle Amplification to Concatemeric Consensus (R2C2) method[76]. R2C2 generates long concatemeric repeats of template DNA, which are then nanopore sequenced, and consensus aligned to improve read accuracy.

For this approach, we designed the payload strands (Figure 3.1) to consist of 3 main components: the strand-specific data payload region (125nt), a partially overlapping Cas9 file

address and universal forward primer sites, and a universal reverse primer site. We included universal forward and reverse primers in our design to allow for PCR amplification of the entire ssDNA pool following massively parallel array-based synthesis of the DNA library. Following synthesis of the ssDNA pool, the entire library was amplified and converted to dsDNA via PCR using the universal primer sequences. The PCR products were then circularized through isothermal assembly into a 430 bp splint vector, RCA amplified to generate long concatemer strands (>10kb), and then dephosphorylated according to the R2C2 protocol. Although this approach includes additional preparation steps compared to previous PCR-based random access studies, these steps are conducted "offline" as they can be carried out prior to storage of the DNA library.

3.2 Results

One-file access

Following the offline preparation steps, we conducted an initial test of our random access approach. File 10 was arbitrarily selected to test single-file access (Methods). For this experiment, a fraction of the DNA pool was incubated with File 10-specific-RNP complex at 37 °C for 30 minute, nanopore adaptor ligated, and sequenced using a MinION sensor (Methods). The resulting data was then used to generate single-molecule payload consensus sequences for each read according to the R2C2 (C3POa) analysis pipeline (Methods). To determine the effectiveness of Cas9 enrichment, we then calculated the enrichment score (ES) of each file. The ES is the ratio of reads obtained from each of the 25 files in this experiment relative to the starting file distribution in the original DNA pool as determined by sequencing of the original pool without dephosphorylation or Cas9 enrichment. While these results showed the

File 10 ES to be enriched, however, file 13 had the highest level of enrichment, represented by a higher enrichment score of file 13 (Figure 3.2). We postulated this could be due to Cas9 off-target activity, since there is only a two-nucleotide sequence difference in the file address between file 10 and 13 (Figure 3.3). To prevent over-cutting of the file of interest, we scaled back the incubation time (1 min) of the reaction and again sequenced the library on the nanopore sensor (Figure 3.2). In this run, file 10 was enriched by two orders of magnitude more than the other files (Figure 3.2). We found that file 10 undergoes higher enrichment from the other files in our DNA pool by two orders of magnitude once incubation time is decreased to prevent over-cutting and off-target activity. These results and observations substantiate a model in which the Cas9 endonuclease is targeting and enriching for the file of interest.

Multiplexed experiments

We next tested the multiplexability of the system by accessing three different files in the same reaction vessel (Figure 3.4). We targeted files 2, 13, and 24 for enrichment in the experiment with an incubation time of 15 minutes followed by sequencing. The results, as expected, showed those files were the only files enriched and their enrichment scores were higher by a minimum of one order of magnitude above the other files in the library (Figure 3.4). After determining that our approach could be used to access three files, we next decided to test the limits of our random access architecture by performing enrichment on all even files (12 total) and 20 different files from our DNA pool. On average, the files that were selected for enrichment saw 100-fold higher enrichment scores, as opposed to the other files in the DNA pool (Figure 3.4**)**. When looking at all three multiplexed experiments at once, we found that the accessed files had higher enrichment

scores than the unaccessed files (Figure 3.4), indicating that this random access architecture allows multiplexing of files in the same reaction.

3.3 Discussion:

In conclusion, we presented a novel DNA data storage random access architecture that enables files to be accessed in multiplex via the CRISPR-Cas9 system and decoded using a nanopore sensor array. This approach has several advantages over previous architectures, such as 1) expanded multiplexability, 2) decreased time-to-decoding. The method presented here pushes the limit of files accessed in the same reaction vessel in the same experiment to nearly two dozen files. Although we have shown that 20 files can be accessed using this approach, we believe this number can be increased further by utilizing CRISPR design tools to design optimal guide sequences for more specific file retrieval[77,78].

3.4 Methods

DNA library synthesis and amplification

DNA pool containing the 25 files was synthesized by Twist Bioscience into 1.6 million DNA strands. The 430bp splint sequence was obtained via PCR from pCDB180. The DNA pool (payload) and the splint were amplified with 2x KAPA Master Mix with 10 uM reverse and forward primers (95 °C for 3 minutes; then 11 (payload) and 15 (splint) cycles of: 98 °C for 20 seconds, 62 °C for 15 seconds, and 72 °C for 30 seconds; followed by final 72 °C for 30 seconds). DNA was purified using the QIAGEN QIAquick® PCR Purification Kit protocol.

Assembly and amplification of circularized DNA (alternatively, library preparation)

Gibson assembly and Rolling Circle Amplification (RCA) of the splint and payload were prepared as described previously[22]. In the Gibson Assembly, 200 ng of both splint and payload were combined with 2x NEBuilder HiFi DNA Assembly Master Mix (NEB) and water, and incubated for 60 min for 55C. The Gibson product was digested with 1 uL each of 1:10 Exonuclease III, Lambda Exonuclease, and Exonuclease I, all from NEB. The circularized DNA was extracted using AMPure XP Beads at a ratio of 1.6 beads:1 sample and eluted in 25 uL water.

In the Rolling Circle Amplification, 10 uL aliquots of circularized DNA were amplified in 50 uL reactions with 5 uL of 10x Phi29 buffer (NEB), 2.5 uL of 2.5 mM dNTPs, 2.5 uL of 100 uM random hexamers (Thermo), 1 uL of Phi29 polymerase (NEB), and volume was adjusted with water. Reactions were incubated overnight at 30 C. RCA product was extracted using AMPure XP Beads at a ratio of 0.5 beads:1 sample. DNA was debranched and eluted by adding 10 uL NEB buffer 2, 2.5 uL T7 Endonuclease and 90 uL of water to the beads which were then incubated on a thermal shaker at 37 °C for 1 hr. The supernatant from the beads was collected on magnets, and the DNA in it was extracted again using 0.5 AMPure VP beads:1 sample and eluted in 15 uL water.

Cas9 enrichment

Cas9 ribonucleoprotein complexes (RNPs) were prepared by combining 3 uL of total 10 uM annealed crRNA-tracrRNA (gRNA) (e.g. 0.5 uM each for 20 guides) with 3 uL 10x CutSmart

buffer (NEB) and 0.3 uL 62 uM HiFi Cas9. [For the 1 uM file 10 enrichment experiment, only a total of 1 uM gRNA was used.] Volume was adjusted to 30 uL with nuclease-free water and reaction was incubated at room temperature for 15 minutes, then kept on ice. High molecular weight (~10 kb) RCA product was dephosphorylated by combining 24 uL (about 5 ug) of RCA pool with 3 uL of 10x CutSmart Buffer (NEB) and 3 uL of Quick CP (NEB) and was incubated at 37 °C for 10 minutes, then 80 °C for 2 minutes, then held at room temperature (20 °C). To cleave and dA-tail the RCA sample, the entire dephosphorylated product was gently mixed with 10 uL of the Cas9 RNPs, 1 uL of 10 mM dATP and 1 uL Taq polymerase (NEB). The reactions were incubated at 37 °C for 1 minute (12-file and single-file access), 15 minutes (3-file and 20-file access) or 30 minutes (single-file access), then at 72 °C for 5 minutes, then held at 4 °C or on ice. Adapter mix was prepared in a separate tube by well-mixing 20 uL of Ligation Buffer, 3 uL of nuclease-free water, 10 uL of NEBNext Quick T4 DNA Ligase and 5 uL of AMX adapters. The adapter mix was combined with the cleaved and dA-tailed product for a total volume of 80 uL and incubated at room temperature for 10 minutes. The ligation yield was purified and concentrated using AMPure XP beads at 0.8 beads:1 sample (ligation yield + 80 uL nuclease-free TE buffer) and eluted in 12 uL Elution Buffer.

Nanopore sequencing

Nanopore sequencing was performed on R9.4.1 flow cells from ONT. The flow cells were primed by loading 800 uL from a mix of 1170 uL of Flush Buffer and 30 uL of Flush Tether into the priming port and waiting 5 minutes. The remaining 200 uL of the priming mix was loaded directly before the sequencing sample: 25 uL of Sequencing Buffer, 13 uL of resuspended Loading Beads, and 12 uL of the eluted DNA library. Sample was loaded dropwise on the

SpotON sample port. Sequencing was run at 37 °C for 20-24 hours. When not in use, flow cells were stored in C18 buffer (150 mM potassium ferrocyanide, 150 mM potassium ferricyanide, 25 mM potassium phosphate, pH 8.0) at 4 °C.

Basecalling and data analysis

Basecalling on sequencing reads was performed using Guppy v3.2.2 (available from ONT) with a quality score cutoff of 9. Reads were then processed using C3POa, which demultiplexes reads into respective files based on file address and generates a consensus sequence for each concatemeric read. Following C3POa, the splint sequences, primer sequences, and file addresses were trimmed off of each read, leaving the payload. Payload sequences were then decoded to recover the original digital files stored in DNA.

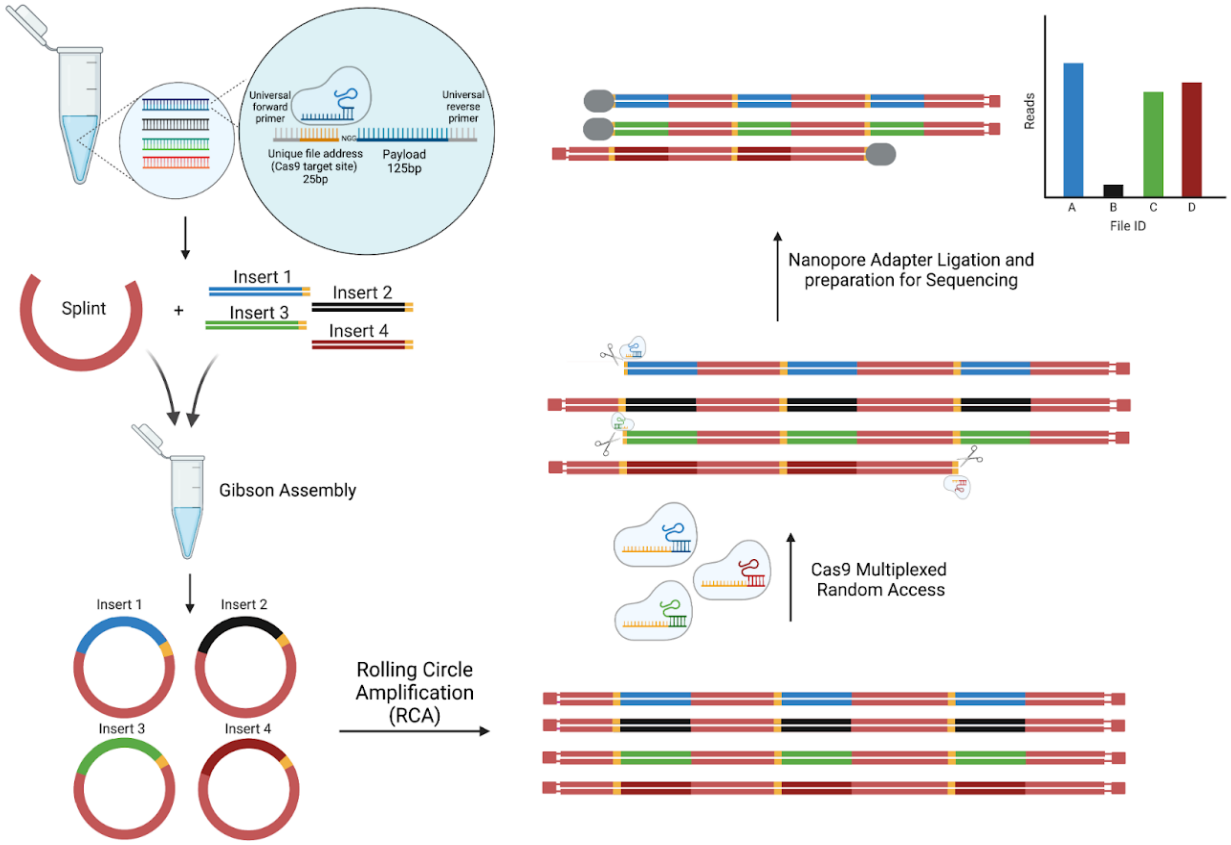


Figure 3.1: Schematic of random access pipeline. Each file (payload) has a file address that contains a unique Cas9 target site including a PAM. On either end of the payload are universal primers which hybridize to the ends of a universal splint during Gibson assembly. Circularized files are then amplified via RCA which results in linear high molecular weight strands. Specific files from a set of 25 are targeted with their complementary Cas9-RNPs which cut the long DNA strands at file address sites. Adapters for nanopore sequencing are ligated at the cut sites, thus only these files are enriched.

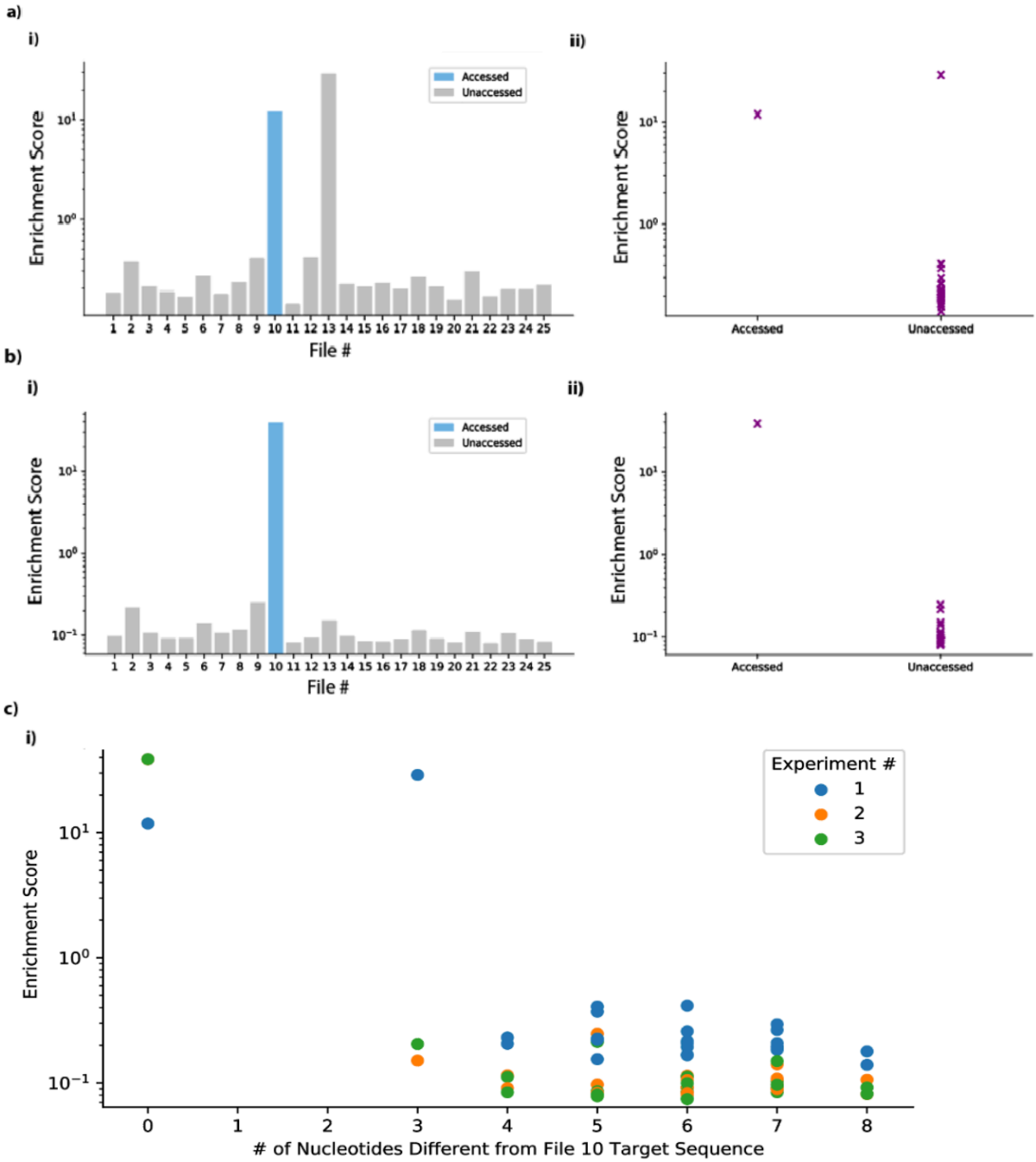


Figure 3.2: **Single-file enrichment of file 10**. a) Initial experiment accessing file 10. 10 uM gRNA was incubated in the Cas9-cleavage reaction for 30 minutes prior to sequencing. File 10 was enriched along with unaccessed file 13 over one order of magnitude higher than the other 23 files. b) To reduce possible Cas9 overactivity, gRNA concentration was decreased to 1 uM and incubated for 1 minute. File 10 was enriched two orders of magnitude over the other files, and no unaccessed files were enriched. c) Comparison of the ES of files with similar Cas9 target sequences to the accessed file as a function of the number of nucleotide differences. As the number of nucleotide differences increases

Figure 3.3: Cas9 gRNA sequences for file 10 AND 13 in library. The different nucleotides between files 10 and 13 are colored orange and the PAM sequences are colored in blue.

> Cas9_guide_10
CTCGCAGAGGTGGCGCATTAAATGG
> Cas9_guide_13
CTCGCAGAGGTGGCGCTTTACAAGG

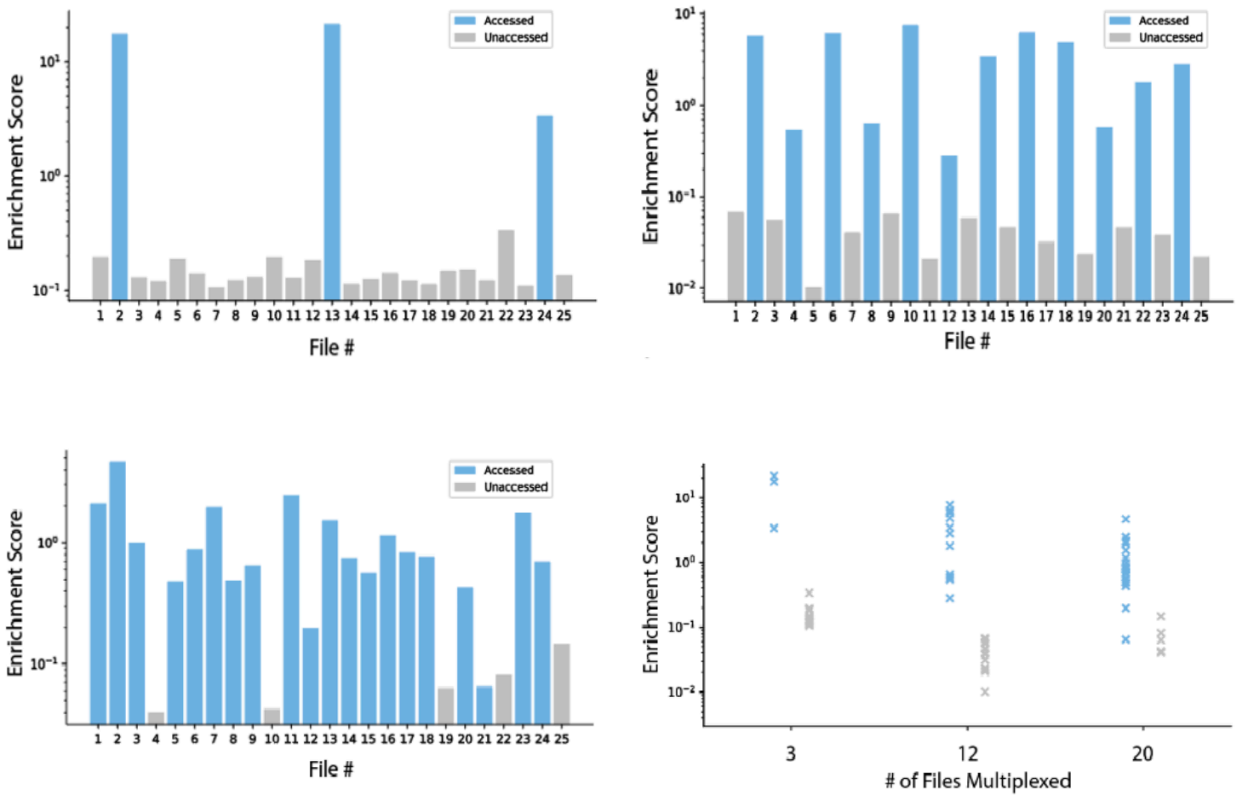


Figure 3.4: Multiplexed enrichment experiments. a) First, multiplexing was tested for three files: 2, 13 and 24 (3.33 uM gRNAs and 15 minute reaction incubation). These three accessed files were enriched at least one order of magnitude more than the unaccessed files. b) Enrichment of all the even files, (0.83 uM gRNAs and 1 minute reaction incubation). All files were successfully enriched, though with a smaller difference in enrichment between accessed and unaccessed files. c) Enrichment of 20 files (0.5 uM gRNAs and 15 minute reaction incubation). In this experiment, there was a single false negative, file 21, but all other 19 files were successfully enriched. Generally, the maximum enrichment decreases for larger multiplexed sets.

Chapter 4

Towards single-molecule protein sequencing on a high-throughput nanopore sensor array

Portions of this chapter were previously published as a manuscript in iScience [79].

3.1 Introduction

Proteins are the major functional molecules involved in essentially every biological process, such as regulating gene expression and powering the immune system. Although genomics and transcriptomics provide fundamental information about cellular history and basal activity, proteomics plays a crucial role in filling the gap between genotype and phenotype as protein activities are more directly related to phenotype. Thus, protein analysis provides valuable information for understanding biological phenomena and disease. Unfortunately, unlike the remarkable technological improvements in DNA and RNA sequencing in recent years, the development of highly sensitive, high-throughput protein sequencing techniques have not yet

been realized. There are two principal methods currently available for protein sequencing/identification that do not use affinity reagents such as antibodies: Edman degradation and mass spectrometry[80,81]. Edman degradation is a useful technique for de novo sequencing, but it is limited to the analysis of homogenous protein samples and read lengths typically <50 amino acids, which are far shorter than the median protein length of eukaryotic (361 amino acid long), bacterial (267 amino acid long), and archaeal organisms (247 amino acid long)[82]. Mass spectrometry allows the analysis of protein mixtures and currently dominates proteomics research. Mass spectrometry has undergone significant improvements in instrumentation and sample preparation over the decades, although it still faces limitations in terms of detection sensitivity, dynamic range, analytical throughput, and instrumentation cost[83].

Although nanopore sensing was initially proposed as a technique for the sequencing of nucleic acid strands[84], it also has great potential for protein analysis. Single-molecule sensitivity, full-length readout, real-time measurement, and device portability is just as, if not more, crucial for proteomics than it is for genomics and transcriptomics[85]. Nanopore sensors have been used for discrimination of peptides and proteins[39,49,86-89], real-time measurement of protein–protein[90] and protein–ligand interactions[91]. Moreover, protein nanopores have shown promise in identifying amino acids and post-translational modifications (PTMs), taking a major step toward single-molecule protein sequencing.

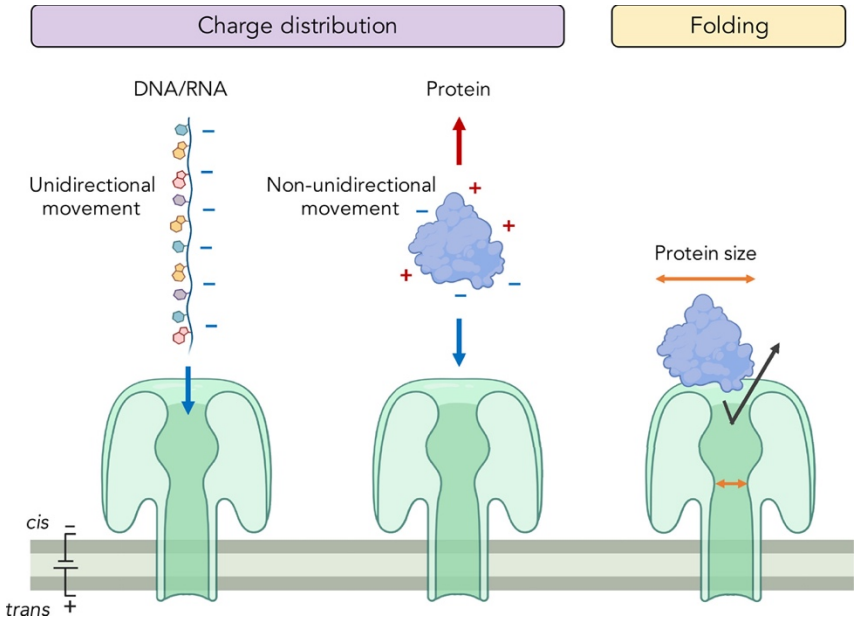


Figure 4.1 Challenges in protein translocation through a nanopore
(Left) While DNA/RNA is uniformly negatively charged, (Middle) proteins can contain both negatively-charged (glutamate and aspartate) and positively-charged (arginine, lysine, and histidine) residues at physiological pH. Unidirectional translocation of proteins in the electric field must be achieved despite their nonuniform charge. (Right) The diameters of folded proteins are typically larger than the constriction of protein nanopores that would be suitable for protein sequencing application. For processive strand analysis, proteins must be unfolded to allow the denatured protein strand to thread through the nanopore with amino acid residues in single-file order.

Despite these promising results, protein sequencing of intact, full-length protein strands using nanopores has been hindered, in part, because of the difficulty in controlling protein translocation through the sensor. This challenge exists because of two major reasons (Figure 4.1): First, the polypeptide backbone is neutrally charged and amino acid side chains can vary in charge state. Thus, electrophoresis-driven unidirectional translocation of peptides or proteins through nanopores cannot be as effectively employed, as it can for uniformly negatively-charged polymers like nucleic acids. Second, most proteins adopt a stable 3-dimensional fold. Thus, disruption of this tertiary structure is required for proteins to translocate through a narrow nanopore constriction for primary sequence analysis. Here, I speak about the

advances and obstacles in controlling protein translocation through a nanopore and highlight label-based approaches that potentially address these challenges.

4.2 Discussion

In 2004, peptide translocation through a nanopore was demonstrated for the first time[92]. This study showed the translocation of short repeats of the collagen-like sequence (GPP) through an α-hemolysin nanopore. Another pioneering study investigated the interactions of an α-hemolysin nanopore with helical peptides containing the $(AAKAA)_n$ sequence[93]. These works laid the foundation for peptide analysis using nanopores, but general approaches for translocation of native peptides/proteins through nanopores are required for the development of single-molecule protein sequencing.

Physical and chemical denaturants

First efforts for protein translocation involved physical and chemical denaturants. For example, groups have demonstrated protein translocation through solid-state pores using sodium dodecyl sulfate (SDS) as a denaturant[94]. SDS further provides a near-uniform negative charge to denatured proteins and promotes the electrical control of the translocation kinetics, though it is unclear if the protein-bound SDS could interfere with the nanopore signal's sensitivity to amino acid sequence. Although such denaturation methods are compatible with solid-state nanopores and show great promise for protein translocation, they cannot be as readily applied to protein nanopore systems that include lipid or lipid-like membranes, which are susceptible to harsh conditions required to completely unfold stable proteins (e.g. high temperature or a high

concentration of denaturants). To overcome this barrier, several label-based translocation strategies that are compatible with protein nanopores have been explored.

In nanopore-based DNA and RNA sensing, nucleic acid strands can be electrophoretically-driven into and through a nanopore unidirectionally by an applied voltage as their phosphodiester backbone is intrinsically negatively charged. Hence, attaching an oligonucleotide strand to a protein is a straightforward way to facilitate electrophoresis-driven protein translocation. This method, however, generates fast translocation events (<1 ms) that may cause poor signal-to-noise ratios and thus make this method less sensitive to protein sequence-level changes. This method may also be ineffective at translocating larger, multi-domain proteins, as the electrophoretic pulling force is largely absent after the oligonucleotide has completely translocated through the nanopore. As molecular motors and nanopore engineering have been employed to reduce the velocity of translocating DNA and RNA in nanopore nucleic acid sequencing[95-97], techniques to regulate the rate of protein translocation may be required for the acquisition of well-resolved and reproducible current signals.

Another approach for protein translocation is based on an unfoldase that enables enzyme-mediated unfolding and translocation of tagged proteins. We previously employed the AAA+ unfoldase ClpX, which specifically unfolds proteins bearing a C-terminal ssrA peptide tag (AANDENYALAA), for processive unfolding of large proteins[48,49]. ClpX generates sufficient mechanical force (~20 pN) to denature stable protein folds and translocates proteins at a rate suitable for nanopore sequencing (up to 80 amino acids per second)[98]. This approach has demonstrated ClpX-mediated translocation of proteins over 700 amino acids in length, including a variety of protein domains, that are genetically fused with the ssrA tag and a polyanion peptide linker designed to promote protein capture and retention in the nanopore electric field. Distinct

protein domains as well as specific point mutations, proteolytic cleavage, and sequence rearrangements in those domains resulted in detectable ionic current pattern changes and single-molecule classification accuracies of 86–99%[49].

While ClpX is capable of unfolding many different types of proteins even with very high stabilities[99], it likely does not generate sufficient force for some protein folds. The force exerted on protein strands by ClpX or by an electric field (tens to several hundred pN) in a typical experimental setup is not able to break the covalent disulfide bond[100]. Although ClpX's ring-like structure is flexible enough to translocate a disulfide-linked beta hairpin into the proteolytic chamber of ClpP[101], it is unlikely that a more narrow, rigid nanopore protein would accommodate such a structure. Thus, the use of reducing agents would assist the linear translocation of proteins with disulfide bonds. Another consideration to the unfoldase approach is the large and variable translocation step size of ClpX. Although the fundamental step size of ClpX is ∼1 nm, this distance corresponds to an irregular number of amino acids that is dependent on the confirmation of the peptide backbone (typically 5–8 amino acids per 1 nm step). ClpX stepping can also occur in quick bursts of up to 4 nm[102]. Although this bursting activity is critical to its unfolding activity, it could complicate sequencing with single-amino-acid resolution. Efforts to explore alternative unfoldase motors that have a more well-defined step size, such as ClpA[103], or proteasome systems[104] may be necessary for building a more robust translocation system with optimal resolution.

To summarize, key advances have been made toward facilitated translocation of proteins and peptides through narrow protein nanopore sensors for the realization of single-molecule protein sequencing. The development of protein translocation systems has been an exciting and active research area and further improvements are anticipated shortly.

4.3 Conclusion

The research described in this dissertation outlines two methods to perform multiplexing for molecular assays. I have described a novel barcoded protein reporter system that incorporated nanopore sensors and utilizes them outside of their intended use case. Additionally, I have shown a DNA data storage random access architecture that is potentially more multiplexable and faster than previous methods. The barcoded protein reporters work presented here shows the potential of using a commercially available nanopore sensor array to develop methods for single-molecule protein sequencing. Post-translational modifications, point-mutations, as well as 12 of the canonical amino acids were identified in this pipeline. Techniques that incorporate a more precise motor protein, improved tagging methods, and more accurate algorithms will still have to be developed but will prove fruitful to accomplish this challenging problem.

1. Posthuma-Trumpie G., Korf J., van Amerongen A. Lateral flow (immuno)assay: Its strengths, weaknesses, opportunities and threats. A literature survey. *Anal. Bioanal. Chem.* 2009

2. Schubert-Ullrich P., Rudolf J., Ansari P., Galler B., Führer M., Molinelli A., Baumgartner S. Commercialized rapid immunoanalytical tests for determination of allergenic food proteins: An overview. *Anal. Bioanal. Chem.* 2009

3. Ngom B., Guo Y., Wang X., Bi D. Development and application of lateral flow test strip technology for detection of infectious agents and chemical contaminants: A review. *Anal. Bioanal. Chem.* 2010

4. Hoffmann, B., Beer, M., Reid, S. M., Mertens, P., Oura, C. A., van Rijn, P. A., et al. (2009). A review of RT-PCR technologies used in veterinary virology and disease control: sensitive and specific diagnosis of five livestock diseases notifiable to the World Organisation for Animal Health. *Vet. Microbiol.*

5. Holland, P. M., Abramson, R. D., Watson, R., and Gelfand, D. H. (1991). Detection of specific polymerase chain reaction product by utilizing the 5′—3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc. Natl. Acad. Sci. U.S.A.*

6. Johnson, G., Nolan, T., and Bustin, S. A. (2013). Real-time quantitative PCR, pathogen detection and MIQE. *Methods Mol. Biol.*

7. A Norris, A. L., Workman, R. E., Fan, Y., Eshleman, J. R., & Timp, W. (2016). Nanopore sequencing detects structural variants in cancer. *Cancer biology & therapy*.

8. Cumbo, C. *et al*. (2018). Genomic BCR-ABL1 breakpoint characterization by a multi-strategy approach for "personalized monitoring" of residual disease in chronic myeloid leukemia patients. *Oncotarget.*

9. Jeck, W. R., Lee, J., Robinson, H., Le, L. P., Iafrate, A. J., & Nardi, V. (2019). A nanopore sequencing–based assay for rapid detection of gene fusions. *The Journal of Molecular Diagnostics.*

10. Deamer, D.W., and Akeson, M. (2000). Nanopores and nucleic acids: prospects for ultrarapid sequencing. Trends Biotechnol. 18

11. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.* (2016).

12. Wanunu, M. Nanopores: A journey towards DNA Sequencing. *Phys. Life Rev.* **9**, 125-158 (2012).

13. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* (2018).

14. Soskine, M. et al. An engineered ClyA nanopore detects folded target proteins by selective external association and pore entry. *Nano Lett.* **12** (2012).

15. Stefureac, R., Waldner, L., Howard, P. & Lee, J. S. Nanopore analysis of a small 86-residue protein. *Small* **4**, (2008).

16. Si, W. & Aksimentiev, A. Nanopore sensing of protein folding. *ACS Nano* **11**, (2017).

17. Robertson, J. et al. The Utility ofNanopore Technology for Protein Sensing. Proteomics (2018).

18. Nivala, J. et al. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. Nat. Biotechnol. 31, (2013).

19. Nivala, J. et al. Discrimination among protein variants using an unfoldase-coupled nanopore. ACS Nano 8, (2014).

20. Rosen, C. et al. Single-molecule site-specific detection of protein phosphorylation with a nanopore. Nat. Biotechnol. (2014).

21. Wood, K. V. (1995) Marker proteins for gene expression. Curr. Opin. Biotechnol. 6, 50-58.

22. Morin, J. G. and Hastings, J. W. (1971) Biochemistry of the bioluminescence of colonial hydroids and other coelenterates. J. Cell Physiol.

23. Morin, J. G. and Hastings, J. W. (1971) Energy transfer in a bioluminescent system. J. Cell Physiol.

24. Rodriguez, E. A. *et al.* The Growing and Glowing Toolbox of Fluorescent and Photoactive Proteins. *Trends in Biochemical Sciences* (2017). doi:10.1016/j.tibs.2016.09.010

25. Martin, L., Che, A. & Endy, D. Gemini, a bifunctional enzymatic and fluorescent reporter of gene expression. *PLoS One* (2009). doi:10.1371/journal.pone.0007569

26. Parrello, D., Mustin, C., Brie, D., Miron, S. & Billard, P. Multicolor whole-cell bacterial sensing using a synchronous fluorescence spectroscopy-based approach. *PLoS One* (2015). doi:10.1371/journal.pone.0122848

27.. Zhirnov, V., Zadegan, R. M., Sandhu, G. S. & Church, G. M. Nucleic acid memory. *Nat. Mater.* 15, 366–370 (2016)

28. Yazdi, S., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A rewritable, random-access DNA-based storage system. *Sci. Rep.* 5, 14138 (2015)

29. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* 494, 77 (2013)

30. Bornholt, J. et al. A DNA-based archival storage system. *ACM SIGARCH Comput. Archit. News* 44, 637–649 (2016)

31. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* 337, 1628 (2012)

32. Organick, L. et al. Random access in large-scale DNA data storage. *Nat. Biotechnol.* 36, 242–248 (2018)

33. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* 355, 950–954 (2017)

34. Yazdi, H. S. M., Gabrys, R. & Milenkovic, O. Portable and error-free DNA-based data storage. *Sci. Rep*. 7, 5011 (2017)

35. Grass, R., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chem.* 54, 2552–2555 (2015).

36. Rutten, M. G. T. A., Vaandrager, F. W., Elemans, J. A. A. W. & Nolte, R. J. M. Encoding information into polymers. *Nat. Rev. Chem.* 2, 365–381 (2018).

37. Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nat. Mater.* 15, 366–370 (2016). This paper presents a detailed analysis of properties of DNA as a data storage medium and compares it with other media.

38. Organick, L., Nguyen, B. H., McAmis, R., Chen, W. D., Kohll, A. X., Dumas, S., Grass, R. N., Ceze, L., Strauss, K., An Empirical Comparison of Preservation Methods for Synthetic DNA Data Storage. *Small Methods* 2021.

39. Cardozo, N., Zhang, K., Doroschak, K. *et al.* Multiplexed direct detection of barcoded protein reporters on a nanopore array. *Nat Biotechnol* **40,** 42–46 (2022).

40. Ghim, C. M., Lee, S. K., Takayama, S. & Mitchell, R. J. The art of reporter proteins in science: Past, present and future applications. *BMB Reports* (2010).

41. Shimo, T., Tachibana, K. & Obika, S. Construction of a tri-chromatic reporter cell line for the rapid and simple screening of splice-switching oligonucleotides targeting DMD exon 51 using high content screening. *PLoS One* (2018)

42. Wroblewska, A. *et al.* Protein Barcodes Enable High-Dimensional Single-Cell CRISPR Screens. *Cell* **175**, (2018). doi: 10.1016/j.cell.2018.09.022

43. He, W., Yuan, S., Zhong, W. H., Siddikee, M. A. & Dai, C. C. Application of genetically engineered microbial whole-cell biosensors for combined chemosensing. *Applied Microbiology and Biotechnology* (2016). doi:10.1007/s00253-015-7160-6

44. Nielsen, A. A. K. *et al.* Genetic circuit design automation. *Science.* **352**, aac7341–aac7341 (2016). doi: 10.1126/science.aac7341

45. Shi, W., Friedman, A. K. & Baker, L. A. Nanopore Sensing. *Anal. Chem.* **89**, 157–188 (2017). doi: 10.1021/acs.analchem.6b04260

46. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.* (2016). doi:10.1186/s13059-016-1103-0l

47. Garalde, D. R. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* (2018). doi:10.1038/nmeth.4577

48. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* (2013). doi:10.1038/nbt.2503

49. Nivala, J., Mulroney, L., Li, G., Schreiber, J. & Akeson, M. Discrimination among protein variants using an unfoldase-coupled nanopore. *ACS Nano* **8**, 12365–12375 (2014). doi: 10.1021/nn5049987

50. Baker, T. A. & Sauer, R. T. ClpXP, an ATP-powered unfolding and protein-degradation machine. *Biochim. Biophys. Acta - Mol. Cell Res.* **1823**, 15–28 (2012). doi: 10.1016/j.bbamcr.2011.06.007.

51. Yim, H. H. & Villarejo, M. osmY, a new hyperosmotically inducible gene, encodes a periplasmic protein in Escherichia coli. *J. Bacteriol.* (1992). doi:10.1128/jb.174.11.3637-3644.1992

52. Kotzsch, A. *et al.* A secretory system for bacterial production of high-profile protein targets. *Protein Sci.* (2011). doi:10.1002/pro.593

53. Goyal, P. *et al.* Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature* (2014). doi:10.1038/nature13768

54. Taylor, S. S. *et al.* PKA: A portrait of protein kinase dynamics. in *Biochimica et Biophysica Acta - Proteins and Proteomics* (2004). doi:10.1016/j.bbapap.2003.11.029

55. Román, R. *et al.* Enhancing heterologous protein expression and secretion in HEK293 cells by means of combination of CMV promoter and IFNα2 signal peptide. *J. Biotechnol.* **239**, 57–60 (2016).

56. Peroutka, R. J., Elshourbagy, N., Piech, T. & Butt, T. R. Enhanced protein expression in mammalian cells using engineered SUMO fusions: Secreted phospholipase A 2 . Protein Sci. (2008). doi:10.1110/ps.035576.108

57. Li, G. W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* (2014). doi:10.1016/j.cell.2014.02.033

58. Owens, N. D. L. *et al.* Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. *Cell Rep.* (2016). doi:10.1016/j.celrep.2015.12.050

59. Gorochowski, T. E. *et al.* Absolute quantification of translational regulation and burden using combined sequencing approaches. *Mol. Syst. Biol.* (2019). doi:10.15252/msb.20188719

60. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* (2016). doi:10.1038/nmeth.3930

61. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* (2019). doi:10.1186/s13059-019-1727-y

62. Rodriguez-Larrea, D. & Bayley, H. Multistep protein unfolding during nanopore translocation. *Nat. Nanotechnol.* (2013). doi:10.1038/nnano.2013.22

63. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to Sequence Learning with Neural Networks. *Proc. NIPS*. 1–9 (2014).

64. Olczak, M. & Olczak, T. Comparison of different signal peptides for protein secretion in nonlytic insect cell system. *Anal. Biochem.* (2006). doi:10.1016/j.ab.2006.09.003

65. Bitter, G. A., Chen, K. K., Banks, A. R. & Lai, P. H. Secretion of foreign proteins from Saccharomyces cerevisiae directed by alpha-factor gene fusions. *Proc. Natl. Acad. Sci.* (2006). doi:10.1073/pnas.81.17.5330

66. Attallah, C., Etcheverrigaray, M., Kratje, R. & Oggero, M. A highly efficient modified human serum albumin signal peptide to secrete proteins in cells derived from different mammalian species. *Protein Expr. Purif.* (2017). doi:10.1016/j.pep.2017.01.003

67. Gorochowski, T. E. *et al.* Genetic circuit characterization and debugging using RNA-seq. *Mol. Syst. Biol.* (2017). doi:10.15252/msb.20167461

68. Gach, P. C. *et al.* A Droplet Microfluidic Platform for Automating Genetic Engineering. *ACS Synth. Biol.* **5**, 426–433 (2016). doi: 10.1021/acssynbio.6b00011

69. Chao, R., Mishra, S., Si, T. & Zhao, H. Engineering biological systems using automated biofoundries. *Metab. Eng.* **42**, 98–108 (2017). doi: 10.1016/j.ymben.2017.06.003

70. Madison, A. C. *et al.* Scalable Device for Automated Microbial Electroporation in a Digital Micro fluidic Platform. *ACS Synth. Biol.* 1701-1709. (2017). doi:10.1021/acssynbio.7b00007

71. Gao, X. J., Chong, L. S. & Kim, M. S. Programmable protein circuits in living cells. *Science*. **361**, 1252–1258 (2018). doi: 10.1126/science.aat5062

72. Schmid-Burgk JL, Gao L, Li D, Gardner Z, Strecker J, Lash B, Zhang F. Highly Parallel Profiling of Cas9 Variant Specificity. Mol Cell. 2020

73. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–1712 (2007).

74. Pickar-Oliver, A., Gersbach, C.A. The next generation of CRISPR–Cas technologies and applications. *Nat Rev Mol Cell Biol* 20, 490–507 (2019).

75. Gilpatrick, T., Lee, I., Graham, J.E. *et al.* Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol* 38, 433–438 (2020).

76. Volden, R., Palmer, T., Byrne, A., Cole, C., Schmitz, R. J., Green, R. E., & Vollmers, C. (2018). Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. *Proceedings of the National Academy of Sciences*, *115*(39), 9726-9731.

77. Schmid-Burgk JL, Gao L, Li D, Gardner Z, Strecker J, Lash B, Zhang F. Highly Parallel Profiling of Cas9 Variant Specificity. Mol Cell. 2020

78. Xu, Q., Schlabach, M. R., Hannon, G. J. & Elledge, S. J. Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc. Natl Acad. Sci. USA* 106, 2289–2294 (2009).

79. Motone, K., Cardozo, N., & Nivala, J. (2021). Herding cats: Label-based approaches in protein translocation through nanopore sensors for single-molecule protein sequence analysis. *Iscience*

80. Edman, P., Högfeldt, E., Sillén, L. G., & Kinell, P. O. (1950). Method for determination of the amino acid sequence in peptides. *Acta chem. scand*, *4*(7), 283-293.

81. Steen, H., & Mann, M. (2004). The ABC's (and XYZ's) of peptide sequencing. *Nature reviews Molecular cell biology*

82. Brocchieri, L., & Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. *Nucleic acids research*

83. Zubarev, R. A. (2013). The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics*

84. Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*

85. Shi, W., Friedman, A. K., & Baker, L. A. (2017). Nanopore sensing. Analytical chemistry

86. Asandei, A., Rossini, A. E., Chinappi, M., Park, Y., & Luchian, T. (2017). Protein nanopore-based discrimination between selected neutral amino acids from polypeptides. Langmuir

87. Huang, G., Willems, K., Soskine, M., Wloka, C., & Maglia, G. (2017). Electro-osmotic capture and ionic discrimination of peptide and protein biomarkers with FraC nanopores. Nature communications

88. Piguet, F., Ouldali, H., Pastoriza-Gallego, M., Manivet, P., Pelta, J., & Oukhaled, A. (2018). Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore. Nature communications

89. Robertson, J. W., & Reiner, J. E. (2018). The utility of nanopore technology for protein and peptide sensing. Proteomics

90. Thakur, A. K., & Movileanu, L. (2019). Real-time measurement of protein–protein interactions at single-molecule resolution using a biological nanopore. Nature biotechnology

91. Harrington, L., Cheley, S., Alexander, L. T., Knapp, S., & Bayley, H. (2013). Stochastic detection of Pim protein kinases reveals electrostatically enhanced association of a peptide substrate. Proceedings of the National Academy of Sciences

92. Sutherland, T. C., Long, Y. T., Stefureac, R. I., Bediako-Amoa, I., Kraatz, H. B., & Lee, J. S. (2004). Structure of peptides investigated by nanopore analysis. Nano letters

93. Movileanu, L., Schmittschmitt, J. P., Scholtz, J. M., & Bayley, H. (2005). Interactions of peptides with a protein pore. Biophysical journal

94. Kennedy, E., Dong, Z., Tennant, C., & Timp, G. (2016). Reading the primary structure of a protein with 0.07 nm 3 resolution using a subnanometre-diameter pore. Nature nanotechnology

95. Cherf, G. M., Lieberman, K. R., Rashid, H., Lam, C. E., Karplus, K., & Akeson, M. (2012). Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. Nature biotechnology

96. Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., ... & Turner, D. J. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nature methods

97. Manrao, E. A., Derrington, I. M., Laszlo, A. H., Langford, K. W., Hopper, M. K., Gillgren, N., ... & Gundlach, J. H. (2012). Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. Nature biotechnology

98. Maillard, R. A., Chistol, G., Sen, M., Righini, M., Tan, J., Kaiser, C. M., ... & Bustamante, C. (2011). ClpX (P) generates mechanical force to unfold and translocate its protein substrates. Cell

99. Kenniston, J. A., Burton, R. E., Siddiqui, S. M., Baker, T. A., & Sauer, R. T. (2004). Effects of local protein stability and the geometric position of the substrate degradation tag on the efficiency of ClpXP denaturation and degradation. Journal of structural biology

100. Baldus, I. B., & Gräter, F. (2012). Mechanical force can fine-tune redox potentials of disulfide bonds. Biophysical journal

101. Burton, R. E., Siddiqui, S. M., Kim, Y. I., Baker, T. A., & Sauer, R. T. (2001). Effects of protein stability and structure on substrate processing by the ClpXP unfolding and degradation machine. The EMBO journal

102. Cordova, J. C., Olivares, A. O., Shin, Y., Stinson, B. M., Calmat, S., Schmitz, K. R., ... & Sauer, R. T. (2014). Stochastic but highly coordinated protein unfolding and translocation by the ClpXP proteolytic machine. Cell

103. Miller, J. M., Lin, J., Li, T., & Lucius, A. L. (2013). E. coli ClpA catalyzed polypeptide translocation is allosterically controlled by the protease ClpP. Journal of molecular biology

104. Zhang, S., Huang, G., Versloot, R., Herwig, B. M., de Souza, P. C. T., Marrink, S. J., & Maglia, G. (2020). Bottom-up fabrication of a multi-component nanopore sensor that unfolds, processes and recognizes single proteins. bioRxiv.