BRIEF COMMUNICATION
https://doi.org/10.1038/s41587-021-01002-6
nature biotechnology

Multiplexed direct detection of barcoded protein reporters on a nanopore array

Nicolas Cardozo[1,2,5], Karen Zhang[1,5], Kathryn Doroschak[1,5], Aerilynn Nguyen[1], Zoheb Siddiqui[1], Nicholas Bogard[3], Karin Strauss[1,4], Luis Ceze[1] and Jeff Nivala[1]✉

Detection of specific proteins using nanopores is currently challenging. To address this challenge, we developed a collection of over twenty nanopore-addressable protein tags engineered as reporters (NanoporeTERs, or NTERs). NTERs are constructed with a secretion tag, folded domain and a nanopore-targeting C-terminal tail in which arbitrary peptide barcodes can be encoded. We demonstrate simultaneous detection of up to nine NTERs expressed in bacterial or human cells using MinION nanopore sensor arrays.

For nearly four decades, reporter proteins have been used as a means to track biological activities such as genetic regulation[1]. 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[2–5]. 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[2]. 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[6], one-pot parallelized whole-cell biosensing[7] and efficient genetic circuit debugging[8].

While biomolecular sensing with nanopore sensors has been explored[9], only recently have high-throughput nanopore sensor platforms emerged for real-time sequencing of DNA[10] and RNA[11]. 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 call NTERs, that use commercially available nanopore sensors (Oxford Nanopore Technologies' MinION device)[10] for multiplexed direct protein reporter detection without the need for any other specialized equipment or laborious sample preparation before analysis (Fig. 1a–d).

To develop this method, we first designed a protein (NTERY00) that would be easily detectable by a nanopore sensor (Fig. 1a, Supplementary Figs 1 and 2 and Supplementary Notes). This protein has three important elements: (1) a long, flexible, negatively charged C-terminal domain (polyGSD) to promote capture of the protein in a nanopore sensor[12,13]; (2) a stable folded domain (Smt3) to sterically inhibit complete translocation of the protein through the pore; and (3) an N-terminal secretion domain (OsmY) to promote transport of the protein to the extracellular medium following expression[14,15]. We expressed and purified NTERY00 from *Escherichia coli* culture supernatant by immobilized metal affinity chromatography (IMAC) and determined whether 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[16]) and a custom MinION run script (Methods). The script applies a constant voltage of −180 mV to all active pores on the flow cell and statically flips the voltage in the reverse direction in 15-s cycles (10 s 'ON' at −180 mV and 5 s 'OFF' or in 'Reverse'; Fig. 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 under 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 (Fig. 1e and Supplementary Fig. 3), signaling the 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 (Fig. 1f and Supplementary Fig. 4). We also explored whether shorter C-terminal NTER 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 (Supplementary Fig. 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.

If this model is correct, we postulated that the ionic current characteristics of the NTERY00 capture state would 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 (Fig. 1g). Tyrosines were chosen because their large side-chain structure was predicted to decrease 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 NTER01–15, 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 (Fig. 1h–j and Supplementary Figs. 3 and 6). These results support a model in which the first ~17 amino acids of the polyGSD tail reside within the CsgG nanopore's sensitive region and contribute to its ionic current signature during a capture event, essentially demarcating

[1]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [2]Molecular Engineering PhD Program, University of Washington, Seattle, WA, USA. [3]Department of Electrical and Computer Engineering, University of Washington, Seattle, WA, USA. [4]Microsoft Research, Redmond, WA, USA. [5]These authors contributed equally: Nicolas Cardozo, Karen Zhang, Kathryn Doroschak. ✉e-mail: jmdn@uw.edu

this segment of the NTER as a nanopore-addressable amino acid 'barcode'. We then further characterized this barcode signal space by assessing how different amino acid types and phosphomimetic point mutations[17] can modulate both the NTER nanopore ionic current signal (Fig. 1k–n, Supplementary Fig. 7 and Supplementary Notes) and NTER capture efficiency into the pore (Supplementary Fig. 8).

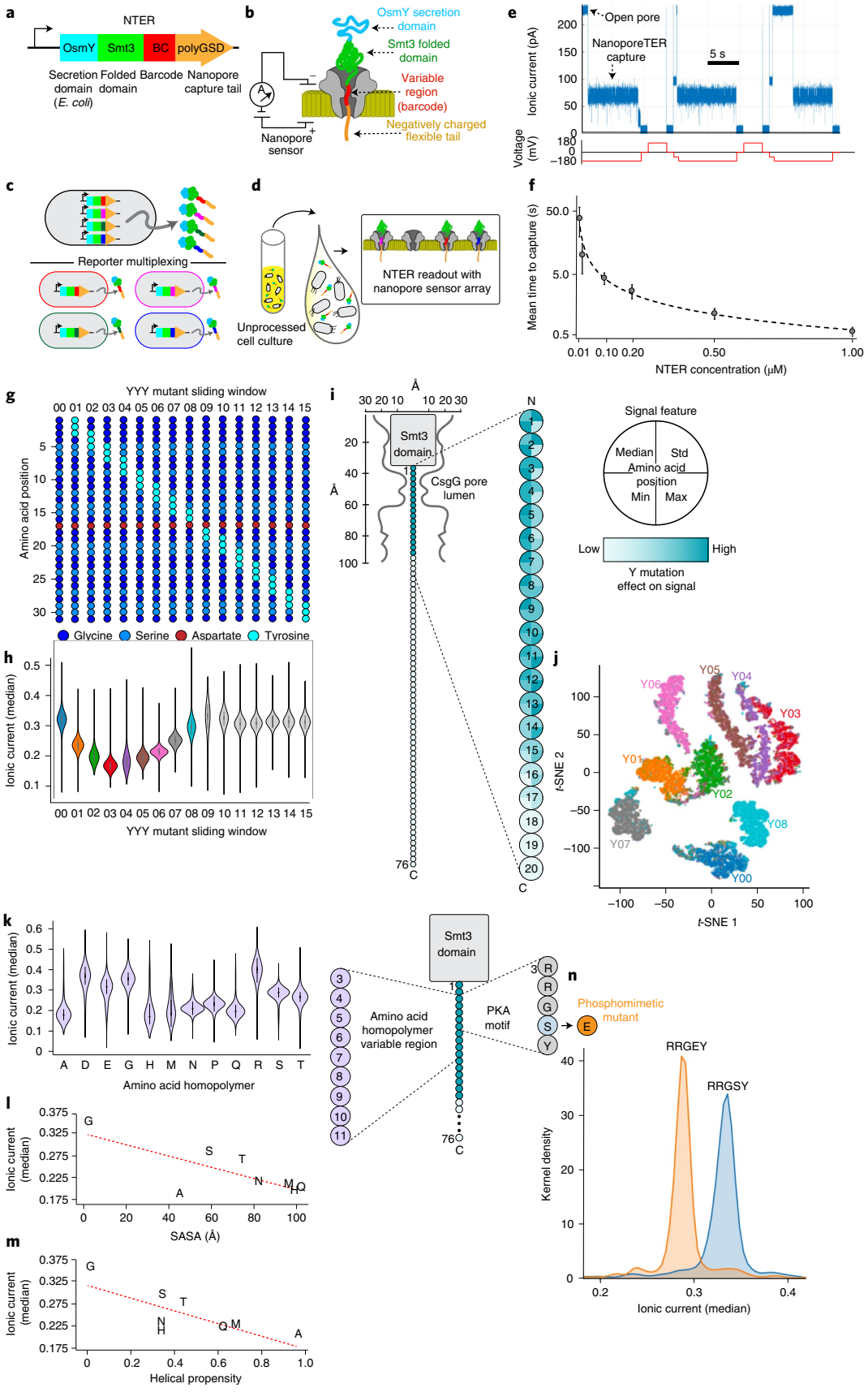
Having explored the potential NTER barcode sequence and signal space, we sought to demonstrate proof-of-principle NTER applications for multiplexed tracking of gene expression. To do this, we first used supervised machine learning to train classifiers that could accurately discriminate amongst combinations of the NTER barcodes explored above (Fig. 2a,b and Supplementary Notes). Using either a set of engineered signal features as input to a random forest classifier or the raw ionic current signal directly into a convolutional neural network (CNN) (Fig. 2a), we used our purified NTER datasets for model training and validation. Both models achieved similar accuracy, ranging from ~80 to 90% depending on the model hyperparameters and barcode set (Fig. 2b and Methods). Next, we used the CNN classifier to assess whether NTERs were immediately recaptured in the nanopore at a non-negligible frequency following their initial capture and ejection, which could lead to molecules being double counted and thus affecting relative NTER quantification. To investigate this, we analyzed the frequency with which successive captures of the same or different barcodes occurred in a mixed-barcode experiment with five different NTER barcodes mixed at varying concentrations (Y00, 0.05 µM; Y02, 0.1 µM; Y05, 0.05 µM; Y07, 0.2 µM; and Y08, 0.1 µM). We found that successive NTER captures of the same barcode were not disproportionately represented, suggesting that immediate recapture of the same NTER molecule was not occurring at a high frequency (Supplementary Fig. 9).

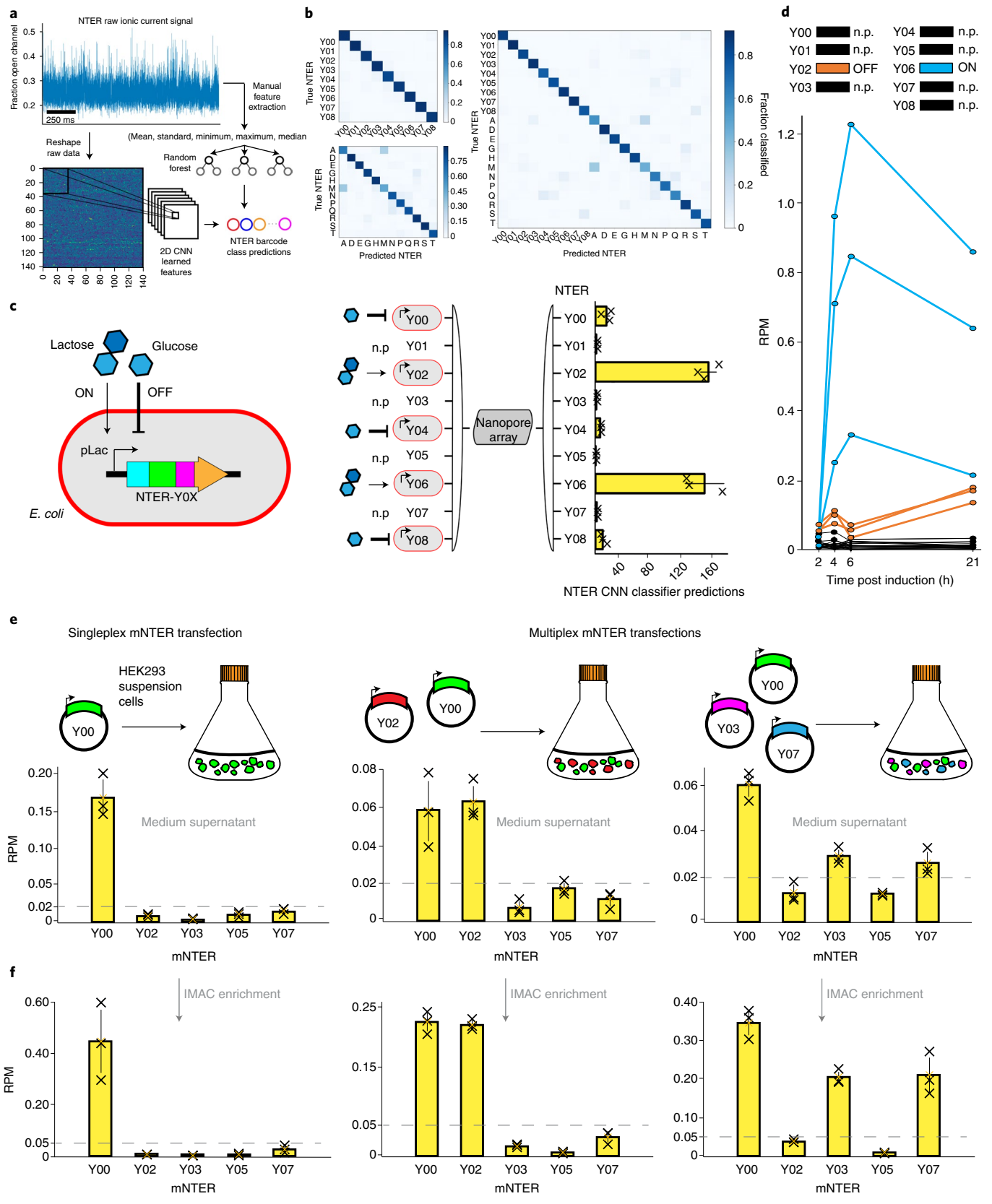
We then used the best-performing classifier that was trained on NTERY00–08 (in addition to a background/noise class; Methods) to determine relative NTER expression levels within bacterial cultures composed of mixed populations of strains engineered with different NTER-tagged, plasmid-based circuits. Specifically, we grew independent mono-barcoded cultures overnight with NTER expression either induced or inhibited (by autoinduction medium containing lactose or lysogeny broth (LB) supplemented with glucose, respectively). In the morning, just before nanopore readout, the cultures were mixed in a single solution, diluted into MinION running buffer and loaded directly into a flow cell for analysis. These cell cultures underwent no processing or purification before analysis, in contrast to our previous experiments. Results from these experiments showed higher classification counts for NTER barcodes for which expression was induced (NTERY02 and Y06), and lower counts for strains that were either inhibited (glucose: NTERY00, Y04 and Y08) or not present at all in the mixed population (NTERY01, Y03, Y05 and Y07) for all replicates (Fig. 2c) over a 10-min MinION runtime. We then conducted a time course experiment in which we tracked the expression of two different NTERs over multiple hours, one of which was induced with isopropyl-β-d-thiogalactopyranoside (IPTG) (NTERY06) and the other in which NTER expression was inhibited with glucose (NTERY02). Again, cultures were grown independently but then mixed just before nanopore readout. Figure 2d shows the results of this time course (and replicates) during 10 min of MinION analysis at 2-, 4-, 6- and 21-h time points following induction (NTERY06) or inhibition (NTERY02) of the NTER circuit. Again, the rate of NTER classification (RPM) was substantially higher for the induced NTERY06 circuits compared to the uninduced NTERY02 circuits. Leaky expression of NTERY02 was still detectable over the background false-positive classification rates for the NTER barcodes that were not present in the experiment (Y00, Y01, Y03, Y04, Y05, Y07 and Y08). These results demonstrate that NTERs can be used as reliable reporters of relative protein expression levels in bacterial cell culture and that cell-based assays can be performed directly on the flow cell itself.

To show that NTERs can also be used for detection of gene expression in alternative cell types, such as mammalian cell cultures, we modified the *E. coli* NTER design to function in HEK293 cells (Supplementary Fig. 1). We did this by making two key changes: (1) we replaced the bacterial secretion domain (OsmY) with a human secretion tag (IFNα2)[18]; and (2) made two mutations to the Smt3 domain that increase its resistance to intracellular degradation in mammalian cells (SUMOstar)[19]. We then cloned several different barcoded versions of this mammalian-optimized NTER design (mNTER) into a vector under the control of a constitutive cytomegalovirus (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 Fig. 2e,f. Specifically, we detected mNTERs directly from medium supernatant collected 3 days after the cultures were transfected with one (Y00), two (Y00 and Y02) or three (Y00, Y03 and Y07) different mNTER barcodes. The number of mNTER counts for each barcode class was reflective of the barcode combinations introduced into each of the cultures, as shown by the substantially higher classification counts for the transfected barcode classes relative to those included in the classifier but absent from the experiment. We observed that, while mNTERs could be detected over background classification levels directly from the raw medium supernatant with no further processing (Fig. 2e), superior classification results were

Fig. 1 | NTERs. a, Gene schematic of NTER design. **b**, Schematic of an NTER captured within a nanopore. **c**, NTERs are designed to enable multiplexed readout of protein expression. **d**, Secretion of NTERs into the extracellular medium. **e**, Example of raw nanopore data generated from a single nanopore showing repeated captures and ejections of NTERY00. **f**, Concentration curve showing the relationship between NTER concentration within a flow cell and the average time between captures or 'reads'. Error bars represent s.d. of $n = 3$ independent nanopore experiments. **g**, Schematic of NTERY00–15 sequences. **h**, Violin plot showing normalized median current level of nanopore capture state for NTERY00–15. Center dot, median; black box, first/third quartiles; whiskers, first/third quartiles ±1.5 interquartile range. Each NTER distribution is composed of $n = ~4{,}000$ events per class. **i**, Model of NTER position within the nanopore during a capture event. Heat map displaying relative change to specific signal features projected onto NTER tail residue positions that were mutated in NTERY00–15, showing the relative magnitude of effect of tyrosine mutations at each residue on the NTER nanopore signal. **j**, *t*-Distributed stochastic neighbor-embedding (*t*-SNE) plot clustering NTER reads based on ionic current signal features, colored by NTER barcode identity (Y00–08). $n = ~4{,}000$ events per class. **k**, Violin plot showing the normalized median ionic current level of the nanopore capture state for amino acid homopolymer NTERs. Center dot, median; black box, first and third quartiles; whiskers, first and third quartiles ±1.5 × interquartile range. Each NTER distribution is composed of ~1,500 single-molecule measurements. **l**, Scatter plot showing the relationship between amino acid solvent accessible surface area (SASA) and the respective amino acid homopolymer NTER mutant's normalized median ionic current level. **m**, Scatter plot showing the relationship between amino acid helical propensity and the respective amino acid homopolymer NTER mutant's normalized median ionic current level. **n**, Kernel density plot comparing the normalized ionic current median of an NTER containing a protein kinase A (PKA) phosphorylation motif within its barcode region with those with a phosphomimetic mutation. Each NTER distribution is composed of ~1,000 single-molecule measurements.

obtained from medium samples following IMAC enrichment (Fig. 2f), indicating that medium contaminants in cell culture led to higher levels of mNTER misclassification events. These results, together with the previous results obtained from *E. coli*, suggest that NTERs will be applicable to a wide diversity of cell types and model systems.





In conclusion, we have laid the foundations for a class of multiplexable protein reporters that can be analyzed using a commercially available nanopore sensor array, the ONT MinION. To support future applications, we have also performed preliminary assessments of the experimental throughput of NTER measurements on the MinION and discuss the scalability of NTER barcode space for future work (Supplementary Figs. 10–13 and Supplementary Notes). We foresee many potential NTER applications, including simultaneous reading of protein-level outputs of many genetically engineered circuit components in one pot, enabling more efficient debugging and tuning

Fig. 2 | Classification and multiplexed detection of NTER expression levels with a MinION. a, Raw ionic current data were classified using either a set of engineered features (mean, standard, minimum, maximum and median) or the unprocessed signal directly, and input into either a random forest or CNN classifier, respectively. **b**, Confusion matrices showing the random forest test set classification accuracies on models using different combination of NTER barcodes. Top left: NTERY00–08; bottom left: amino acid homopolymer mutants A, D, E, G, H, M, N, P, Q, R, S and T; right: both NTERY00–08 and amino acid homopolymer mutants. **c**, Schematic showing the gene construct used for controllable NTER expression. Lactose is used to induce NTER expression (ON) using a lactose-inducible promoter (pLac) while glucose inhibits expression (OFF). The diagram and bar plot on the right show the results of a mixed-culture experiment in which NTER expression was induced for NTERY02 and Y04 and inhibited for NTERY00, Y02 and Y08. NTERY01, Y03, Y05 and Y07 were not present (n.p.) in the experiment as negative controls. Bars show the average total number of reads classified as each NTER barcode during MinION analysis of three technical replicates (error bars show s.d.). **d**, Line plot showing time course of NTER expression levels as determined by the rate of classified reads (reads per pore min^{-1} (RPM)) for each NTER barcode. NTERY06 was induced while NTERY02 was inhibited. The other NTERs were not present (n.p.) as negative controls and show false-positive classification rates. Three technical replicates for each condition are plotted. **e**, Bar plots showing the results of singleplex and multiplex 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 RPM for each barcode during MinION analysis. Three technical replicates for each experiment are plotted (error bars show s.d.). **f**, As in **e**, but with the addition of an IMAC purification step before MinION analysis.

than permitted by current analysis methods. For instance, in comparison to traditional sets of fluorescent protein reporters, NTERs 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). While RNA-sequencing can be used to measure the transcriptional output of many circuits in parallel[20], our method has these advantages: (1) little to no sample preparation, which makes it more amenable to automation[21–23] and reduces both time to analysis (latency) and cost; (2) a nondestructive readout, enabling the profiling of expression dynamics in living cells; and (3) direct detection of outputs at the protein level. The last of these creates new opportunities for engineering of reporters with NTER barcodes that can report simultaneously on both protein expression and specific post-translational modifications. We anticipate that this capability will be especially useful for synthetic protein-level circuit engineering[24].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-01002-6.

Received: 16 October 2019; Accepted: 28 June 2021;
Published online: 12 August 2021

References

1. Ghim, C. M., Lee, S. K., Takayama, S. & Mitchell, R. J. The art of reporter proteins in science: past, present and future applications. *BMB Rep.* **43**, 451–460 (2010).
2. Rodriguez, E. A. et al. The growing and glowing toolbox of fluorescent and photoactive proteins. *Trends Biochem. Sci* **42**, 111–129 (2017).
3. Martin, L., Che, A. & Endy, D. Gemini, a bifunctional enzymatic and fluorescent reporter of gene expression. *PLoS ONE* **4**, e7569 (2009).
4. Parrello, D., Mustin, C., Brie, D., Miron, S. & Billard, P. Multicolor whole-cell bacterial sensing using a synchronous fluorescence spectroscopy-based approach. *PLoS ONE* **10**, e0122848 (2015).
5. 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* **13**, e0197373 (2018).
6. Wroblewska, A. et al. Protein barcodes enable high-dimensional single-cell CRISPR screens. *Cell* **175**, 1141–1155 (2018).
7. He, W., Yuan, S., Zhong, W. H., Siddikee, M. A. & Dai, C. C. Application of genetically engineered microbial whole-cell biosensors for combined chemosensing. *Appl. Microbiol. Biotechnol.* **100**, 1109–1119 (2016).
8. Nielsen, A. A. K. et al. Genetic circuit design automation. *Science* **352**, aac7341 (2016).
9. Shi, W., Friedman, A. K. & Baker, L. A. Nanopore sensing. *Anal. Chem.* **89**, 157–188 (2017).
10. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).
11. Garalde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
12. Nivala, J., Marks, D. B. & Akeson, M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat. Biotechnol.* **31**, 247–250 (2013).
13. 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).
14. Yim, H. H. & Villarejo, M. osmY, a new hyperosmotically inducible gene, encodes a periplasmic protein in *Escherichia coli*. *J. Bacteriol.* **174**, 3637–3644 (1992).
15. Kotzsch, A. et al. A secretory system for bacterial production of high-profile protein targets. *Protein Sci.* **20**, 597–609 (2011).
16. Goyal, P. et al. Structural and mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature* **516**, 250–253 (2014).
17. Taylor, S. S. et al. PKA: a portrait of protein kinase dynamics.*Biochim. Biophys. Acta Proteins Proteom.* **1697**, 259–269 (2004).
18. 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).
19. 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.* **17**, 1586–1595 (2008).
20. Gorochowski, T. E. et al. Genetic circuit characterization and debugging using RNA-seq. *Mol. Syst. Biol.* **13**, 952 (2017).
21. Gach, P. C. et al. A droplet microfluidic platform for automating genetic engineering. *ACS Synth. Biol.* **5**, 426–433 (2016).
22. Chao, R., Mishra, S., Si, T. & Zhao, H. Engineering biological systems using automated biofoundries. *Metab. Eng.* **42**, 98–108 (2017).
23. Madison, A. C. et al. Scalable device for automated microbial electroporation in a digital micro fluidic platform. *ACS Synth. Biol.* **6**, 1701–1709 (2017).
24. Chen, Z. & Elowitz, E. B. Programmable protein circuit design. *Cell* **184**, 2284–2301 (2021).

Methods

NTER construction, cloning, expression and purification. The initial NTER 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 New England Biolabs' cloning protocol. Sequence verification was obtained through Genewiz Inc. Expression of the NTER protein was done in the 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 method used the associated buffer set from Takara, following their specified protocol. Proteins were concentrated using Amicon Ultra 0.5-ml centrifugal filters with Ultracel 30 K (Amicon). The final concentration of proteins averaged ~7 mg ml^{-1} from 5-ml overnight cultures. Purified proteins were stored over the long term at −80 °C in 10-µl aliquots, and over the short term at 4 °C.

***E.coli* raw culture-mixing experiments.** Cultures were picked from single colonies on plates and used to inoculate 3 ml of LB supplemented with 0.5 mM isopropylthiogalactoside (IPTG) and kanamycin (induced), or 3 ml of LB supplemented with 0.2% glucose and kanamycin (inhibited). After overnight incubation at 37 °C with shaking, cultures were equally mixed (45 µl of culture, 50 µl of 4x C17 buffer (2 M KCl, 100 mM HEPES, pH 8.0) and 105 µl of water (total volume, 200 µl)). 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 30 µl of overnight culture (LB) into 3 ml of fresh LB supplemented with 0.5 mM IPTG and kanamycin (induced), or 3 ml of fresh LB supplemented with 0.2% glucose and kanamycin (inhibited). The cultures were placed in a shaker/incubator at 37 °C to facilitate culture growth. Samples were then collected at 2-, 4-, 6- and 21-h time points. At each time point, cultures were equally mixed (10 µl of culture, 50 µl of 4x C17 buffer and 140 µl of water (total volume, 200 µl)). 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 °C 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 dilution of analyte solution into C17 buffer for a final concentration of 0.5 M KCl and 25 mM HEPES (pH 8.0) into the flow cell priming port. Flow cells were run on the MinION at a temperature of 30 °C and a run voltage of −180 mV with 10-khz sampling frequency and 15-s 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 with 1 ml of C17 buffer three times between experiments. Flow cells were stored at 4 °C in C18 buffer (150 mM potassium ferrocyanide, 150 mM potassium ferricyanide, 25 mM potassium phosphate, pH 8.0) when not in use.

Nanopore signal analysis, quantification and classification. The analysis pipeline for a NTER sequencing run begins with extraction of the segments of the raw nanopore signal that contain capture events. A capture is defined as a region where the signal current falls to <70% of the open-pore current for a duration of at least 1 ms. 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 NTER captures from noise captures based on features of the normalized raw current (mean, standard deviation, minimum, maximum and 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 captures within each NTER class are subsequently fed to a quantifier that 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 RPM per class per active pore. In addition to the NTER datasets, a background/noise class dataset was also used in training the models to recognize data generated from non-NTER-specific pore blockages that progressed through the filtering step. These data were collected from experiments in which only running buffer, LB medium and/or NTER-free *E. coli* or HEK293 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 was a CNN implemented in PyTorch. An 80/20 training/test split was used to generate the classification accuracy estimations and confusion matrix results. For both models, only the first 2 s of each capture was considered for analysis. The random forest was trained on an array composed of the mean, standard deviation, minimum, maximum and median of that 2-s window. Default random forest hyperparameters were modified to: n_estimators=300 and max_depth=100. The CNN used the 2 s of raw signal directly as input following reshaping of the one-dimensional signal into a two-dimensional (2D) structure. The neural network was composed of four 2D convolutional layers, each with rectified linear unit activation and maximum 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 (for example, assigning labels for only those events with >95% confidence in a single class). Confidence thresholds of 95 and 90% were used for the *E.* coli and HEK293 cell culture experiments, respectively. Full model details and code can be found at https://github.com/uwmisl/NanoporeTERs.

While both classifiers (random forest and CNN) achieved similar accuracy on the training/test data, the CNN is more likely to be extensible as we increase the number of barcodes. With the random forest, the information content available for classification is reduced to just a few manually extracted summary statistics (ionic current blockade mean, median, maximum, minimum and standard deviation), leaving other distinguishing signal characteristics behind. In contrast, the CNN classifies barcodes directly from the ionic current traces. As the number of barcodes increases, a CNN will probably be able to take better advantage of features embedded in the signal itself, including variation in the noise pattern for a specific barcode.

Materials availability. All plasmids and cells lines used in this study are available upon request.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data are available upon request and on can be found on github.com/uwmisl/NanoporeTERs.

Code availability

Codes are available upon request and can be found on github.com/uwmisl/NanoporeTERs. Custom MinION MinKNOW runscripts can also be obtained from Oxford Nanopore Technologies upon request.

Acknowledgements

We thank additional members of the Molecular Information Systems Lab for helpful discussion and feedback on this work. The OsmY expression plasmid was generously provided by C. Bryan and L. Carter (Institute for Protein Design, University of Washington). We also thank A. Heron and R. Gutierrez (Oxford Nanopore Technologies) for providing the configurable MinION run script and discussions on its use, and M. Jain (UCSC) for a custom Matlab script that facilitated visualization of the raw MinION data. This work was supported in part by NSF EAGER Award no. 1841188 and NSF CCF Award no. 2006864 to L.C. and J.N., an NIH/NCI Cancer Center Support Grant (no. P30 CA015704) Pilot Award and NSF Award 2021552 to J.N. and a sponsored research agreement from Oxford Nanopore Technologies.

Author contributions

N.C., K.Z., A.N. and N.B. performed wet laboratory experiments. K.Z. and K.D. developed the data analysis pipeline and performed computational analyses. Z.S.

implemented the machine learning approach. N.B., K.S., L.C. and J.N. supervised the project. J.N. conceived and directed the project. All authors contributed to writing and editing of the manuscript.

Competing interests

A provisional patent has been filed by the University of Washington covering aspects of this work (Patent Application no. 17/283,007). K.S. is an employee of Microsoft. J.N. is a consultant to Oxford Nanopore Technologies. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-01002-6.

Correspondence and requests for materials should be addressed to J.N.

Peer review information *Nature Biotechnology* thanks Yi-Tao Long, Meni Wanunu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature research

Corresponding author(s): Jeff Nivala

Last updated by author(s): Apr 27, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	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.
Data analysis	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 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.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data and code can be found at https://github.com/uwmisl/NanoporeTERs.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed.
Data exclusions	Data were not excluded from analysis except as described in Data Analysis, which includes filtering steps to remove nanopore signal noise.
Replication	Experiments were conducted in triplicate to assess reproducibility.
Randomization	Not applicable to this study; does not include subjects that require allocation into experimental groups.
Blinding	Not applicable to this study; does not include subjects that require allocation into blinded experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HEK FreeStyle 293-F (Thermo catalog #R79007)
Authentication	cell lines were not authenticated
Mycoplasma contamination	cell lines were not tested for mycoplasma
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*