PEPTIDE SEQUENCING

# Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device

Brian D. Reed[1]*, Michael J. Meyer[1], Valentin Abramzon[1], Omer Ad[1], Pat Adcock[1], Faisal R. Ahmad[1], Gün Alppay[1], James A. Ball[1], James Beach[1], Dominique Belhachemi[1], Anthony Bellofiore[1], Michael Bellos[1], Juan Felipe Beltrán[1], Andrew Betts[1], Mohammad Wadud Bhuiya[1], Kristin Blacklock[1], Robert Boer[1], David Boisvert[1], Norman D. Brault[1], Aaron Buxbaum[1], Steve Caprio[1], Changhoon Choi[1], Thomas D. Christian[1], Robert Clancy[1], Joseph Clark[1], Thomas Connolly[1], Kathren Fink Croce[1], Richard Cullen[1], Mel Davey[1], Jack Davidson[1], Mohamed M. Elshenawy[1], Michael Ferrigno[1], Daniel Frier[1], Saketh Gudipati[1], Stephanie Hamill[1], Zhaoyu He[1], Sharath Hosali[1], Haidong Huang[1], Le Huang[1], Ali Kabiri[1], Gennadiy Kriger[1], Brittany Lathrop[1], An Li[1], Peter Lim[1], Stephen Liu[1], Feixiang Luo[1], Caixia Lv[1], Xiaoxiao Ma[1], Evan McCormack[1], Michele Millham[1], Roger Nani[1], Manjula Pandey[1], John Parillo[1], Gayatri Patel[1], Douglas H. Pike[1], Kyle Preston[1], Adeline Pichard-Kostuch[2], Kyle Rearick[1], Todd Rearick[1], Marco Ribezzi-Crivellari[2], Gerard Schmid[1], Jonathan Schultz[1], Xinghua Shi[1], Badri Singh[1], Nikita Srivastava[1], Shannon F. Stewman[1], T. R. Thurston[1], Philip Trioli[1], Jennifer Tullman[1], Xin Wang[1], Yen-Chih Wang[1], Eric A. G. Webster[1], Zhizhuo Zhang[1], Jorge Zuniga[1], Smita S. Patel[3], Andrew D. Griffiths[2], Antoine M. van Oijen[4], Michael McKenna[1], Matthew D. Dyer[1], Jonathan M. Rothberg[1]

Studies of the proteome would benefit greatly from methods to directly sequence and digitally quantify proteins and detect posttranslational modifications with single-molecule sensitivity. Here, we demonstrate single-molecule protein sequencing using a dynamic approach in which single peptides are probed in real time by a mixture of dye-labeled N-terminal amino acid recognizers and simultaneously cleaved by aminopeptidases. We annotate amino acids and identify the peptide sequence by measuring fluorescence intensity, lifetime, and binding kinetics on an integrated semiconductor chip. Our results demonstrate the kinetic principles that allow recognizers to identify multiple amino acids in an information-rich manner that enables discrimination of single amino acid substitutions and posttranslational modifications. With further development, we anticipate that this approach will offer a sensitive, scalable, and accessible platform for single-molecule proteomic studies and applications.

Measurements of the proteome provide deep and valuable insight into biological processes. However, methods with higher sensitivity are needed to fully understand the complex and dynamic states of the proteome in cells and changes to the proteome that occur in disease states and to make this information more accessible. The complex nature of the proteome and the chemical properties of proteins present several fundamental challenges to achieving sensitivity, throughput, cost, and adoption on par with DNA sequencing technologies (*1*, *2*). These challenges include the large number of different proteins per cell (>10,000) and yet larger number of proteoforms (*3*); the very wide dynamic range of protein abundance in cells and biological fluids (*4*, *5*) and lack of correlation with transcript levels (*6*); the costs and high detection limits of current methods of mass spectrometry (*2*); and the inability to copy or amplify proteins. Methods to directly sequence single protein molecules offer the maximum possible detection sensitivity, with the potential to enable single-cell inputs, digital quantification based on read counts, detection of posttranslational modifications (PTMs) and low-abundance or aberrant proteoforms, and cost and throughput levels that favor broad adoption.

Here, we present a single-molecule protein sequencing approach and integrated system for proteomic studies. We immobilize peptides in nanoscale reaction chambers on a semiconductor chip and detect N-terminal amino acids (NAAs) with dye-labeled NAA recognizers in real time. Aminopeptidases sequentially remove individual NAAs to expose subsequent amino acids for recognition, eliminating the need for complex chemistry and fluidics (Fig. 1). We built a benchtop device with a 532-nm pulsed laser source for fluorescence excitation and electronics for signal processing (fig. S1A). Our semiconductor chip uses fluorescence intensity and lifetime, rather than emission wavelength, for discrimination of dye labels. Our recognizers detect one or more types of NAAs and provide information for peptide identification based on the temporal order of NAA recognition and the kinetics of on-off binding.

[1]Quantum-Si, Inc., Guilford, CT 06437, USA. [2]Laboratoire de Biochimie, ESPCI Paris, Université PSL, CNRS UMR 8231, Paris, France. [3]Department of Biochemistry and Molecular Biology, Rutgers University, Piscataway, NJ 08854, USA. [4]Molecular Horizons, University of Wollongong, Wollongong, NSW 2522, Australia.
*Corresponding author. Email: breed@quantum-si.com

### A complementary metal-oxide semiconductor chip and integrated system for single-molecule measurements

We used complementary metal-oxide semiconductor fabrication technology to build a custom time domain–sensitive semiconductor chip with nanosecond precision, containing fully integrated components for single-molecule detection, including photosensors, optical waveguide circuitry, and reaction chambers for biomolecule immobilization (fig. S1, B and C). We achieve observation volumes of <5 attoliters through evanescent illumination at reaction-chamber bottoms from the nearby waveguide, enabling sensitive single-molecule detection in the context of high concentrations (>1 μM) of freely diffusing dye.

The semiconductor chip uses a filterless system that excludes excitation light on the basis of photon arrival time, achieving >10,000-fold attenuation of incident excitation light. Elimination of an integrated optical filter layer increases the efficiency of fluorescence collection and enables scalable manufacturing of the chip. To discriminate fluorescent dye labels attached to NAA recognizers by fluorescence lifetime and intensity, the chip rapidly alternates between early and late signal collection windows associated with each laser pulse, thereby collecting different portions of the exponential fluorescence lifetime decay curve. The relative signal in these collection windows (termed "bin ratio") provides a reliable indication of fluorescence lifetime (fig. S1, D to H, and materials and methods).

### Ordered recognition and cleavage of NAAs on single peptide molecules in real time

For NAA binding proteins to function as recognizers, the recognizer-peptide complex must remain bound long enough (typically >120 ms on average) to generate detectable single-molecule binding events. We first focused on proteins from the N-end rule adapter family ClpS that naturally bind to N-terminal phenylalanine, tyrosine, and tryptophan (*7*–*9*). Using PS610, a recognizer we derived from ClpS2 from *Agrobacterium tumefaciens* (table S1), we established that this recognizer binds detectably to immobilized peptides with these NAAs. We also determined that the kinetics of binding differ for each NAA. To demonstrate these properties, we incubated immobilized peptides containing the initial N-terminal sequences FAA, YAA, or WAA (A, alanine; F, phenylalanine; W, tryptophan; Y, tyrosine) on separate chips with PS610 and collected data for 10 hours (see materials and methods). We observed NAA recognition by PS610, characterized by continuous on-off binding during the incubation period, with a distinct pulse duration (PD) for each

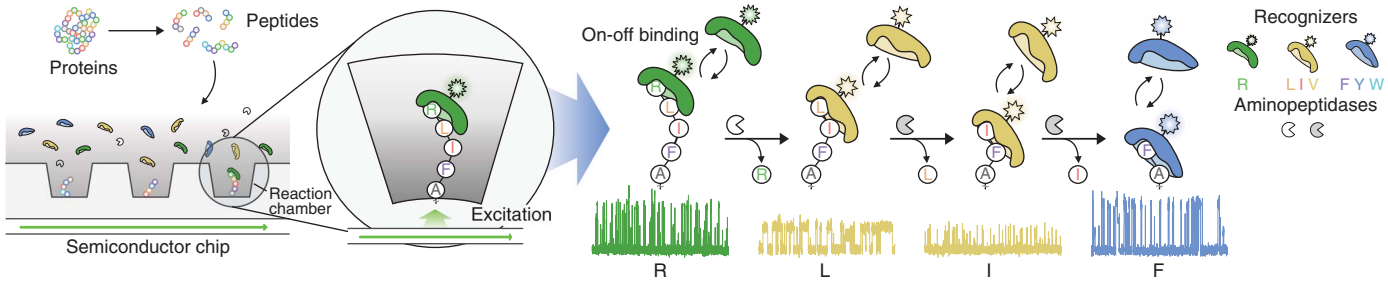


**Fig. 1. Overview of real-time dynamic protein sequencing.** Protein samples are digested into peptide fragments, immobilized in nanoscale reaction chambers, and incubated with a mixture of freely diffusing NAA recognizers and aminopeptidases that carry out the sequencing process. The labeled recognizers bind on and off to the peptide when one of their cognate NAAs is exposed at the N terminus, thereby producing characteristic pulsing patterns. The NAA is cleaved by an aminopeptidase, exposing the next amino acid for recognition. The temporal order of NAA recognition and the kinetics of binding enable peptide identification and are sensitive to features that modulate binding kinetics, such as PTMs.

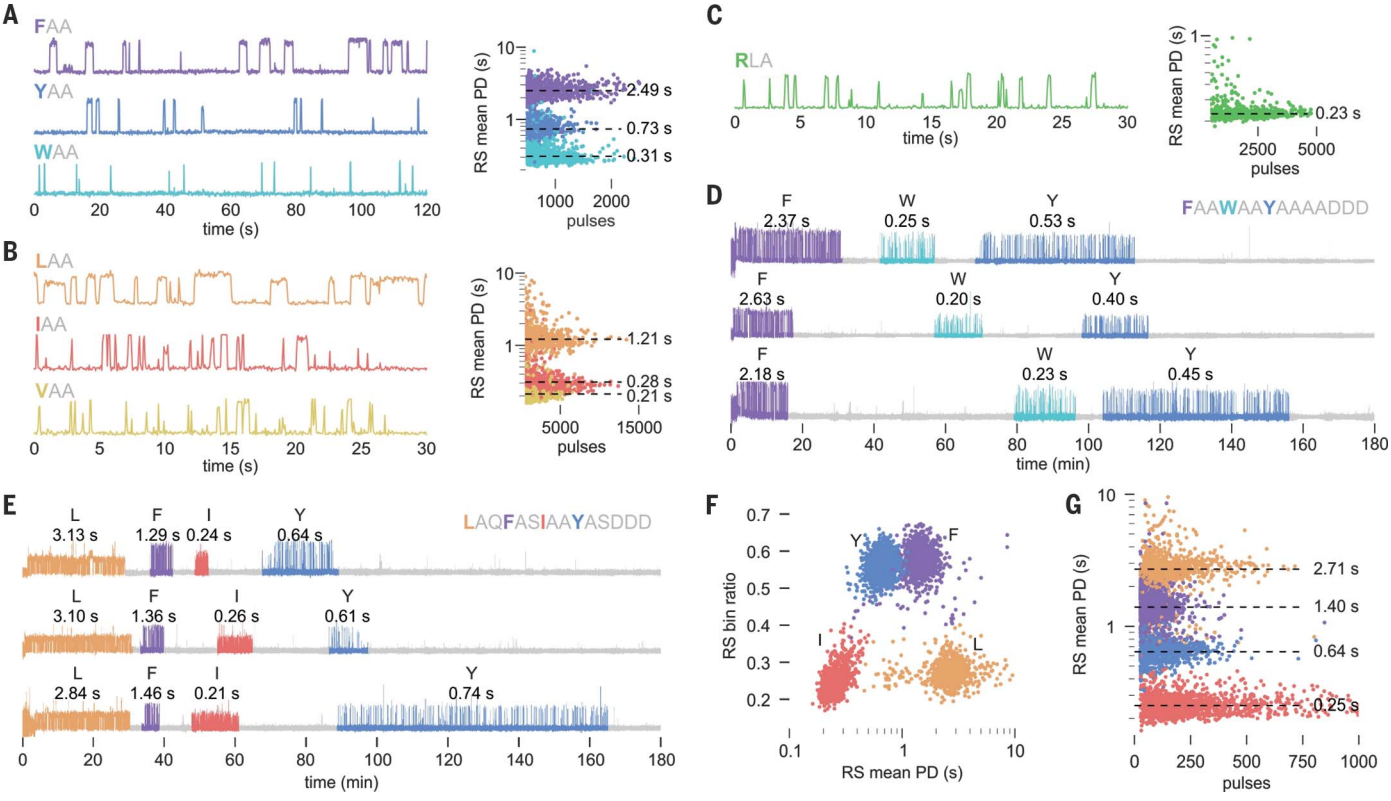


**Fig. 2. NAA recognition and dynamic sequencing.** (**A** to **C**) Example traces demonstrating single-molecule N-terminal recognition by PS610 (A), PS961 (B), and PS691 (C). Scatterplots of the number of pulses per RS versus RS mean PD are displayed for each peptide in (A) to (C), with median PD indicated. (**D**) Example traces from dynamic sequencing of the synthetic peptide FAAWAAYAADDD. Median PD is indicated above each RS. (**E** to **G**) Dynamic sequencing of the synthetic peptide LAQFASIAAYASDDD using PS610 and PS961. Example traces are shown in (E). A scatterplot of RS mean PD versus bin ratio illustrating discrimination of recognizers by bin ratio and NAAs by PD is shown in (F). A scatterplot of the number of pulses per RS versus RS mean PD, grouped by the amino acid label assigned to the RS, is shown in (G).

peptide (Fig. 2A). Median PDs were 2.49, 0.73, and 0.31 s for FAA, YAA, and WAA, respectively. These values reflect differences in binding affinity driven by different dissociation rates for each type of protein-NAA interaction (*7*) (fig. S2, A and B).

To expand the set of recognizable NAAs, we further investigated N-end rule pathway proteins as a source of additional recognizers. In a comprehensive screen of diverse ClpS family proteins, we discovered a group of ClpS proteins from the bacterial phylum Planctomycetes with native binding to N-terminal leucine, isoleucine, and valine. We applied directed evolution techniques to generate a Planctomycetes ClpS variant—PS961 (table S1)—with submicromolar affinity to N-terminal leucine, isoleucine, and valine, and demonstrated recognition of these

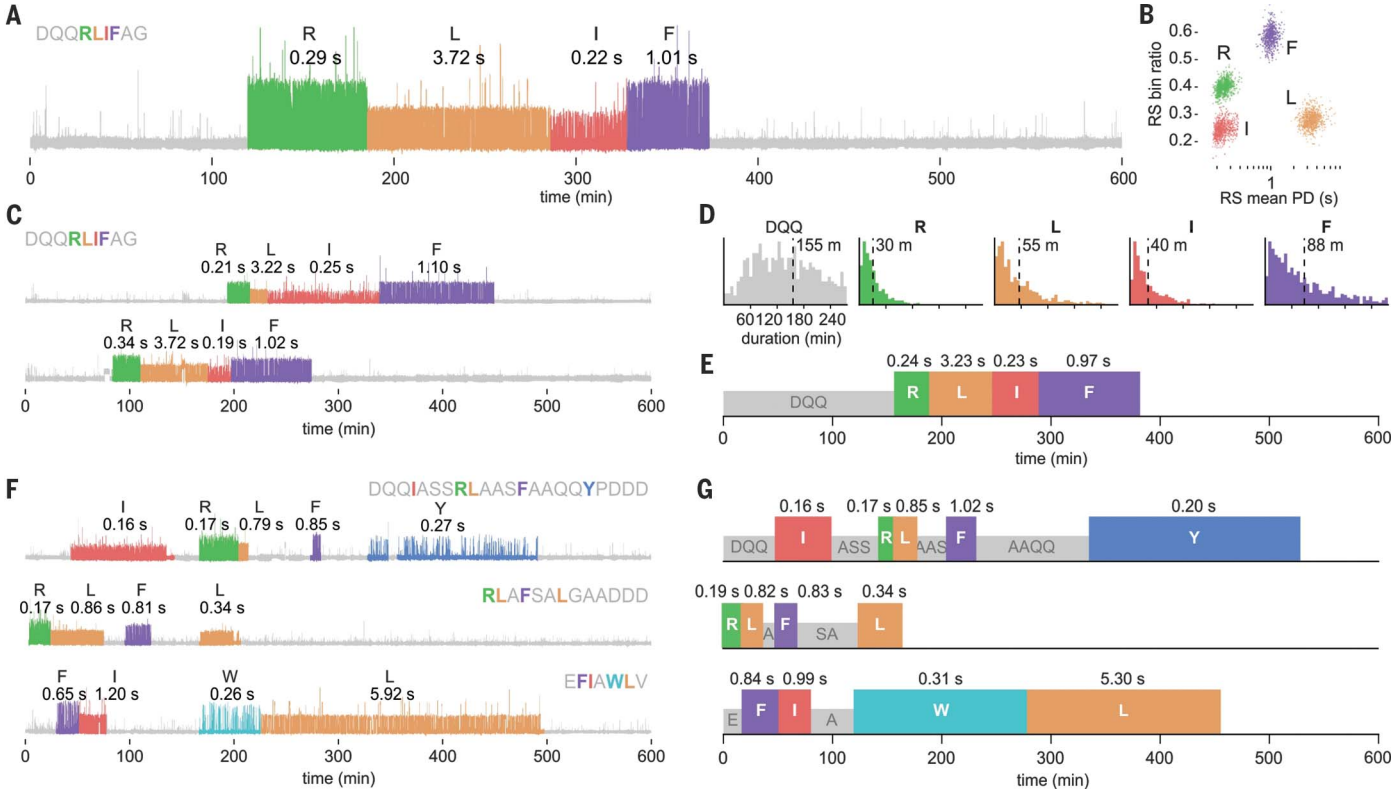


**Fig. 3. Dynamic sequencing of diverse peptides with high-precision kinetic outputs.** (**A** to **E**) Dynamic sequencing of the peptide DQQRLIFAG. An example trace is shown in (A). A scatterplot of RS mean PD versus bin ratio is shown in (B). Shown in (C) are additional example traces of dynamic sequencing of DQQRLIFAG peptide. Shown in (D) are distributions of the duration of each RS and NRS acquired during sequencing, with mean durations indicated. Kinetic signature plots summarizing the characteristic sequencing behavior of DQQRLIFAG peptide are shown in (E). (**F** to **G**) Dynamic sequencing of the synthetic peptides DQQIASSRLAASFAAQQYPDDD (top), RLAFSALGAADDD (middle), and EFIAWLV (bottom). Example traces for each peptide are shown in (F). Corresponding kinetic signature plots are shown in (G).

NAAs (Fig. 2B). The median PD of binding to peptides with N-terminal LAA, IAA, and VAA (I, isoleucine; L, leucine; V, valine) was 1.21, 0.28, and 0.21 s, respectively, in agreement with bulk characterization (fig. S2C).

In a separate screen, we investigated a diverse set of UBR-box domains from the UBR family of ubiquitin ligases that natively bind N-terminal arginine, lysine, and histidine (*10*). The UBR-box domain from the yeast *Kluyveromyces marxianus* UBR1 protein (table S1) exhibited the highest affinity for N-terminal arginine, and we used this protein to generate an arginine recognizer, PS691. PS691 recognized arginine in a peptide with N-terminal RLA (R, arginine) with a median PD of 0.23 s (Fig. 2C). Lower-affinity binding to N-terminal lysine and histidine (fig. S2, D and E) was insufficient for single-molecule detection.

To demonstrate that amino acids in a single peptide molecule can be sequentially exposed by aminopeptidases and recognized in real time with distinguishable kinetics, we incubated an immobilized peptide containing the initial sequence FAAWAAYAA with PS610 for 15 min, followed by the addition of PhTET3, an aminopeptidase from *Pyrococcus horikoshii* (*11*). The collected traces consisted of regions of distinct pulsing, which we refer to as recognition segments (RSs), separated by regions lacking recognition pulsing [nonrecognition segments (NRSs)]. We developed analysis software to automatically identify pulsing regions and transition points within traces on the basis of fluorescence properties and pulsing kinetics (see materials and methods). Traces began with the recognition of phenylalanine with a median PD of 2.36 s (Fig. 2D), in agreement with the PD observed for FAA in recognition-only assays. This pattern terminated after aminopeptidase addition (on average, 11 min after addition) and was followed by the ordered appearance of two RSs with median PDs of 0.25 and 0.49 s (Fig. 2D), corresponding to the short and medium PDs obtained in our YAA and WAA recognition-only assays. Thus, the introduction of aminopeptidase activity to the reaction resulted in the sequential appearance of discrete RSs with the expected kinetic properties in the correct order.

To demonstrate dynamic sequencing with two NAA recognizers, we labeled PS610 and PS961 with the distinguishable dyes atto-Rho6G and Cy3, respectively, and exposed an immobilized peptide of sequence LAQFASIAAYASDDD (D, aspartate; Q, glutamine; S, serine) to a solution containing both recognizers. After 15 min, we added two *P. horikoshii* aminopeptidases with combined activity covering all 20 amino acids—PhTET2 and PhTET3 (*11*, *12*). The collected traces displayed discrete segments of pulsing alternating between PS961 and PS610 according to the order of recognizable amino acids in the peptide sequence (Fig. 2E). The average bin ratio and average PD associated with each RS readily distinguished the two dye labels and four types of recognized NAAs (Fig. 2F). Median PDs were 2.71, 1.40, 0.25, and 0.64 s for N-terminal LAQ, FAS, IAA, and YAS, respectively (Fig. 2G).

Previous studies have shown that NAA-bound ClpS and UBR proteins also make contacts with the residues at position 2 (P2) and position 3 (P3) from the N terminus that influence binding affinity (*9*, *13*, *14*). These influences are reflected in the modulation of PD depending on the downstream P2 and P3 residues, as we observed above for LAA (1.21 s) compared with

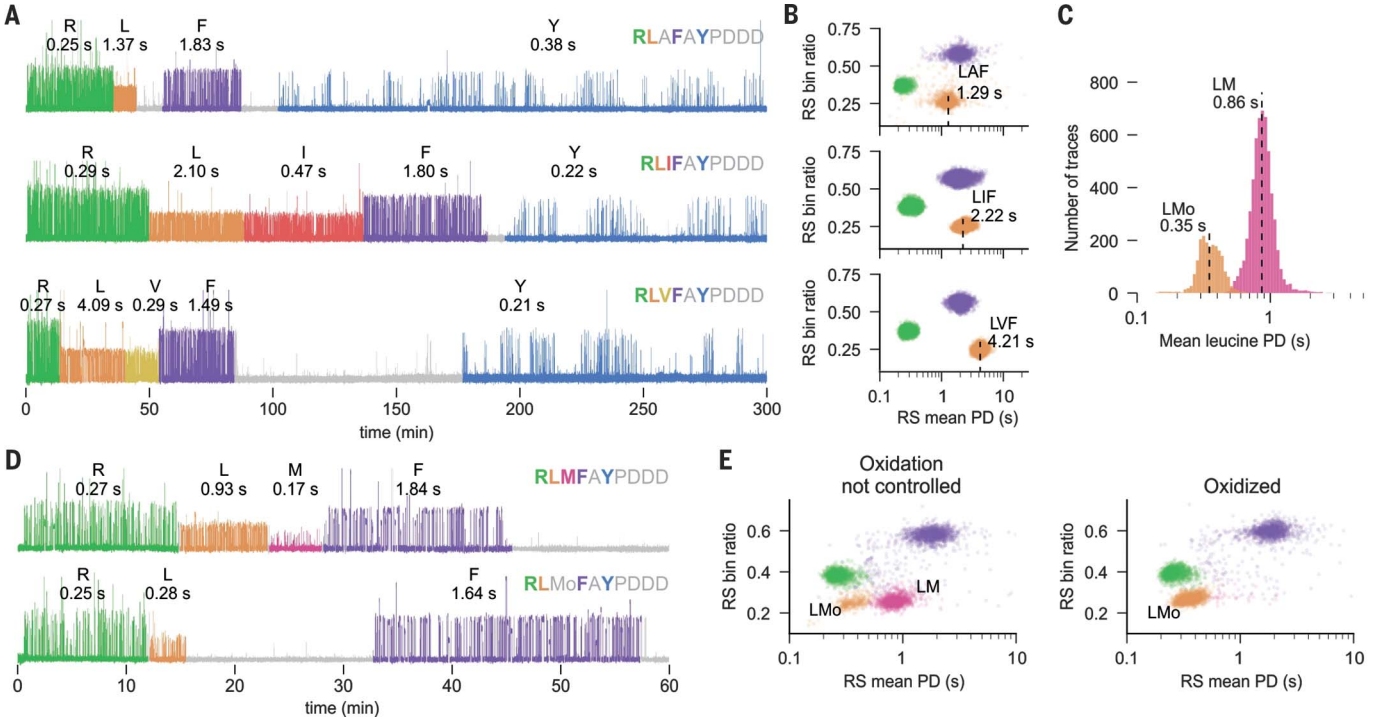


**Fig. 4. Detection of single amino acid changes and PTMs.** (**A** and **B**) Dynamic sequencing of synthetic peptides that differ by a single amino acid: RLAFAYPDDD (top), RLIFAYPDDD (middle), and RLVFAYPDDD (bottom). Example traces are shown in (A). Scatterplots of RS mean PD versus bin ratio are shown in (B). (**C** and **D**) Detection of oxidized methionine in the peptide RLMFAYPDDD. Distributions of mean PD for leucine are shown in (C); labels indicate populations with leucine followed by methionine (LM) or methionine sulfoxide (LMo). Shown in (D) are example traces in which methionine is recognized by PS961 and leucine exhibits a long PD (top) or in which methionine is not recognized, owing to oxidation, and in which leucine exhibits a short PD (bottom). (**E**) Scatterplots of RS mean PD versus bin ratio for runs in which oxidation was not controlled (left) or in which methionine was fully oxidized (right).

LAQ (2.70 s). We find that these influences on PD vary within informatically advantageous ranges and can be determined empirically or approximated in silico to model peptide sequencing behavior a priori (fig. S2, F to H). A powerful feature of this recognition behavior with respect to peptide identification is that each RS contains information about potential downstream P2 and P3 residues or PTMs, regardless of whether these positions are the targets of an NAA recognizer.

### Principles of dynamic protein sequencing illustrated with model peptides

To evaluate the kinetic principles of our dynamic sequencing method when applied to diverse sequences, we first characterized the synthetic peptide DQQRLIFAG (G, glycine), corresponding to a segment of human ubiquitin (Fig. 3, A to E). We performed sequencing reactions through a combination of three differentially labeled recognizers—PS610, PS961, and PS691—and two aminopeptidases—PhTET2 and PhTET3 (see materials and methods). The example trace in Fig. 3A starts with an NRS that corresponds to the time interval during which residues in the initial DQQ motif are present at the N terminus. The first RS starts at 120 min, upon exposure of N-terminal arginine to recognition by PS691. Subsequent cleavage events sequentially expose N-terminal leucine, isoleucine, and phenylalanine to their corresponding recognizers, with fast transitions (average <10 s) from one RS to the next. The transition from leucine to isoleucine recognition by PS961 is readily identified as a sharp change in average PD. This overall pattern is replicated across many instances of sequencing of the same peptide, with similar PD statistics across traces, as each peptide molecule follows the same reaction pathway over the course of the sequencing run (Fig. 3, B and C). Owing to the stochastic timing of cleavage events, each trace displays distinct start times and durations for each RS (Fig. 3C).

This approach reports the binding kinetics at each recognizable amino acid position and the kinetics of aminopeptidase cleavage along the peptide sequence. High-precision kinetic information on binding is obtained from a single trace because each RS typically contains tens to hundreds of on-off binding events, resulting in a distribution of PD and interpulse duration (IPD) measurements that can be analyzed statistically. The repetitive probing of each NAA also provides accurate recognizer calling because calls are not based on the error-prone detection of a single event associated with one fluorophore molecule (fig. S1F). Recognizer on-rate and concentration govern IPD for each RS; higher recognizer concentrations result in shorter average IPDs and faster rates of pulsing (fig. S3, A and B). Higher recognizer concentrations, however, increase the fluorescence background from freely diffusing recognizers, resulting in lower pulse signal-to-noise ratios, and can compete with aminopeptidases for N-terminal access. In practice, IPDs in the range of ~2 to 10 s provide a favorable balance among these factors.

The distribution of RS durations across an ensemble of replicate traces defines the rate of cleavage of each recognizable NAA. For DQQRLIFAG peptide, we observed average cleavage times of 30, 55, 40, and 88 min for N-terminal arginine, leucine, isoleucine, and phenylalanine, respectively, with approximate single-exponential decay statistics for each position (Fig. 3D and fig. S3C). The distribution of NRS durations reports the cleavage rate of a run of one or more nonrecognized NAAs. The average NRS duration for the initial DQQ motif was 155 min (Fig. 3D). Average cleavage rates are a key parameter and are controlled by the aminopeptidase concentration in the assay (fig. S3, D and E). Given the exponential behavior, we target average RS durations of 10 to 40 min

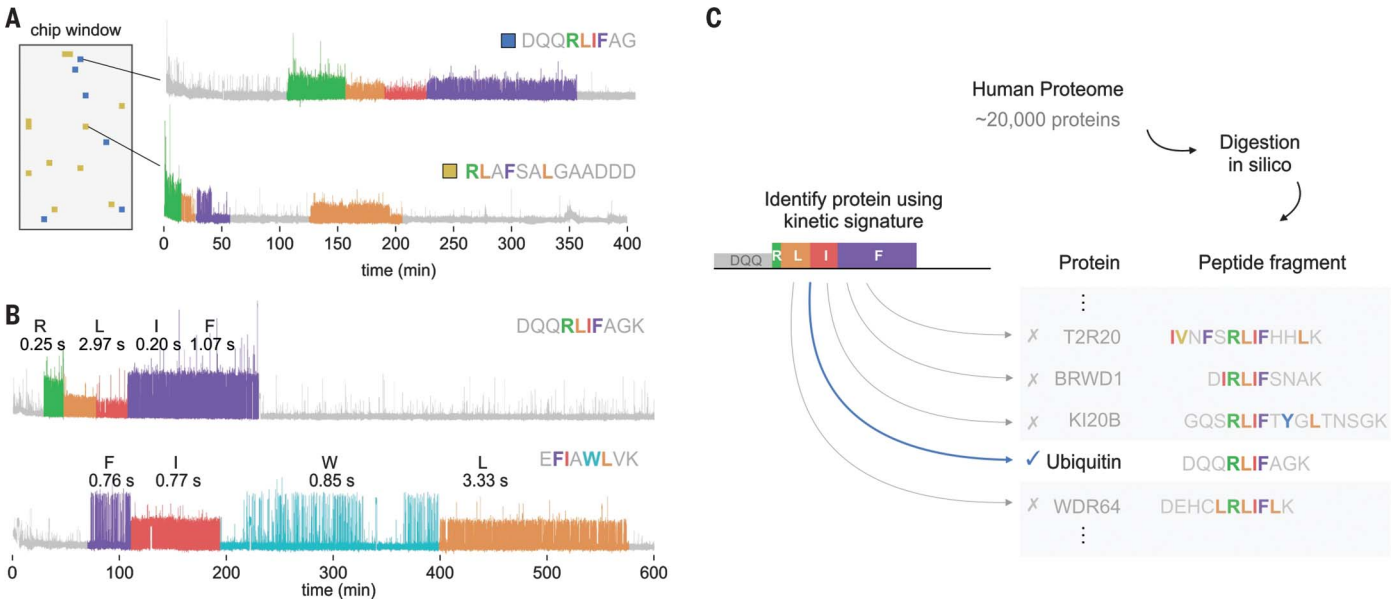


**Fig. 5. Discrimination of peptides in mixtures and mapping peptides to the human proteome.** (**A**) Example traces from sequencing a mixture of the peptides DQQRLIFAG and RLAFSALGAADDD on the same chip; the chip window indicates the location of reaction chambers producing a sequencing readout for each peptide. (**B**) Example traces from the dynamic sequencing of two peptides, DQQRLIFAGK (top) and EFIAWLVK (bottom), isolated from the recombinant human proteins ubiquitin and GLP-1, respectively. (**C**) Diagram illustrating the identification of the protein ubiquitin as a match to the kinetic signature from DQQRLIFAGK peptide in an in silico digest of the human proteome based on kinetic information.

to provide sufficient time for pulsing data collection, avoid missed RSs due to rapid cleavage, and minimize excessively long RS durations. We found it helpful to visualize the sequencing profiles of peptides as kinetic signature plots—simplified trace-like representations of the time course of complete peptide sequencing containing the median PD for each RS and the average duration of each RS and NRS (Fig. 3E). These highly characteristic features provide a wealth of sequence-dependent information for mapping traces from peptides to their proteins of origin.

To demonstrate that this core methodology and its kinetic principles apply to a wide range of peptide sequences, we sequenced the synthetic peptides DQQIASSRLAASFAAQQYPDDD, RLAFSALGAADDD, and EFIAWLV (E, glutamate; P, proline)—a segment of human glucagon-like peptide–1 (GLP-1)—under the same sequencing conditions used for DQQRLIFAG (Fig. 3F). Each peptide generated a characteristic kinetic signature in accordance with its sequence (Fig. 3G). We obtained readouts as far as position 18 (the furthest recognizable amino acid) in the peptide DQQIASSRLAASFAAQQYPDDD, illustrating that the method is compatible with long peptides and capable of deep access to sequence information in peptides.

### Distinctive kinetic signatures from single amino acid changes and PTMs

To illustrate how the kinetic parameters acquired from sequencing are sensitive to changes in sequence composition, we performed sequencing with a set of three peptides—RLAFAYPDDD, RLIFAYPDDD, and RLVFAYPDDD—that differ only at a single position, located immediately downstream from the PS961 N-terminal target leucine (Fig. 4A). Each type of amino acid at this position had a distinct effect on the PD acquired during recognition of N-terminal leucine by PS961. We observed median PDs of 1.29, 2.22, and 4.21 s for LAF, LIF, and LVF, respectively (Fig. 4B). In addition to differences in PD for leucine, each peptide displayed a characteristic RS or NRS in the interval between leucine and phenylalanine recognition (Fig. 4A and fig. S4A). These results demonstrate the sensitivity of the sequencing readout to variation at a single position and illustrate that both directly recognized NAAs and adjacent residues can influence the full kinetic signature obtained from sequencing.

Because the aminoacyl-proline bond of the YP motif in peptides such as RLIFAYPDDD cannot be cleaved by the PhTET aminopeptidases (*11*, *12*), observation of YP pulsing at the end of a trace ensures that cleavage has progressed completely from the first to last recognizable amino acid. The sequencing output from RLIFAYPDDD, therefore, provided a convenient dataset for examining biochemical sources of nonideal behavior that could lead to errors in peptide identification. The main sources of incomplete information in traces were deletions of expected RSs due to the stochastic occurrence of rapid sequential cleavage events (fig. S4B) and early termination of reads resulting from photodamage or surface detachment (fig. S4C).

In addition to changes in amino acid sequence composition, sequencing readouts are sensitive to changes due to PTMs. As an example, we examined methionine oxidation. The thioether moiety of the methionine side chain is susceptible to oxidation during peptide synthesis and sequencing. We determined that PS961 binds a peptide with N-terminal methionine with a dissociation constant ($K_d$) of 947 ± 47 nM (fig. S4D) and hypothesized that oxidation, resulting in a polar methionine sulfoxide side chain, would eliminate binding and reduce NAA binding affinity when located at P2. We determined computationally that methionine sulfoxide is highly unfavorable in the PS961 NAA binding pocket and that nonpolar residues are preferred at P2 (fig. S4E and fig. S2H). We sequenced the synthetic peptide RLMFAYPDDD (M, methionine) and observed two populations of traces with distinct kinetic signatures—a first population containing leucine recognition with a median PD of 0.86 s and a second population with a median PD of 0.35 s (Fig. 4C). Traces from the first population also displayed methionine recognition with a short PD in the time interval between leucine and phenylalanine recognition (Fig. 4D). Methionine recognition was absent in traces from the second population (Fig. 4D), indicating that the methionine side chain in these peptides was not capable of recognition by PS961. When we fully oxidized methionine by preincubation with hydrogen peroxide (see materials and methods), we observed elimination of both methionine recognition and the leucine

recognition cluster with a long median PD, as expected (Fig. 4E). These results demonstrate the capability for extremely sensitive detection of PTMs owing to their kinetic effects on recognition.

### Sequencing peptide mixtures and mapping peptides derived from human proteins

Proteomics applications require identification of peptides in mixtures derived from biological sources. To extend our results to peptide mixtures and biologically derived peptides, we performed two experiments. First, we mixed DQQRLIFAG and RLAFSALGAADDD peptides, immobilized them on the same chip, and performed a sequencing run. Data analysis (see materials and methods) identified two populations of traces corresponding to each peptide, with kinetic signatures in close agreement with those identified in runs with individual peptides (Fig. 5A and fig. S4F). Second, to demonstrate that our method extends to biologically derived peptides, we performed sequencing runs with peptide libraries generated using a simple workflow from recombinant human ubiquitin (76 amino acids) and GLP-1 (37 amino acids) proteins digested with AspN/LysC and trypsin, respectively (see materials and methods). For both libraries, data analysis readily identified traces matching the expected recognition pattern for the protease cleavage products DQQRLIFAGK and EFIAWLVK (K, lysine) for ubiquitin and GLP-1, respectively, and produced kinetic signatures in agreement with synthetic versions of these peptides (Fig. 5B and fig. S4G). We identified matches to the kinetic signature of the ubiquitin peptide DQQRLIFAGK across the human proteome, taking advantage of simple sequence constraints provided by kinetic information (see materials and methods). We found only one protein other than ubiquitin that contained a peptide that could potentially match this signature (Fig. 5C); thus, even short signatures can exhibit proteome abundance of $<1$ in $10^4$ proteins. These results illustrate the potential of the full kinetic output from sequencing to enable digital mapping of peptides to their proteins of origin.

### Conclusions

Our simple, real-time dynamic approach differs markedly from other recently described single-molecule approaches that rely on complex, iterative methods involving stepwise Edman chemistry or hundreds of cycles of epitope probing (*15*–*17*). Nanopore approaches offer the potential for real-time readouts and simplicity but face substantial challenges related to the size and biophysical complexity of polypeptides (*18*–*20*). Our sequencing technology is readily expanded in its capabilities, and there are multiple areas for improvement. Expansion of proteome coverage can be achieved through directed evolution and engineering of recognizers. The NAA targets demonstrated here make up ~35.6% of the human proteome, but lower-affinity NAA targets require longer PDs to enable detection in all sequence contexts. Recognizers for new amino acids or PTMs can be evolved from current recognizers or identified in screens of other scaffolds, such as other types of NAA- or PTM-binding proteins or aptamers. Extension to detection of all 20 natural amino acids and multiple PTMs is feasible for de novo sequencing; however, partial sequences are sufficient for most proteomics applications, which rely on mapping to predefined sets of candidate proteins (*21*). Aminopeptidases can be engineered to optimize cleavage rates and minimize RS deletions from rapid sequential cleavage. We envision that the dynamic range of samples and the applications most suitable for the system will tend to scale with the number of reaction chambers on the chip and that compression of dynamic range will be necessary for certain applications. We anticipate that future developments of the platform will increase the accessibility of proteomics studies and enable discoveries in biological and clinical research.

### REFERENCES AND NOTES


1. S. Goodwin, J. D. McPherson, W. R. McCombie, *Nat. Rev. Genet.* **17**, 333–351 (2016).
2. W. Timp, G. Timp, *Sci. Adv.* **6**, eaax8978 (2020).
3. R. Aebersold *et al*., *Nat. Chem. Biol.* **14**, 206–214 (2018).
4. M. Beck *et al*., *Mol. Syst. Biol.* **7**, 549 (2011).
5. N. L. Anderson, N. G. Anderson, *Mol. Cell. Proteomics* **1**, 845–867 (2002).
6. Y. Liu, A. Beyer, R. Aebersold, *Cell* **165**, 535–550 (2016).
7. J. Tullman, N. Callahan, B. Ellington, Z. Kelman, J. P. Marino, *Appl. Microbiol. Biotechnol.* **103**, 2621–2633 (2019).
8. D. A. Dougan, B. G. Reid, A. L. Horwich, B. Bukau, *Mol. Cell* **9**, 673–683 (2002).
9. B. J. Stein, R. A. Grant, R. T. Sauer, T. A. Baker, *Structure* **24**, 232–242 (2016).
10. T. Tasaki, Y. T. Kwon, *Trends Biochem. Sci.* **32**, 520–528 (2007).
11. M. A. Durá *et al*., *Mol. Microbiol.* **72**, 26–40 (2009).
12. M. A. Durá *et al*., *Biochemistry* **44**, 3477–3486 (2005).
13. W. S. Choi *et al*., *Nat. Struct. Mol. Biol.* **17**, 1175–1181 (2010).
14. J. Muñoz-Escobar, E. Matta-Camacho, C. Cho, G. Kozlov, K. Gehring, *Structure* **25**, 719–729.e3 (2017).
15. J. Swaminathan *et al*., *Nat. Biotechnol.* **36**, 1076–1082 (2018).
16. J. D. Egertson *et al*., bioRxiv 2021.10.11.463967 [Preprint] (2021); https://doi.org/10.1101/2021.10.11.463967.
17. M. Chee, K. Gunderson, M. P. Weiner, Macromolecule analysis employing nucleic acid encoding (2019); https://pdfaiw.uspto.gov/.aiw?Docid=20190145982.
18. S. Zhang *et al*., *Nat. Chem.* **13**, 1192–1199 (2021).
19. H. Brinkerhoff, A. S. W. Kang, J. Liu, A. Aksimentiev, C. Dekker, *Science* **374**, 1509–1513 (2021).
20. J. A. Alfaro *et al*., *Nat. Methods* **18**, 604–617 (2021).
21. J. Swaminathan, A. A. Boulgakov, E. M. Marcotte, *PLOS Comput. Biol.* **11**, e1004080 (2015).
22. B. D. Reed, M. J. Meyer, J. F. Beltrán, Code and data for "Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device." Zenodo (2022); https://doi.org/10.5281/zenodo.7017750.



### ACKNOWLEDGMENTS

We thank E. Chen for helpful discussions in preparing the manuscript. **Funding:** This work was funded by Quantum-Si, Inc. **Author contributions:** A.D.G., A.L., A.M.v.O., A.P.-K., B.D.R., B.S., C.L., D.H.P., E.M., F.L., G.P., H.H., J.A.B., J.P., J.T., J.Z., K.B., K.F.C., M.Mi., M.M.E., M.P., M.R.-C., M.W.B., N.D.B., N.S., O.A., R.B., R.N., S.Ha., S.G., S.S.P., T.D.C., X.S., and Y.-C.W. developed methods and reagents for protein sequencing; A. Bu., D. Be., D.F., G.A., G.K., J.F.B., M.D., M.J.M., S.F.S., S.L., V.A., and Z.Z. developed software; and A. Bel., A. Bet., A.K., B.L., C.C., D.Bo., E.A.G.W., F.R.A., G.S., J.B., J.C., J.D., J.S., K.P., K.R., L.H., M.B., M.F., P.A., P.L., P.T., R.Cl., R.Cu., S.Ho., S.C., T.C., T.R., T.R.T., X.M., X.W., and Z.H. developed semiconductor chips, nanophotonics, lasers, and instruments. J.M.R., M.Mc., T.R., B.D.R., M.D., G.S., M.F., P.L., and M.D.D. supervised and/or acquired resources and funding. B.D.R., M.J.M., M.P., and B.S. designed experiments and/or generated the single-molecule recognition and sequencing data presented in the figures. B.D.R., M.J.M., and J.F.B. analyzed sequencing data and prepared the figures. M.P. and G.P. characterized recognizer ensemble kinetic properties. O.A. generated libraries from recombinant proteins. H.H. and Y.-C.W. generated model peptides. D.H.P. performed computational modeling. B.D.R. led the study and wrote the manuscript with review and commentary from coauthors. All authors met the criteria for authorship and contributed critically to the development of the sequencing method and platform by conducting experiments, developing methods and concepts, analyzing and interpreting data, developing software, supervising research, or acquiring funding. **Competing interests:** All authors affiliated with Quantum-Si, Inc., along with A.D.G. and A.M.v.O., are shareholders and/or are listed as inventors on patents owned by Quantum-Si, Inc.; A.D.G. and A.M.v.O. are on the scientific advisory board of Quantum-Si, Inc., and are paid consultants. The technology presented in this paper is the subject of numerous pending or awarded patents filed by Quantum-Si, Inc., with the US Patent and Trademark Office and international offices. **Data and materials availability:** Data and custom code used in this paper are available for download online at Zenodo (*22*). Sequencing reagents and instruments are available from Quantum-Si, Inc., under a material transfer agreement.



### SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abo7651
Materials and Methods
Figs. S1 to S4
Table S1
References (*23*–*27*)

View/request a protocol for this paper from *Bio-protocol*.

Submitted 28 February 2022; accepted 13 September 2022
10.1126/science.abo7651

# Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device

Brian D. Reed, Michael J. Meyer, Valentin Abramzon, Omer Ad, PhD, Omer Ad, Pat Adcock, Faisal R. Ahmad, Gn Alppay, James A. Ball, James Beach, Dominique Belhachemi, Anthony Bellofiore, Michael Bellos, Juan Felipe Beltrn, Andrew Betts, Mohammad Wadud Bhuiya, Kristin Blacklock, Robert Boer, David Boisvert, Norman D. Brault, Aaron Buxbaum, Steve Caprio, Changhoon Choi, Thomas D. Christian, Robert Clancy, Joseph Clark, Thomas Connolly, Kathren Fink Croce, Richard Cullen, Mel Davey, Jack Davidson, Mohamed M. Elshenawy, Michael Ferrigno, Daniel Frier, Saketh Gudipati, Stephanie Hamill, Zhaoyu He, Sharath Hosali, Haidong Huang, Le Huang, Ali Kabiri, Gennadiy Kriger, Brittany Lathrop, An Li, Peter Lim, Stephen Liu, Feixiang Luo, Caixia Lv, Xiaoxiao Ma, Evan McCormack, Michele Millham, Roger Nani, Manjula Pandey, John Parillo, Gayatri Patel, Douglas H. Pike, Kyle Preston, Adeline Pichard-Kostuch, Kyle Rearick, Todd Rearick, Marco Ribezzi-Crivellari, Gerard Schmid, Jonathan Schultz, Xinghua Shi, Badri Singh, Nikita Srivastava, Shannon F. Stewman, TR Thurston, T. R. Thurston, Philip Trioli, Jennifer Tullman, Xin Wang, Yen-Chih Wang, Eric A. G. Webster, Zhizhuo Zhang, Jorge Zuniga, Smita S. Patel, Andrew D. Griffiths, Antoine M. van Oijen, Michael McKenna, Matthew D. Dyer, and Jonathan M. Rothberg

*Science*, **378** (6616), .
DOI: 10.1126/science.abo7651

**Single-molecule reading of proteins**

Modern DNA-sequencing methods can interrogate single molecules in extremely high throughput, but protein sequencing typically uses ensemble techniques and requires larger amounts of relatively pure material. Reed *et al*. generated a set of labeled proteins that recognize the first few amino acids at the N terminus of a peptide immobilized on an optical chip. Transient binding yields spectral signals and association and dissociation rates that can be used to identify the terminal amino acid. Multiple amino acids on a single molecule can then be read by adding a protease that gradually reveals the next amino acid at the terminus. —MAF