


Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures

Jagannath Swaminathan^{1,5}, Alexander A Boulgakov^{1,5}, Erik T Hernandez^{2,5}, Angela M Bardo^{1,5} , James L Bachman², Joseph Marotta^{1,4}, Amber M Johnson², Eric V Anslyn² & Edward M Marcotte^{1,3} 

The identification and quantification of proteins lags behind DNA-sequencing methods in scale, sensitivity, and dynamic range. Here, we show that sparse amino acid–sequence information can be obtained for individual protein molecules for thousands to millions of molecules in parallel. We demonstrate selective fluorescence labeling of cysteine and lysine residues in peptide samples, immobilization of labeled peptides on a glass surface, and imaging by total internal reflection microscopy to monitor decreases in each molecule’s fluorescence after consecutive rounds of Edman degradation. The obtained sparse fluorescent sequence of each molecule was then assigned to its parent protein in a reference database. We tested the method on synthetic and naturally derived peptide molecules in zeptomole-scale quantities. We also fluorescently labeled phosphoserines and achieved single-molecule positional readout of the phosphorylated sites. We measured >93% efficiencies for dye labeling, survival, and cleavage; further improvements should enable studies of increasingly complex proteomic mixtures, with the high sensitivity and digital quantification offered by single-molecule sequencing.

Proteins often exist in extremely complex mixtures; for example, a typical human cell contains >10,000 unique proteins and perhaps ten times as many post-translationally modified proteoforms. Each protein potentially varies in abundance from 1 to 10⁹ copies, in a manner often poorly predicted by mRNA-transcript levels¹. The inability to comprehensively sequence such complex protein samples, and especially to quantify and identify proteins having low abundance or post-translational modifications, is a major roadblock in protein-biomarker discovery². Currently, mass spectrometry is the method of choice for large-scale protein identification, but it is limited in its ability to analyze low-abundance samples and to map rare amino acid variants^{3–5}. These limitations could be addressed by successful development of highly parallel single-molecule protein sequencing^{6–12}, a concept analogous to nucleic acid technologies that sequence millions to billions of oligonucleotides in complex mixtures in parallel. The approach would offer an improvement of more than one million-fold in sensitivity over conventional technologies and would allow for millions of distinct peptide molecules to be sequenced in parallel, identified, and digitally quantified (**Fig. 1a**). Here, we describe an implementation of protein fluorosequencing by directly visualizing individual fluorescently labeled peptide or protein molecules as they are subjected to the classic protein sequencing chemistry, Edman degradation¹³.

In the protein fluorosequencing concept, one or more amino acid types are selectively labeled with a specific identifier fluorophore¹⁴. After the immobilization of millions of labeled peptides on a glass

coverslip, each molecule’s fluorescence is monitored through total internal reflection fluorescence (TIRF) microscopy after consecutive rounds of N-terminal amino acid removal through Edman chemistry¹³ (**Fig. 1b**). The sequence positions of the labeled amino acids are identified for each peptide molecule, thus providing a partial sequence. These sequences of fluorescent amino acids are compared against a reference proteome for assignment to their proteins of origin. Although only labeled amino acids are visualized, the results can nonetheless be very rich in information, because the sequence positions of the labeled amino acids are precisely determined, the identities of the terminal amino acids can be constrained by the choice of proteolytic enzyme and surface-attachment chemistry, and the identities of the intervening amino acids are partly constrained (because they were not the labeled, cleaved, or attached amino acid types). The information richness of the readout is illustrated in a plot of the proportions of human proteins in an assortment of subcellular compartments that can be uniquely identified by using only a two-color code (**Fig. 1c**, modeling the labeling of cysteines and lysines on peptides generated by proteolysis after glutamate or aspartate). Even a two-color code can be sufficient to uniquely identify most proteins in mixtures of moderate complexity comprising as many as ~1,000 human proteins (**Fig. 1c** and refs. 6,7). Monte Carlo simulations predict that the use of additional labels (for example, as established for aspartate/glutamate and tryptophan¹⁴) should be sufficient to identify most proteins in the human proteome, even despite the expected effects of experimental errors due to, for example, photo/chemical dye

¹Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, USA. ²Department of Chemistry, University of Texas at Austin, Austin, Texas, USA. ³Department of Molecular Biosciences, University of Texas at Austin, Austin, Texas, USA. ⁴Present address: Luminex Corporation, Austin, Texas, USA. ⁵These authors contributed equally to this work. Correspondence should be addressed to E.M.M. (marcotte@icmb.utexas.edu) or E.V.A. (anslyn@austin.utexas.edu).

Received 5 September 2017; accepted 21 September 2018; published online 22 October 2018; doi:10.1038/nbt.4278

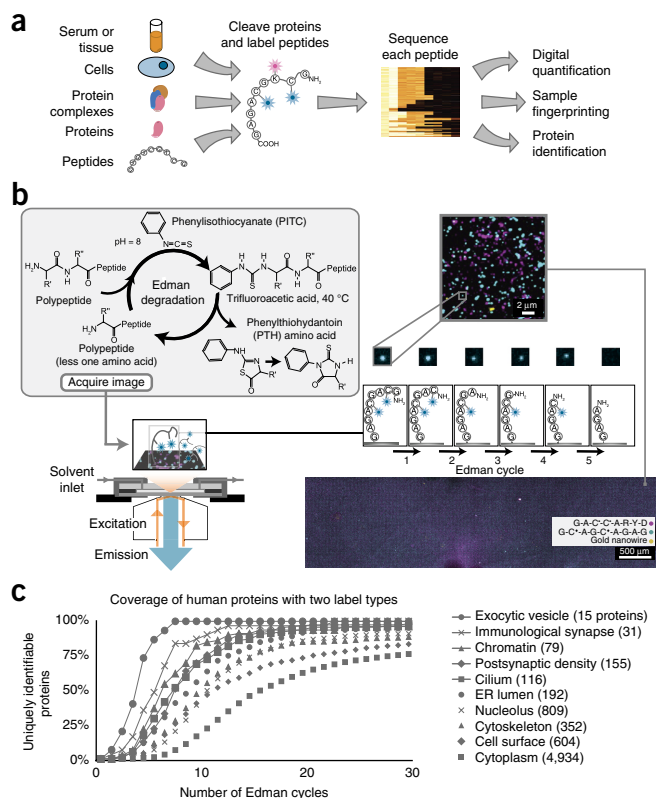


Figure 1 Overview of single-molecule fluorosequencing. **(a)** Summary of the approach for protein and peptide analyses. **(b)** Peptides are covalently labeled with amino acid-specific fluorescent dyes and immobilized in a TIRF single-molecule-microscope stage perfusion chamber. Through TIRF, each peptide is imaged, and its N-terminal amino acid is chemically removed via Edman degradation, thus leaving each peptide one amino acid shorter and regenerating its free N terminus. Repeated cycles of chemistry (each removing one amino acid) and imaging reveal the positions of fluorescent dyes within each molecule. Millions of individual peptide molecules can be analyzed in parallel at reasonable attachment densities, shown for approximately 3 million peptides in an approximately 1.3 × 5 mm area of the coverslip. Bullet indicates TMR conjugated to cysteine; diamond indicates Atto647N conjugated to cysteine; gold nanowires serve as fiducial markers. **(c)** Even a relatively modest amino acid-labeling scheme can be sufficiently information rich to identify proteins, as illustrated by a calculation of the proportions of human proteins in specific subcellular compartments (defined by Gene Ontology Cellular Component annotations; numbers indicate protein counts) that are uniquely identifiable with a two-color code. Each curve plots coverage of uniquely identifiable proteins as a function of read length (Edman cycles performed), considering the scenario of labeling only cysteines and lysines on peptides formed by GluC proteolysis, which cleaves after glutamate or aspartate. ER, endoplasmic reticulum.

inactivation, incomplete fluorescence labeling, and sporadic failures of Edman reaction cycles⁶.

Here, we experimentally implemented the fluorosequencing concept by labeling and discriminating peptides and simple peptide mixtures, a process that required developing instrumentation and methods; extensive testing of fluorophores, microfluidic design, chemistry of peptide immobilization and Edman degradation; creating image-processing algorithms for monitoring individual peptides' fluorescence intensity; and classifying and modeling the sources of errors. We analyzed samples of increasing complexity, from singly labeled peptide samples to peptides labeled at as many as three positions, both individually and in

simple peptide mixtures, and we distinguished specific phosphoserine post-translational modifications by sequencing.

RESULTS

Instrumentation for single-molecule fluorescent-peptide imaging and Edman sequencing

Because Edman sequencing uses harsher reagents than conventional aqueous microscopy experiments (including strong organic acids, bases, solvents, and heat), we identified fluorescent dyes able to survive the chemistry (**Supplementary Fig. 1**); adapted a microscope-stage perfusion chamber with chemically resistant tubing, connectors, and perfluoroelastomer gaskets (**Supplementary Fig. 2a,b**); and automated chemical manipulations within the chamber by using computer-controlled pumps and valves to exchange reagents under nitrogen (**Supplementary Fig. 2c**). Tests of bulk fluorescent peptides on beads confirmed that the dyes did not strongly affect Edman degradation (**Supplementary Fig. 3**). We next confirmed that the fluorescent peptides could be covalently tethered via aminosilane to a glass coverslip and survive extended imaging (**Supplementary Fig. 4**), exposure to Edman solvents, and heat, without a significant loss of fluorescence (**Supplementary Fig. 5**). Covalently tethered gold nanowires additionally provided unique constellations of fiducial markers in each field of view (**Fig. 1b**, showing ~3 million peptides in an approximately 1.3 mm × 5 mm area of coverslip). Thus, even reasonably sparse peptide densities allow for millions of individual peptide molecules to be imaged in the apparatus, and the immobilized peptides and dyes survive the necessary reagents.

Identifying positions of single labels within peptide molecules

To demonstrate that consecutive cycles of Edman chemistry could be performed on peptides with high efficiency in the apparatus, we considered a series of experiments with control peptides of increasing sample and label complexity. To interpret these experiments, we developed custom image-processing algorithms (**Supplementary Figs. 6 and 7**). These algorithms (i) identified individual fluorescent molecules within each micrograph, (ii) aligned fluorescent peaks from the same field of view, imaged across consecutive Edman cycles, by using fiducial markers to correct for microscope-stage variation, and then (iii) identified peptides whose fluorescence signals were stable and successfully removed by the final Edman cycle, thus computationally flagging contaminating fluorescent objects and nonsequenced peptides.

We first compared a uniform population of copies of the peptide GK[†]AGAG (where † indicates the fluorophore Atto647N covalently coupled via *N*-hydroxysuccinimide (NHS) ester to the lysine side chain) to a second uniform population of that peptide blocked from sequencing by *N*-terminal acetylation, which served as a negative control (**Fig. 2a**). We performed several cycles of Edman chemistry with all reagents and incubation steps but omitting the key reagent, phenylisothiocyanate (PITC). Dyes disappearing during these 'mock' Edman cycles allowed us to estimate background dye-loss rates of approximately 7% per cycle, from a combination of photobleaching (**Supplementary Fig. 4**), chemical destruction, and loss of noncovalently bound molecules. Subsequent Edman cycles incorporating PITC confirmed that peptides most frequently lost dyes at the expected second cycle, in contrast to blocked negative-control peptides, thus demonstrating successful identification of the dye position for 98,945 individual molecules (out of 238,503 molecules analyzed over three replicate experiments, each imaging 100 image fields), in a manner requiring a free peptide N terminus.

We further confirmed the apparatus and chemistry by analyzing a control mixture of many copies of two peptides distinguishable

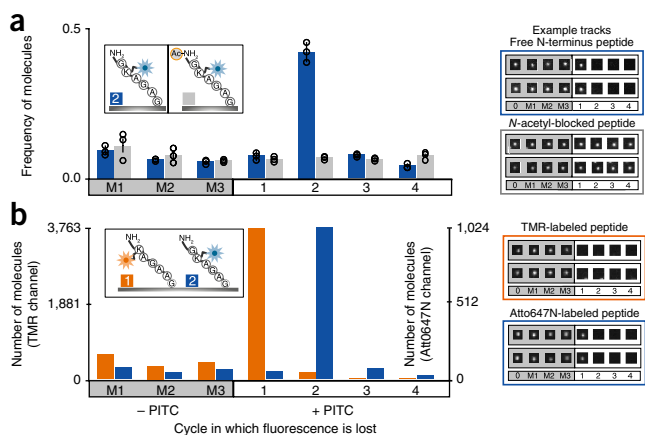


Figure 2 Fluorescent amino acid positions can be determined at single-molecule sensitivity. **(a)** Sequencing of individual peptide molecules requires a free N terminus, as shown by comparing fluorescent sequences of the hexamer GK[†]AGAG and its nonsequenceable N-terminally acetylated version. The histogram at left plots the relative frequencies of peptide molecules exhibiting dye loss at each Edman cycle (mean \pm s.d. of 3 technical replicates; $n = 59,434$, $80,541$, or $98,528$ molecules measured across 100 image fields each). M1, M2, and M3 denote negative-control (mock) Edman cycles in which PITC was omitted. Individual traces are illustrated at right, with extracted TIRF images for four individual molecules (two blocked and two unblocked) across cycles. **(b)** The sequence positions of dye-labeled amino acids can be accurately determined within individual peptide molecules, as shown by deconvolution of a mixture of two control peptides differing in both label position and color. The histogram displays counts of individual molecules of each color, K^{*}AGAAG and GK[†]AGAG (where [†] indicates Atto647 conjugated to lysine, and * indicates TMR conjugated to lysine) ($n = 5,683$ and $n = 1,598$ molecules measured across 20 fields), indicating the cycle numbers at which the dyes are removed. Example single-molecule TIRF images are shown at right. Fluorescence-intensity measurements are provided for all single-molecule image tracks in **Supplementary Data 1**.

both by fluorophore color and by sequence position, with one set labeled by the red-emitting dye tetramethylrhodamine (TMR) at position 1 and the other set labeled by far-red Atto647N at position 2. We determined the positions of dye loss for approximately 5,000 individual peptide molecules across nine cycles of chemistry (three mock-Edman and six complete Edman cycles). The predominant patterns observed were PITC dependent and matched the expected positions for each dye (**Fig. 2b**). Similarly to the results in **Figure 2a**, we observed a low background rate of dye loss per cycle, which was consistent with nonspecific, PITC-independent dye destruction. Because each fluorescence channel independently reports on a different dye, the sequence positions of multiple amino acid types on a single peptide can be determined by labeling each type with a different fluorophore (as in **Supplementary Fig. 8**). Overall, the efficiencies of Edman degradation, dye attachment, detection, and stability, as well as peptide surface-attachment chemistry, all appeared sufficiently robust to support fluorosequencing.

Determining the precise amino acid positions of dyes within multiply labeled peptide molecules

Determining the positions of multiple dyes within one peptide requires accurately determining which Edman cycles elicit stepwise intensity decreases in that molecule's fluorescence; each step corresponds to the removal of one or more dye molecules. We demonstrated this key requirement by determining the positions of two

labeled cysteine amino acids within many identical copies of peptide GC[♦]AGC[♦]AGAG (where [♦] indicates Atto647N coupled by iodoacetamide to cysteine). For each copy of GC[♦]AGC[♦]AGAG, we expected losses of the fluorescent cysteines after the second and fifth Edman cycles (**Fig. 3a**).

Indeed, monitoring an individual peptide molecule (**Fig. 3b**) and measuring its fluorescence after every Edman cycle revealed clear stepwise decreases in its intensity after the second and fifth cycles (**Fig. 3c**, orange diamonds). We collated such intensity patterns for all 1,695 individual doubly labeled peptide molecules (**Fig. 3d**) and observed that the largest proportion of the peptide tracks (675 molecules) had distinct decreases in intensity after the second and fifth Edman cycles (**Fig. 3c**, box plots). Thus, by noting which Edman cycle elicited a stepwise intensity decrease for a peptide molecule, we were able to correctly localize the two cysteine-coupled dyes within each individual molecule, in a manner sufficient to infer the sequence xCxxCxxx (where C represents cysteine, and x represents any amino acid except C).

To better interpret data for other dye positions and counts, we empirically determined single-peptide-molecule fluorescence-intensity distributions, then used these empirical distributions as the basis for a maximum-likelihood statistical model for assigning the most probable dye positions to an observed peptide-intensity track (**Supplementary Fig. 9** and Online Methods). We found it useful to summarize these sequence assignments across a population of molecules by representing them as a heat map of counts of peptides with given dye positions. Such heat maps allowed us to quickly determine the most prevalent sequences and to assess systematic errors. We plotted a histogram corresponding to the GC[♦]AGC[♦]AGAG experiment described above (**Fig. 4a**). Notably, 675 molecules (also presented in **Fig. 3d**) were correctly determined to have dyes at the expected 2 and 5 positions, corresponding to the peaks of the doubly labeled sequences (illustrated schematically).

In parallel, to isolate and quantify specific sources of sequencing error, we tested N-terminally acetylated versions of the peptide. The histogram (**Fig. 4b**) arising from background dye/molecule-loss rates established an empirical baseline for correcting the observed sequence frequencies for losses by chance and allowed us to calculate the signal relative to the expected background (**Fig. 4b**).

Characterizing and modeling errors

Although we observed the correct sequence in approximately 40% of the examples (**Figs. 3** and **4**), the results were accompanied by certain previously expected systematic errors⁶. These errors arose from defective dyes or failed dye attachment (collectively referred to as 'dud dyes'), molecule-by-molecule failures of Edman chemistry resulting in missed cleavage events, position-independent background rates of dye or molecule loss in each cycle (directly quantified by N-acetylated peptide controls), and assignment errors in computing dye positions from the observed fluorescent sequences. Each type of error introduces a distinct bias into the sequencing histogram (**Fig. 4a**, inset), thus allowing us to estimate error rates by comparing our observed signal to that obtained from Monte Carlo simulations of sequencing with errors. Simulations of the GC[♦]AGC[♦]AGAG sequencing experiment (**Supplementary Fig. 10**) agreed well with observed sequences with low residuals (35%), thus confirming high rates of Edman cleavage (94% efficiency per cycle), 95% of dyes surviving per cycle, and molecular surface retention of 95%, with the largest error arising from dud dyes (7%). Because we chromatographically purified doubly labeled peptides and verified their labels by mass spectrometry before analysis (**Supplementary Fig. 11**), this effect was attributable

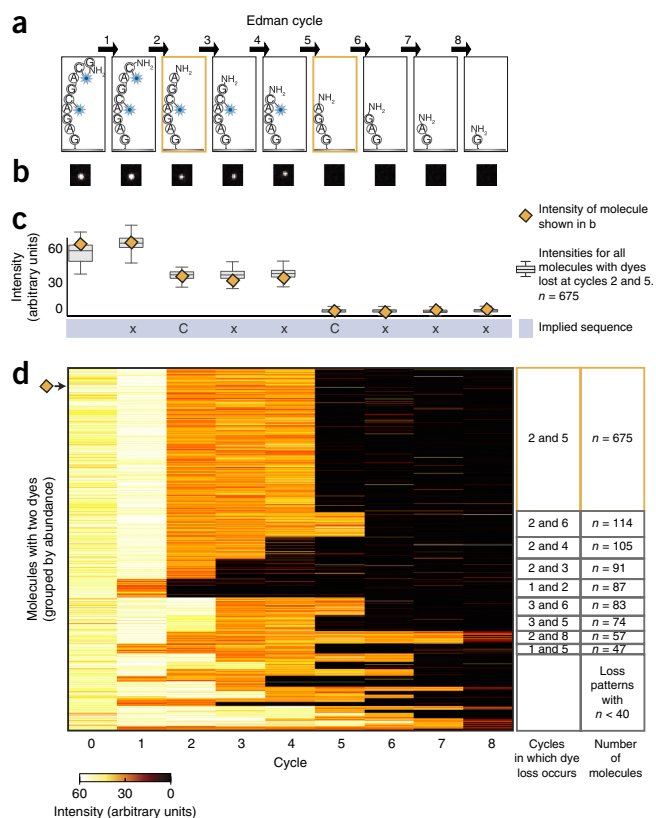


Figure 3 Stepwise decreases in fluorescence intensity occur at the Edman cycles that correspond to the removal of the dye-labeled amino acids. (a) Schematic of the peptide molecule, GC*AGC*AGAG (where ♦ indicates Atto647N conjugated to cysteine), losing dye-labeled amino acids at the second and fifth Edman degradation cycles. (b) The decrease in fluorescence intensity accompanying dye loss, illustrated in a representative set of TIRF images for a single peptide molecule. (c) Intensities for the representative molecule shown in (b) (orange diamonds) and a box plot of intensities (center line, median; limits, 75% and 25%; whiskers, ± 1.5 interquartile range) for all 675 molecules collected across all 49 images correctly identified as having decreases at amino acid positions 2 and 5. By noting the Edman cycle corresponding to the stepwise intensity decrease, the partial sequence of the peptide (xCxxCxxx) can be inferred. (d) The heat map of fluorescence-intensity values for each of the 1,695 peptides with two dyes, observed after every Edman cycle, showing that the predominant pattern corresponds to dye losses after the second and fifth cycles ($n = 675$ peptide molecules).

to correctly coupled dyes that did not fluoresce. A survey of multiple dyes and manufacturer batches revealed this outcome to be a feature of several commercial dyes, thus indicating a clear need for future improvement. Finally, to confirm that these experimental conditions allowed for high rates of Edman cleavage independently of amino acid composition, we studied peptides containing proline, which have been historically characterized by lower Edman cleavage rates¹⁵. We observed only a modest decrease in cleavage efficiency to 91%, as compared with 95% for alanine and 97% for repetitive glycine/alanine residues (Supplementary Fig. 12).

Deconvolution of peptide mixtures into groups of individual molecules

We next performed experiments on peptides in simple mixtures and from naturally occurring proteins, and demonstrated the identification

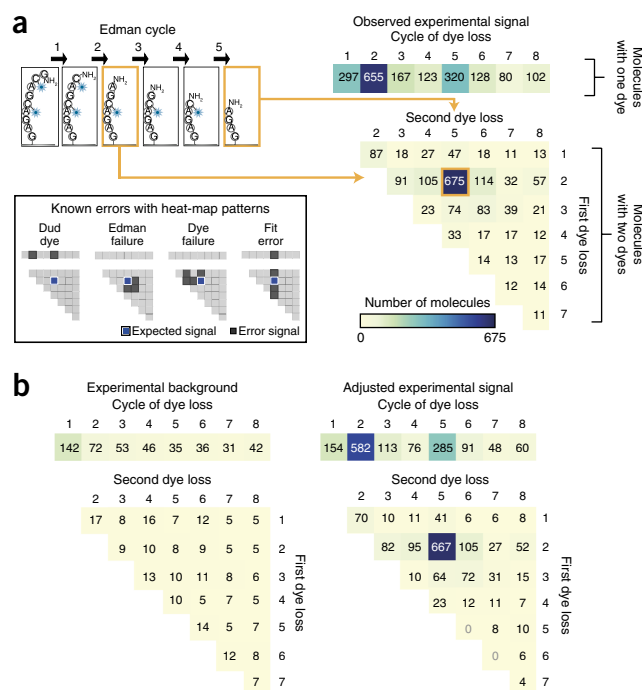
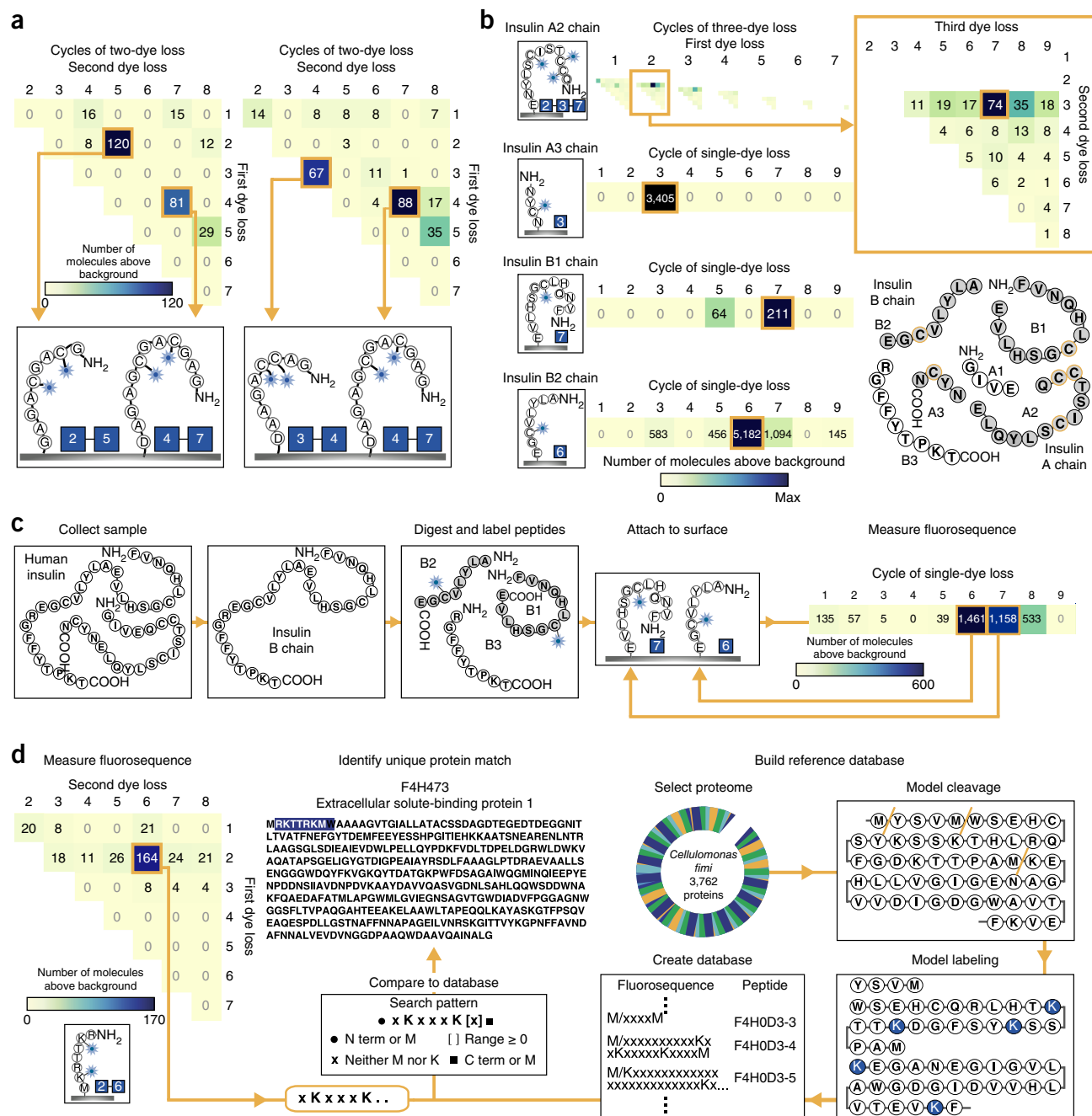


Figure 4 Fluorescent sequences can be interpreted computationally to identify dye positions and quantify errors. A maximum-likelihood statistical model allows for correct sequencing of multilabeled peptides. (a) Histograms of the fit fluorescent sequences obtained for GC*AGC*AGAG (where ♦ indicates Atto647N conjugated to cysteine) (right, based on 49 image fields). We summarized dye-loss positions for peptides with only one detectable dye as a 1D histogram (top-right histogram) and dye-loss positions for doubly labeled peptides as a 2D histogram (bottom-right histogram). As an aid for interpreting the 2D histogram, the example at top left shows a schematic of a peptide exhibiting dye losses at the second and fifth cycles, which correspond to the second row and fifth column of the 2D histogram; 675 peptide molecules exhibited this pattern. In this experiment, all other patterns correspond to specific sequencing errors, as illustrated graphically in the inset at left. (b) Through sequencing of an N-terminally acetylated population over 49 image fields of the same sequence, background observations expected from non-Edman events were determined (left), and the foreground counts were adjusted to determine the signal above background (right; calculation in Online Methods).

of a protein from a database. In addition, because all acetylated-control experiments exhibited similar sequencing patterns (because they simply lost dyes at background rates, owing to chemical destruction and other factors), we collected all acetylated experiments with each given dye count to obtain general background distributions that we could subtract from any experimental histogram to better estimate the sequenced label positions (Online Methods).

We performed zeptomole-scale experiments on two mixtures of peptide pairs (Fig. 5a and Supplementary Fig. 13a), thus clearly distinguishing several hundred peptide molecules labeled on cysteines at the (2,5), (4,7), and (3,4) positions. To demonstrate identification of peptides derived from a natural human protein, we also synthesized peptides corresponding to GluC protease-digested insulin fragments fluorescently labeled on cysteine residues. The major peak in each histogram corresponded to the correct dye-labeled amino acid positions; thus, this method correctly detected the specific labeling patterns of two singly labeled insulin B-chain fragments, one singly labeled A-chain fragment, and one triply labeled A-chain fragment (Fig. 5b and Supplementary Fig. 14).



Notably, the fluorescent sequences of these four peptides, when considered together, were sufficient to uniquely identify insulin in the human proteome. We obtained equivalent results for biologically derived insulin (Fig. 5c), by using a peptide mixture obtained after GluC protease digestion of recombinant insulin B chain.

Protein identification

To illustrate how a single experimentally determined partial sequence might be used to identify a parent protein from a reference proteome, we studied the peptide RK[†]TTRK[†]M (where † indicates Atto647N coupled to lysine residues) from the bacterium *Cellulomonas fimi*, modeling a scenario in which peptides are generated by cyanogen bromide proteolysis, which cleaves proteins after methionines, and subsequent fluorescence labeling of lysine residues (Fig. 5d. and Supplementary Fig. 13b). The observed partial sequence XKXXXX[X]_{≥0} (where K indicates lysine, and X represents any amino acid except lysine or methionine), when constrained by knowledge of the proteolysis cleavage specificity (i.e., adjacent to a methionine or protein terminus), was found to occur only once in a database of all 3,762 proteins from the bacterium *C. fimi* (strain ATCC 484), thus uniquely identifying the protein as extracellular solute-binding protein family 1 (UniProt database identifier F4H473_CELFA). Thus, even for simple labeling schemes, there exist peptides for which partial sequencing is sufficient to uniquely identify their parent protein from a reference proteome. In practice, the identification of proteins in a reference database will be limited by sequencing errors. A computational model incorporating our experimentally determined error rates (Supplementary Fig. 15) suggests that the technique is currently sufficiently empowered to discriminate proteins in samples of tens to hundreds. Incorporating additional labels or information-rich constraints from proteolysis or attachment specificity should increase the power of this approach.

Single-molecule sequencing of serine phosphorylation sites

We demonstrated identification of the specific amino acid positions of phosphoserine residues at single-molecule sensitivity. We considered the peptide YSPTSPSK, found in high-copy tandem repeats within the C-terminal domain of RNA polymerase II and whose phosphorylation patterns on Ser2 and Ser5 have been implicated in transcriptional regulation¹⁶. To selectively label serine or threonine phosphorylation sites, we used a scheme (Fig. 6a) consisting of beta-elimination followed by conjugate addition via thiols¹⁷ to substitute thiol-linked fluorophores in place of phosphates. Analysis of the peptides YpS[°]PTSPSK and YSPTp[°]PSK (where ° indicates Atto647N coupled at phosphoserine residues) clearly discriminated serine phosphorylation sites within three amino acids of one another at single-molecule sensitivity (Fig. 6b).

DISCUSSION

Single-molecule protein sequencing combines aspects of DNA sequencing, mass spectrometry proteomics, and classic Edman sequencing; thus, comparing this method to its constituent technologies is useful to obtain a sense of its likely scalability, limits of dynamic range, applications, and other properties. Broadly, the approach shares upstream protein isolation and proteolysis with shotgun mass spectrometry, as well as computational matching of peptide-sequence-dependent patterns (fluorescent sequences versus spectra) to a reference proteome database, and combining evidence from peptide identifications into protein identifications. Thus, single-molecule protein sequencing should be able to take advantage of established protocols for these procedures. However, because the sensitivity of the approach is inherently single molecule, in contrast to the attomole-to-femtomole (10⁶–10⁹

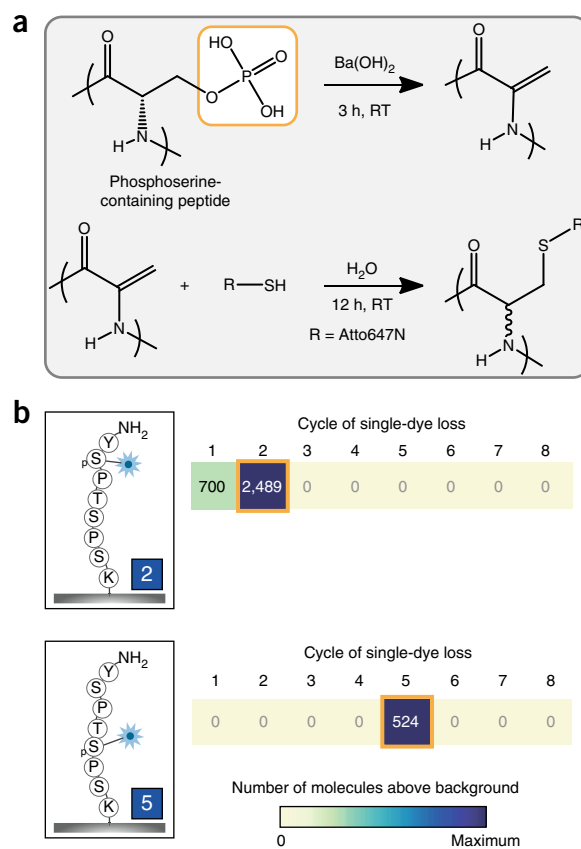


Figure 6 Direct single-molecule sequencing of phosphoserine positions within RNA polymerase II C-terminal-domain repeat peptides.

(a, b) Phosphorylated serines or threonines can be specifically labeled with fluorescent dyes by beta-elimination and conjugate addition¹⁷, then sequenced to determine the amino acid positions of the phosphorylated residues within each molecule (a), as demonstrated in b for C-terminal-domain repeat peptides phosphorylated (p) at either Ser2 (top) or Ser5 (bottom). Histograms in a and b each report observations from 49 image fields.

molecules) scales typically analyzed by a conventional Orbitrap mass spectrometer^{3,4}, there are reasonable prospects for decreasing sample volumes and protein abundance requirements. Provided that the challenge of fluorescently labeling low-abundance proteins can be met, this method may have potential for applications in, for example, single-cell proteomics experiments¹⁸.

In other respects, the approach resembles DNA- and RNA-sequencing pipelines, whose basis is the acquisition of large numbers of (often short) reads in parallel. Similarities include that the data are intrinsically amenable to digital quantification simply by counting reads and that longer reads tend to be richer in information. In principle, the method should work for both peptides (short reads) and full proteins (long reads). Currently, the partial sequence information gained by knowing protease specificity and the observed dye-destruction rates make application to peptides more practical. Parallel efforts are underway to develop long-read single-molecule protein sequencing based on nanopores^{7–10,19–21}.

The error spectrum of the method strongly resembles that of nucleic acid sequencing, because it is characterized by insertions/deletions (indels) and substitutions, rather than the attribution errors that predominate in mass spectrometry as a result of isobaric amino acids

or peptides. Many of the same concerns apply for this method and traditional Edman sequencing for optimization of PITC attachment and cleavage of PTH amino acids (Fig. 1c), and similar optimizations are required for temperatures and reagent-incubation times for efficient cleavage¹⁵. However, unlike Edman sequencing, this method does not rely on detecting PTH amino acids and thus is not affected by many challenges to the traditional method, including inefficient extraction and detection of PTH molecules and amino acid modification effecting PTH retention times^{13,22}. In addition, whereas the Edman method has the drawbacks of lags and decreased repetitive yield caused by loss of population synchrony, our approach differs in that a missed cleavage on one molecule has no effect on a different molecule, and cleavage efficiencies (91–97%) are simply modeled into database lookup probabilities.

Although we did not evaluate peptide quantification here, the intrinsically digital nature of the data offers both advantages and disadvantages over mass spectrometry. In contrast to mass spectrometry, in which assay dynamic ranges are largely set by mass-detector dynamic ranges of 10^3 – 10^4 (as for Orbitrap detectors^{3,4}) or counts of mass spectra collected (typically no more than 10^5), in single-molecule protein identification, the dynamic ranges should scale in a manner similar to those for imaging-based nucleic acid-sequencing methods, as set fundamentally by the surface area of the flow cell, density of attached molecules, and imaging times. Current-generation Illumina sequencers routinely collect hundreds of millions of reads per run, and one previously developed single-molecule DNA-sequencing instrument using a TIRF-microscope setup similar to those in this study has reported scaling to >1 billion molecules sequenced²³. In principle, single-molecule protein sequencing should scale similarly, offering multiple order-of-magnitude increases in dynamic range over current-generation proteomics platforms. However, a potential confounding issue distinct from DNA and RNA sequencing is the substantially larger dynamic range exhibited by some natural proteomes; for example, plasma protein concentrations can vary by more than 12 orders of magnitude²⁴. In such cases, approaches will be needed to simplify the samples, such as affinity-based subtraction of highly abundant proteins or biochemical fractionation before sequencing. Finally, directions for future development include methods for multiplexing samples, preparing low-abundance proteins/peptides, and expanding the palettes and stabilities of dyes and labelable amino acids or their modifications.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the [online version of the paper](#).

ACKNOWLEDGMENTS

We thank B. Cannon and R. Russell for early assistance with single-molecule imaging, M. Gadush for assistance with peptide synthesis, I. Riddington, J. Dinser, and K. Suhr for assistance in mass spectrometry analysis of fluorescently labeled peptides, Z. Simpson and J. Rybarski for assistance with image analysis, A. Ellington for many fruitful discussions, and the Texas Advanced Computing Center for high-performance computing. This work was supported by fellowships from the HHMI (to J.S.) and NSF (DGE-1610403 to A.A.B.), and by grants from DARPA (N66001-14-2-4051 to E.V.A. and E.M.M.), NIH (DP1 GM106408, R01 GM076536, and R35 GM122480 to E.M.M.), CPRIT (to E.M.M.), and the Welch foundation (F-1515 to E.M.M. and F-0046 to E.V.A.).

AUTHOR CONTRIBUTIONS

J.S., A.A.B., E.T.H., A.M.B., J.L.B., A.M.J., E.V.A., and E.M.M. designed and analyzed the experiments or interpreted the data. J.S., E.T.H., A.M.B., J.L.B., and J.M. performed the experiments. J.S., A.A.B., E.T.H., A.M.B., E.V.A., and E.M.M. wrote and edited the manuscript.

COMPETING INTERESTS

J.S., A.M.B., E.M.M., and E.V.A. are cofounders and shareholders of Erisyon Inc. J.S., E.M.M., and E.V.A. are co-inventors on granted US patent PCT/US2012/043769. J.S., A.A.B., E.T.H., J.L.B., A.M.J., E.V.A., and E.M.M. are co-inventors on pending US patent PCT/US2015/050099.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Liu, Y., Beyer, A. & Aebersold, R. On the dependency of cellular protein levels on mRNA abundance. *Cell* **165**, 535–550 (2016).
- da Costa, J.P., Santos, P.S.M., Vitorino, R., Rocha-Santos, T. & Duarte, A.C. How low can you go? A current perspective on low-abundance proteomics. *Trends Analyt. Chem.* **93**, 171–182 (2017).
- Makarov, A. *et al.* Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal. Chem.* **78**, 2113–2120 (2006).
- Makarov, A., Denisov, E., Lange, O. & Horning, S. Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. *J. Am. Soc. Mass Spectrom.* **17**, 977–982 (2006).
- Hawkrige, A.M. in *Quantitative Proteomics* (eds. Eyers, C.E. & Gaskell, S.) 3–21 (The Royal Society of Chemistry, Cambridge, 2014).
- Swaminathan, J., Boulgakov, A.A. & Marcotte, E.M. A theoretical justification for single molecule peptide sequencing. *PLoS Comput. Biol.* **11**, e1004080 (2015).
- Yao, Y., Docter, M., van Ginkel, J., de Ridder, D. & Joo, C. Single-molecule protein sequencing through fingerprinting: computational assessment. *Phys. Biol.* **12**, 055003 (2015).
- Zhao, Y. *et al.* Single-molecule spectroscopy of amino acids and peptides by recognition tunnelling. *Nat. Nanotechnol.* **9**, 466–473 (2014).
- Wilson, J., Sloman, L., He, Z. & Aksimentiev, A. Graphene nanopores for protein sequencing. *Adv. Funct. Mater.* **26**, 4830–4838 (2016).
- Kennedy, E., Dong, Z., Tennant, C. & Timp, G. Reading the primary structure of a protein with 0.07 nm³ resolution using a subnanometre-diameter pore. *Nat. Nanotechnol.* **11**, 968–976 (2016).
- Sampath, G. Amino acid discrimination in a nanopore and the feasibility of sequencing peptides with a tandem cell and exopeptidase. *RSC Advances* **5**, 30694–30700 (2015).
- Kolmogorov, M., Kennedy, E., Dong, Z., Timp, G. & Pevzner, P.A. Single-molecule protein identification by sub-nanopore sensors. *PLoS Comput. Biol.* **13**, e1005356 (2017).
- Edman, P. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* **4**, 283–293 (1950).
- Hernandez, E.T., Swaminathan, J., Marcotte, E.M. & Anslyn, E.V. Solution-phase and solid-phase sequential, selective modification of side chains in KDYWEK and KDYWE as models for usage in single-molecule protein sequencing. *New J. Chem.* **41**, 462–469 (2017).
- Hermodson, M.A., Ericsson, L.H., Titani, K., Neurath, H. & Walsh, K.A. Application of sequenator analyses to the study of proteins. *Biochemistry* **11**, 4493–4502 (1972).
- Phatnani, H.P. & Greenleaf, A.L. Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev.* **20**, 2922–2936 (2006).
- Stevens, S.M. Jr. *et al.* Enhancement of phosphoprotein analysis using a fluorescent affinity tag and mass spectrometry. *Rapid Commun. Mass Spectrom.* **19**, 2157–2162 (2005).
- Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
- Ohshiro, T. *et al.* Detection of post-translational modifications in single peptides using electron tunnelling currents. *Nat. Nanotechnol.* **9**, 835–840 (2014).
- Nivala, J., Marks, D.B. & Akeson, M. Unfoldase-mediated protein translocation through an α -hemolysin nanopore. *Nat. Biotechnol.* **31**, 247–250 (2013).
- Rosen, C.B., Rodriguez-Larrea, D. & Bayley, H. Single-molecule site-specific detection of protein phosphorylation with a nanopore. *Nat. Biotechnol.* **32**, 179–181 (2014).
- Wettenhall, R.E., Aebersold, R.H. & Hood, L.E. Solid-phase sequencing of 32P-labeled phosphopeptides at picomole and subpicomole levels. *Methods Enzymol.* **201**, 186–199 (1991).
- Pushkarev, D., Neff, N.F. & Quake, S.R. Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* **27**, 847–850 (2009).
- Anderson, N.L. & Anderson, N.G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).

ONLINE METHODS

Fluorophore selection. We observed that many commonly available fluorophores underwent substantial spectral shifts (>100 nm) or irreversible fluorescence loss after exposure to the Edman reagents, primarily the trifluoroacetic acid (TFA) and PITC/pyridine mixture. We screened 26 fluorophores (**Supplementary Table 1**) to identify those most resistant to the Edman solvents by covalently attaching the dyes to Tentagel beads (Chem-Impex International, 04773) and measuring their fluorescence after a 24-h incubation with TFA or pyridine/PITC in a 9:1 ratio at 40 °C (**Supplementary Fig. 1**). Nonspecifically bound fluorophores were removed by repeated washing with dimethylformamide (DMF), dichloromethane, and methanol. Atto647N, Alexa555, and rhodamine variants including TMR showed minimal (<5%) changes in fluorescence and had quantum yields sufficiently high for effective sequencing. We used Atto647N (quantum yield = 0.65) and rhodamine variants, including the improved TMR analog JF594 (quantum yield = 0.88), for all subsequent experiments.

Wide-field microscopy for bead-based assays. Beads labeled with fluorescent dyes or peptides were suspended in 20 μ L of phosphate-buffered saline (PBS, pH 7.2) and added to a glass coverslip. The samples were imaged with an Apo 60 \times /NA 0.95 objective mounted on an Eclipse TE2000-E inverted microscope (Nikon) equipped with a Cascade II 512 camera (Photometrics), a Lambda LS Xenon light source and a Lambda 10-3 filter-wheel control (Sutter Instrument), and a motorized stage (Prior Scientific), all operated via Nikon NIS Elements Imaging Software. Images were acquired at one frame per second through a 89000ET filter set (Chroma Technology) with channels 'DAPI' (excitation 350/50, emission 455/50), 'FITC' (excitation 490/20, emission 525/36) 'TRITC' (excitation 555/25, emission 605/52), and 'Cy5' (excitation 645/30 emission 705/72), and bead fluorescence was quantified from the images.

Peptide synthesis, purification, and labeling. All peptides were synthesized with a standard automated solid-phase peptide synthesizer (Liberty Blue microwave peptide synthesizer; CEM Corporation) and purified by analytical high-performance liquid chromatography (HPLC) (Shimadzu) with an Agilent Zorbax column (4.6 \times 250 mm) operating at a 10 mL/min flow rate, with elution with a gradient of 5–95% acetonitrile (0.1% TFA) over 90 min. Solvents used were HPLC grade. Peptides were labeled with fluorophores with standard coupling schemes¹⁴ by reaction with Atto647N-NHS, Atto647N-iodoacetamide, TMR-NHS, or JF549-NHS, as appropriate, to label lysines (via NHS) or cysteines (via iodoacetamide) (**Supplementary Table 1**). Purities, including the presence and count of fluorescent labels, were confirmed by mass spectrometry (6530 Accurate Mass QToF/MS, Agilent Technologies). N-terminal amines of synthetic peptides were typically blocked with a *tert*-butyloxycarbonyl (boc) or a fluorenylmethyloxycarbonyl (fmoc) protecting group before immobilization of peptides, thus preventing peptide concatenation of the activated C termini with free peptide N termini.

Labeling phosphoserines. Phosphorylated serines were fluorescently labeled (**Fig. 6**) by mixing solubilized phosphopeptide with a saturated solution of barium hydroxide and sodium hydroxide for 3 h at room temperature for beta-elimination of the phosphate¹⁷. Atto647N-NHS was reacted with cystamine to produce Atto647N-S-S-Atto647N, which was subsequently incubated overnight with the peptide solution and Tris(2-carboxyethyl)phosphine hydrochloride in DMF to fluorescently label the relevant serines. Peptides were purified by HPLC, and labeling was verified by mass spectrometry. Notably, this chemistry is known to additionally label phosphothreonines¹⁷ and also has the potential to eliminate O-glycans or to eliminate water from hydroxy amino acids²⁵.

Fluidics. We adapted an FCS2 temperature-controlled perfusion chamber (Biopetech), substituting the gaskets with custom gaskets die-cut from 0.05-mm-thick Kalrez-0040 rubber (Dupont), because of its compressibility and inertness to the Edman reagents (**Supplementary Fig. 2b**). We used a USB-controlled piston syringe (Cavro) and ten-port valve (Valco) to dispense reagents through polytetrafluoroethylene tubing into the perfusion chamber, which was affixed on the microscope stage (**Supplementary Fig. 2c**).

Tentagel-bead-based confirmation of Edman sequencing through fluorescent amino acids. Because the prior literature was unclear regarding the applicability of Edman chemistry to fluorescent-dye-modified amino acid residues, we used bead-based assays to test whether Edman sequencing could be observed in bulk studies of fluorescent peptides. Synthetic peptides with known positions of TMR-labeled lysine residues were covalently coupled to Tentagel beads via EDC/NHS chemistry (described below). We measured the decrease in peripheral bead fluorescence (attributable to covalent binding) after consecutive Edman cycles adapting established protocols²⁶, observing ~80% efficiency per amino acid residue without optimization, thus confirming the general ability of the Edman degradation chemistry to sequence peptides with bulky and hydrophobic fluorophore-tagged residues (**Supplementary Fig. 3**). We did not attempt to further optimize Edman chemistry on bulk peptides or beads.

Reducing photobleaching. We took advantage of the perfusion chamber's compatibility with diverse solvents to optimize the solution conditions for single-molecule imaging, testing imaging quality (dye brightness and half-life) in a range of organic and aqueous solvents. We observed optimal performance from methanol with 1 mM Trolox (Sigma, 238813-1G), purged 30 min with nitrogen gas. The methanol/Trolox imaging solution increased the half-life of the TMR and Atto647N fluorophores to 105 and 110 s, respectively, corresponding to >100 Edman cycles, assuming a 1-s exposure per cycle (**Supplementary Fig. 4**).

Peptide surface immobilization. For single-molecule Edman sequencing, a #1 (1.7-mm) glass cover-slip surface was first cleaned with UV/ozone (Jelight Company) and functionalized through amino-silanization with aminopropyltriethoxysilane (Gelest, SIA0610.1) with the vendor-supplied protocol (<http://www.gelest.com/wp-content/uploads/09Apply.pdf>). The slide surfaces were further passivated (for experiments in **Figs. 3–6** and **Supplementary Figs. 10, 13 and 14**) by overnight incubation with polyethylene glycol (PEG)-NHS solution, which was prepared by dissolving a mixture of 80 mg mPEG-SVA and 4 mg tBOC-PEG-SVA (Laysan Bio, MPEG-SVA-2000 and tBOC-NH-PEG-SVA-5K, respectively) in sodium bicarbonate solution, pH 8.2. Functionalized slides were stored in a vacuum desiccator until use. The *t*-butyloxycarbonyl protecting groups were removed by incubating a slide with 90% TFA (vol/vol in water) for 5 h before use, thus exposing free amine groups for peptide immobilization. Additionally, to aid in surface passivation, PEG sides were optionally treated with a 2% solution of Tween 20 (Bio-Rad, 170-6531) in Tris for 30 min (as for experiments in **Fig. 5c**). In control experiments, we confirmed that an amino-silanized glass surface was stable to multiple cycles of Edman degradation and after washes with wash buffer (1% sodium dodecyl sulfate and 0.1% Triton X-100 in PBS), as determined by assaying the retention of NHS-derivatized Atto647N covalently attached to free amines on the surface (**Supplementary Fig. 5**).

For a typical single-molecule peptide-sequencing experiment, peptides were covalently coupled to the cover-slip surface via amide bonds between the carboxylic acid of the C-terminal amino acid residue and the glass surface amines. Fresh solutions of 4 mM of 1-ethyl-3-(3-dimethylamino) propyl carbodiimide, hydrochloride (EDC; Sigma, 03449-1G) and 10 mM of NHS (Sigma; 130672-5G) or *N*-hydroxysulfosuccinimide (Thermo, PG82071) were made in buffer with 0.1 M 2-(*N*-morpholino)ethanesulfonic acid (MES, Pierce, 28390) immediately before use (notably, use of fresh EDC was critical). A solution of fluorescently labeled peptide (typically 200 μ M) was diluted with EDC-NHS solution (1:1 (vol/vol)) to a final concentration of 20 μ M peptide, 1.6 mM EDC, and 4 mM NHS. This solution was mixed for 4 h at room temperature before an initial dilution series was prepared in 0.1 M MES. We titrated peptides from a secondary dilution series to between 20 pM and 2 nM peptide in 0.1 M NaHCO₃ to provide an attachment density on the slide of approximately ten molecules per square nanometer (**Fig. 1b**). Peptides were typically incubated on the slide for 20 min before being washed with water and methanol to remove unbound peptide. Additionally, 1- μ m-long 12-mercaptododecanoic acid NHS ester-functionalized gold nanorods (Nanopartz, B14-1000-12CNHS-0.25-DMF) were covalently attached via the amines to serve as fiducial markers for focusing and image registration. After attachment of peptides and nanorods, the slide was incubated in 90% TFA (vol/vol in water)

for 5 h and then rinsed with methanol to remove boc groups and expose the peptides' free amino termini. Alternatively, to remove fmoc protecting groups, the peptides were incubated for 1 h in 20% piperidine solution (in DMF), then washed with DMF and methanol to remove residual piperidine. An optional 1-h incubation of 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU, Sigma, 33682) was used to remove peptides nonspecifically bound to the surface experiments in **Figures 5c** and **6** and **Supplementary Figure 8**. To aid in focus stability, the chamber and microscope were allowed to equilibrate at 40 °C during deblock- and up to an additional 12 h.

Total internal reflection microscopy. Single-molecule TIRF microscopy experiments were performed with two similar systems, each with a Nikon Ti-E inverted microscope equipped with a CFI Apo 60×/1.49 NA oil-immersion objective lens, a motorized stage with a 100-nm-resolution linear encoder (ProScan II; Prior Scientific), an iXon3 DU-897E 512 × 512 EMCCD detector (Andor) operated at -70 °C, and a MLC400B (Keysight) laser combiner with 561-nm (N1245AL34) and 647-nm (N1245AL44 and N1245BL56, systems A and B, respectively) lasers, as diagrammed in **Supplementary Figure 2c**. Fluorescence from Atto647N was excited with 6.0 mW (50%, system A) or 2.8 mW (12.5%, system B) of 647-nm laser power via a 647LP dichroic filter and collected through 665LP and 705/72BP emission filters. TMR fluorescence was excited with 2.7 mW (35%) of 561-nm laser power via a 561LP dichroic filter and collected through 575LP and 600/50BP emission filters. Gold-nanorod reflection was excited with <0.01 mW (3%) of 561-nm laser light with a 95/5 reflectance cube. To increase the number of pixels in an individual diffraction-limited spot and to maximize the flat-field portion of the image collected, an additional 1.5× tube lens was inserted into the beam path. Laser powers were measured before the objective. All data presented in figures, except those in **Figure 5b** and **Supplementary Figures 8, 12** and **14**, were collected with system A. All peptide-sequencing results were independently confirmed for both systems.

Automated Edman degradation. For single-molecule Edman sequencing experiments, the sample temperature was maintained at 40 °C by heating both the perfusion chamber and microscope objective. Edman reagents were bubbled with dry nitrogen gas for 10 min, and then kept under nitrogen gas throughout the experiment. Solvent exchanges in the fluidic device were controlled with in-house Python scripts and coordinated with image acquisition via custom macros in the Nikon Elements software package. Reagents (of the highest purity available from Sigma) were introduced to the perfusion chamber in the steps described in **Supplementary Table 2**.

'Wash' denotes exchanging the solvents in the flow chamber (approximately 3 min). On system A, free-base solution 1 was used for experiments in **Figures 1–4** and **5a,d** and **Supplementary Figures 10** and **13**, and free-base solution 2 was used for experiments in **Figures 5c** and **6**. On system B, free-base solution 1 was used for **Figure 5b** and **Supplementary Figures 12b** and **14**; free-base solution 2 was used for **Supplementary Figures 8** and **12a**. Typically, to distinguish signal loss specifically due to Edman chemistry, as many as four mock Edman cycles with all reagents except PITC were performed before Edman cycles. In total, steps 1–11 took approximately 1 to 1.5 h.

Image processing and photometry. Images of each field of view taken after each consecutive Edman cycle and stored as PNG files, with sets of images from each Edman cycle (henceforth, 'frames') sequentially collated into fields of view by file name.

To identify individual peptide molecules in frames, we applied a median filter to locate candidate fluorescent point sources in images (**Supplementary Fig. 6**). Candidate point sources were then fit with a two-dimensional Gaussian as an approximation to their Airy disc, as implemented in AGPY (authored by A. Ginsburg; downloaded 7 April 2015 from <https://github.com/keflavich/>), and an R^2 quality of fit was assessed, retaining point sources with $R^2 > 0.7$. Further criteria were applied as described below to remove potential contaminants from analysis.

To track individual fluorescent point sources through an experiment, the frames of each field of view were aligned pairwise across cycles with fast Fourier transform cross-correlation²⁷ (implemented in Python with scikit-image; <http://scikit-image.org/>) of the gold-nanorod reflection-channel

images, if present, or one of the fluorescence channels otherwise. We collated instances of each fluorescent point source across aligned frames by matching their coordinates with the alignment offsets, within error tolerance. If a fluorescent point source was absent in one or more frames, its position was extrapolated to those frames with the alignment offsets. Point sources that mapped outside of a frame in any imaging cycle were discarded.

We quantified the fluorescence intensity of each point-source across frames with Mexican hat photometry. In each frame, we summed the innermost 7 × 7 pixels centered about the point source to obtain its raw photometry, then subtracted the median of the enclosing 19 × 19 pixel area (excluding the 7 × 7 center) to adjust for background. Any point source whose Mexican hat was not contained entirely in all frames was discarded.

For each point source's progression through frames, we constructed a Boolean logic sequence consisting of two possible states: ON and OFF (**Supplementary Fig. 7**). A point source was considered ON in the frames in which a 2D Gaussian fit with $R^2 > 0.7$ was found or was considered OFF otherwise. For example, a point source that was well fitted with a 2D Gaussian in frames 1–3, was not detected in frames 4–6, and was again fittable in frames 7–10 would be assigned a sequence [ON, ON, ON, OFF, OFF, OFF, ON, ON, ON, ON]. Only point sources that turned off monotonically were considered validly sequenced peptides; i.e., they started in the ON state, and if they turned OFF in any frame, they then remained OFF for the rest of the experiment. Fluorescent point sources that turned ON after being OFF at any point were discarded from further consideration. For each point source, the sequence of its Boolean states and its Mexican hat photometries was collated. This collated sequence is termed the point source's track.

Before further analysis, the dye photometries were adjusted to account for frame-to-frame focus variations. Tracks with the ON state across all frames ('remainders') were collated for each field. The percentage deviation in fluorescence intensity was determined at each cycle for each remainder track. The average remainder deviation for each cycle was then applied to all tracks within that field. Fields with fewer than five remainders were removed from further analysis.

Overview of maximum-likelihood assignment of dye positions. For each peptide track, we sought to infer the number of dyes remaining on the peptide in each sequencing frame. For example, a track's dye count might be written as [3, 3, 2, 2, 1, 0], representing a peptide that started with three dyes, decreased to two dyes after two Edman cycles, decreased to one dye after another three cycles, and finally decreased to zero dyes after one more cycle.

Dye counts are not directly observable, but instead must be inferred from the measured photometries and their stepwise intensity losses^{28,29}. Our general strategy to infer a sequence of dye counts d_i from a sequence of photometries φ_i across frames 1, 2, ..., i was as follows:

1. A peptide had a dye count of 0 in a frame if and only if it was in the OFF state as defined above.
2. We considered all possible monotonically decreasing dye-count sequences as competing explanations for the observed sequence of photometries. We considered a maximum of five dyes, allowing multiple simultaneous dye losses per cycle.
3. The probability of the observed intensities being generated from each dye-count sequence was calculated as a quality-of-fit score $S(d|\varphi) = \prod_i S(d_i|\varphi_i)$, with i indexing the track's frames. The per-frame scoring function $S(d_i|\varphi_i)$ is the probability density function $\rho(\varphi_i|d_i)$ of a point source with d_i dyes yielding photometry φ_i . This probability density function is log-normal, as described below. The dye-count sequence d maximizing this score was taken as the best explanation for observed photometries φ .
4. To guard against poorly behaved fluorescent point sources, if the best-fitting dye-count sequence d had any frame for which $S(d_i|\varphi_i)$ was below a threshold (defined below), we considered the track uninterpretable and discarded it from further consideration.

Single-molecule dye fluorescence intensities are log-normally distributed. Tracks from each experiment represented a population of fluorescent point sources that could be characterized in bulk. Here and in subsequent analysis, the distribution of photometries was binned with the optimal histogram-binning algorithm from Shimazaki *et al.*³⁰.

We first characterized the intensities of peptides with only one dye remaining. Because each track contains a sequence of ON/OFF states, we can assume that the last ON state of each track before an OFF is, in most cases, caused by loss of a single dye regardless of how many dyes the peptide began with. This assumption is valid on a population basis because the probability of two or more dyes turning OFF in a single cycle is low relative to that of one dye. We defined φ_{final} as the set of photometries of the last ON frames that are followed by an OFF frame across all tracks. To maximize the ON/OFF transitions caused by a single dye loss and not whole-molecule loss, tracks with OFF transitions in the first three frames (typically mock cycles) were excluded from this definition. We found the distribution of φ_{final} to be log-normal, matching observations by Mutch *et al.*³¹ (Supplementary Fig. 9a). A log-normal distribution for one fluorophore can be written as a probability density function ρ of intensity φ :

$$\rho(\varphi) = \frac{1}{\varphi \times \sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\ln \varphi - \mu)^2}{2\sigma^2}\right)$$

where the scale parameter μ and shape parameter σ completely characterize the distribution.

For simplicity, we henceforth considered the logarithmic space $\varphi^* = \ln \varphi$, with the corresponding transformed probability density function and parameters:

$$\rho^*(\varphi^*) = \frac{1}{\sqrt{2\pi\sigma^{*2}}} \exp\left(\frac{-(\varphi^* - \mu^*)^2}{2\sigma^{*2}}\right)$$

Following Mutch *et al.*³¹, the log-normal distribution can be expanded to cases of multiple (c) dyes by increasing the scale parameter μ^* to $\mu^* + \ln c - Q_c$, where Q_c is a dye-dye-interaction factor; the shape parameter σ^* is held constant, per ref. 31. Thus, the probability density function for a point source with multiple dyes can be written as:

$$\rho^*(\varphi^*|c) = \frac{1}{\sqrt{2\pi\sigma^{*2}}} \exp\left(\frac{-(\varphi^* - \mu^* - \ln c + Q_c)^2}{2\sigma^{*2}}\right)$$

In this context, step 4 of the general fitting strategy (thresholding) was based on the deviations of a track's observed photometries from the log-normal model, $\frac{|\varphi^* - \mu^*|}{\sigma^*}$. If this deviation was greater than three in any frame, the track was discarded.

Inference of log-normal fluorescence parameters via simulation. The parameter μ^* in $\rho^*(\varphi^*|c)$ can be obtained directly by setting $\mu^* = \langle \varphi_{\text{final}}^* \rangle$. Parameters σ^* and Q_c are more challenging to extract by applying a straightforward function to data points φ_i^* . Instead, we used forward simulation to find a combination of parameters under which our fitted model best matched our data. Specifically, we started with fluorosequencing data from a doubly labeled peptide GC*AGC*AGAG (where ♦ indicates Atto647N conjugated to cysteine) experiment and calculated its μ^* . We then computationally generated each possible monotonically decreasing dye count that dropped to 0 within the experiment's number of cycles. Iterating over 225 parameter combinations of σ^* and Q_c , we generated 10^5 tracks for each of the possible dye counts as follows: the intensity of each frame in a track was randomly drawn from the distribution $\rho^*(\varphi^*|c)$, with c determined according to the corresponding dye count d_i in that frame. We applied the general fitting strategy to both these simulated tracks and experimental tracks with σ^* and Q_c . To gauge whether a particular pair of parameters σ^* and Q_c recapitulated the distribution of photometries well, we collated the dye sequences fitted to the experimental data with their simulated counterparts and compared the distribution of photometries in each frame. Supplementary Figure 9b shows the distribution of photometries in each frame for dye sequence [2, 2, 2, 2, 1, 1, 1, 0, 0, 0] (mocks included; frames with 0 dyes are OFF and are omitted) under an overestimated shape parameter σ^* and an underestimated dye-dye-interaction factor Q_2 . Supplementary Figure 9c shows the corresponding distributions for $\sigma^* = 0.20$ and $Q_2 = 0.30$, which fit with an average R^2 of 0.87 across all eight frames. Repeating this parameter sweep for multiple experiments showed these values for σ^* and Q_c to be generally valid for Atto647N, and we used them for all subsequent analyses.

With these parameters, for any given track φ , we could thus infer the underlying dye-count sequence d by maximizing the fit score:

$$\max_d S(d|\varphi) = \prod_i S(d_i|\varphi_i) = \prod_i \frac{1}{\sqrt{2\pi\sigma^{*2}}} \exp\left(\frac{-(\varphi_i^* - \mu^* - \ln d_i + Q_{d_i})^2}{2\sigma^{*2}}\right)$$

Estimation of sequencing errors via Monte Carlo simulations. Peptide fluorescent sequences are subject to experimental error. We took advantage of our previously developed computational model of the likeliest sources of experimental error (Edman failure, photobleaching, and dud dyes)⁶, for which we had developed an extensible Monte Carlo model of fluorosequencing. We added to our previous model an additional source of error—whole-molecule loss—to reflect our observations that the reagent that flushes through the perfusion chamber could remove labeled but nonspecifically bound peptides from the slide surface, especially during the first few experimental cycles. With this error model, we were able to simulate the molecular state of a peptide after each experimental cycle, thus providing a simulated dye-count sequence for a given peptide as it undergoes sequencing. By chaining together this existing framework with our log-normal model of fluorescence, we thus simulated the complete experimental observations (tracks) that we would expect for any given peptide sequence. Applying our dye-count-inference fitter to the simulated data, we thus obtained a set of modeled fluorescent sequences for comparison to an actual experiment.

The primary sources of error are modeled as follows (more detailed discussion in ref. 6):

- Edman failure is modeled as a Bernoulli variable. The probability of an amino acid being removed after every Edman cycle is p and is independent of all other events.
- Photobleaching is modeled as an exponential decay. The probability of a dye photobleaching after any experimental cycle is e^{-b} .
- The probability of a dye being a nonfluorescing dud before the experiment begins is a Bernoulli variable with dud probability u .
- Whole-molecule loss is modeled as a bimodal Bernoulli variable, with probability d_{initial} of a whole molecule being removed after every cycle during the initial c cycles, and probability $d_{\text{subsequent}}$ per cycle thereafter. According to experimental observations, $d_{\text{initial}} \geq d_{\text{subsequent}}$.

We were able to recapitulate experiments with simulations (for example, Supplementary Fig. 10), and found that the error parameters were broadly conserved across multiple experiments.

Adjustment of sequencing histograms for expected background rates of dye destruction and whole-molecule loss. We compiled data across multiple acetylated-peptide-sequencing experiments to establish a background rate of nonspecific dye destruction and whole-molecule loss, and we adjusted the sequencing histograms (where indicated) to account for this background. Importantly, acetylated-control experiments exhibited fluorescent sequencing patterns in a sequence-independent manner, depending only on the per-molecule dye count; hence, we were able to pool all acetylated experiments with a given dye count to obtain a general background distribution, which could be used to adjust histograms from a sequencing experiment for observations expected by chance.

First, we standardized each acetylated-control experiment by converting the counts at each histogram position to relative frequencies, by dividing each count by the total number of observations in the experiment. We considered all step-drop patterns that dropped to a dye count of 0 by the end of the experiment, including those that had a total of four or five drops; step-drop patterns that remained above 0 in the last frame (i.e., remainders) were omitted. Multiple standardized acetylated experiments were then averaged on a per-histogram position basis to obtain the average background rate, i.e., the normalized count of each step-drop pattern expected by chance as a result of nonsequencing-related experimental losses. Likewise, we obtained the variance in the background rate on a per-histogram position basis. We assumed the background rate at each histogram position to follow a normal distribution, defined by the average and variance obtained from multiple acetylated-control experiments.

We then adjusted the sequencing-experiment histograms for expected background with the following iterative algorithm:

1. Standardize the sequencing histogram as for individual acetylated histograms above.
2. For each position in the standardized sequencing histogram with standardized frequency S , compute its z score against the background distribution's mean μ and s.d. σ :

$$z = \frac{S - \mu}{\sigma}$$

3. Define a smoothing operation for sequencing histogram position $H = (i, j, k, \dots)$ as replacing its raw counts with the average of counts at all positions within a Hamming distance of 1. For example, smoothing at position 6 would entail averaging counts at positions 5 and 7, and smoothing at position (3, 4) would entail averaging counts at all eight positions satisfying $(3 \pm 1, 4 \pm 1)$. Of note, after a smoothing operation, a sequencing histogram must be re-standardized with the updated counts to compute its scores.

4. Score all peaks in the sequencing histogram for the largest decrease in z score that would result from background correction, with smoothing from adjacent histogram positions to compensate for outliers, calculated as follows:

$$\max_H \Delta Z = \max_H Z_H - Z'_H = \max_H \left(\frac{S_H - \mu_H}{\sigma_H} - \frac{S'_H - \mu_H}{\sigma_H} \right) = \max_H \frac{S_H - S'_H}{\sigma_H}$$

where S_H , μ_H , and σ_H are the histogram value, background mean, and background s.d. at position H , and S'_H is the corresponding value for the histogram smoothed at position H .

5. Update the histogram by smoothing at the position yielding the best improvement in step 4.

6. Repeat from step 1 until the highest z score in step 2 is below a specified threshold (for example, $z = 1$), or no further interpolation can be made that lowers a z score.

This procedure generates a smoothed estimate of the expected background counts for a given sequencing experiment; we simply subtracted these counts from raw foreground sequencing counts to obtain the adjusted foreground counts, setting any negative entries to 0. Of note, the z -score threshold applied in step 6 effectively considers any peaks whose z score is below it as background and thus removes them from the final results.

Effect of experimental errors on protein identification. To assess the potential effect of our observed experimental error rates on protein identification,

we re-simulated the cellular compartments considered under ideal conditions in **Figure 1c** with error rates, calculated with the Monte Carlo-simulation algorithm as described above and in ref. 6. We considered the case for rates measured for the experimental samples in **Figures 3 and 4**, and **Supplementary Figure 10** of 94% Edman efficiency, 5% dye destruction, 5% surface degradation, and 7% dud dyes. For each set of proteins, we simulated 10,000 copies of each protein in a Monte Carlo fashion for 30 Edman cycles and tabulated their resulting fluorescent sequences. We defined a protein as being uniquely identified if it yielded a fluorescent sequence at least ten times (out of 10,000) for which no more than 10% of the counts of that fluorescent sequence were emitted by other proteins (protein-coverage curves in **Supplementary Fig. 15**).

Statistics and reproducibility. Replicate data are summarized for all figures in **Supplementary Data 1** and the Life Sciences Reporting Summary.

Code availability. Computer code is freely downloadable from GitHub, including image-processing algorithms and Monte Carlo simulations, at <https://github.com/marcottelab/FluorosequencingImageAnalysis/> or in **Supplementary Software**.

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

Data availability. All single-molecule-imaging data sets and an image-processing tutorial are available via Zenodo (<http://doi.org/10.5281/zenodo.782860>).

25. McLachlin, D.T. & Chait, B.T. Improved beta-elimination-based affinity purification strategy for enrichment of phosphopeptides. *Anal. Chem.* **75**, 6826–6836 (2003).
26. Laursen, R.A. Solid-phase Edman degradation: an automatic peptide sequencer. *Eur. J. Biochem.* **20**, 89–102 (1971).
27. Guizar-Sicairos, M., Thurman, S.T. & Fienup, J.R. Efficient subpixel image registration algorithms. *Opt. Lett.* **33**, 156–158 (2008).
28. Cannon, B., Pan, C., Chen, L., Hadd, A.G. & Russell, R. A dual-mode single-molecule fluorescence assay for the detection of expanded CGG repeats in Fragile X syndrome. *Mol. Biotechnol.* **53**, 19–28 (2013).
29. Das, S.K., Darshi, M., Cheley, S., Wallace, M.I. & Bayley, H. Membrane protein stoichiometry determined from the stepwise photobleaching of dye-labelled subunits. *ChemBioChem* **8**, 994–999 (2007).
30. Shimazaki, H. & Shinomoto, S. A method for selecting the bin size of a time histogram. *Neural Comput.* **19**, 1503–1527 (2007).
31. Mutch, S.A. *et al.* Deconvolving single-molecule intensity distributions for quantitative microscopy measurements. *Biophys. J.* **92**, 2926–2943 (2007).

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](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
- An indication of 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 statistics 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
- Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

Nikon Ti Elements software was used to capture the microscope images and image data converted in TIFF format.

Data analysis

Computer code is freely downloadable from GitHub, including image processing algorithms and Monte Carlo simulations, at <https://github.com/marcottelab/FluorosequencingImageAnalysis>.

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 upon request. 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

All single molecule imaging datasets and an image processing tutorial are available via Zenodo (10.5281/zenodo.782860).

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

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

Sample size	Not Applicable. However for most data presented, the number of image fields were between 50 - 100, as described in each figure text.
Data exclusions	p. 15, lines 1-6; p. 15, lines 27-30
Replication	Yes, all data was replicated multiple times independently. p 13, lines 40-42, All peptide fluorosequencing results were independently confirmed on two separate microscopes by two independent investigators. Replicates are listed in Supp. File 1.
Randomization	Not applicable. No experimental trials requiring randomization were performed
Blinding	Not applicable. Due to the nature of the experiments performed, blinding studies are not applicable, and results were computed algorithmically.

Behavioural & social sciences study design

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

Study description	<input type="text"/>
Research sample	<input type="text"/>
Sampling strategy	<input type="text"/>
Data collection	<input type="text"/>
Timing	<input type="text"/>
Data exclusions	<input type="text"/>
Non-participation	<input type="text"/>
Randomization	<input type="text"/>

Ecological, evolutionary & environmental sciences study design

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

Study description	<input type="text"/>
Research sample	<input type="text"/>
Sampling strategy	<input type="text"/>
Data collection	<input type="text"/>
Timing and spatial scale	<input type="text"/>
Data exclusions	<input type="text"/>
Reproducibility	<input type="text"/>
Randomization	<input type="text"/>

Blinding

Did the study involve field work? Yes No

Field work, collection and transport

Field conditions

Location

Access and import/export

Disturbance

Reporting for specific materials, systems and methods

Materials & experimental systems

- | | |
|-------------------------------------|------------------------------------------------------|
| n/a | Included in the study |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Unique biological materials |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Human research participants |

Methods

- | | |
|-------------------------------------|-------------------------------------------------|
| n/a | Included in the study |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials

Antibodies

Antibodies used

Validation

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

Authentication

Mycoplasma contamination

Commonly misidentified lines
(See [ICLAC](#) register)

Palaeontology

Specimen provenance

Specimen deposition

Dating methods

 Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

Wild animals

Field-collected samples

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

Recruitment

ChIP-seq

Data deposition

Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).

Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

Files in database submission

Genome browser session

(e.g. [UCSC](#))

Methodology

Replicates

Sequencing depth

Antibodies

Peak calling parameters

Data quality

Software

Flow Cytometry

Plots

Confirm that:

The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

All plots are contour plots with outliers or pseudocolor plots.

A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Instrument

Software

Cell population abundance

Gating strategy

 Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Design specifications

Behavioral performance measures

Acquisition

Imaging type(s)

Field strength

Sequence & imaging parameters

Area of acquisition

Diffusion MRI

 Used Not used

Preprocessing

Preprocessing software

Normalization

Normalization template

Noise and artifact removal

Volume censoring

Statistical modeling & inference

Model type and settings

Effect(s) tested

Specify type of analysis: Whole brain ROI-based BothStatistic type for inference
(See [Eklund et al. 2016](#))

Correction

Models & analysis

n/a | Involved in the study

 Functional and/or effective connectivity Graph analysis Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Graph analysis

Multivariate modeling and predictive analysis