

Sampling the proteome by emerging single-molecule and mass spectrometry methods

Michael J. MacCoss, Javier Antonio Alfaro, Danielle A. Faivre, Christine C. Wu, Meni Wanunu & Nikolai Slavov



Mammalian cells have about 30,000 times as many protein molecules as mRNA molecules, which has major implications in the development of proteomics technologies. We discuss strategies that have been helpful for counting billions of protein molecules by liquid chromatography–tandem mass spectrometry and suggest that these strategies can benefit single-molecule methods, especially in mitigating the challenges posed by the wide dynamic range of the proteome.

The ubiquitous roles of proteins in biomedicine are well appreciated and have motivated technologies seeking to advance two key parameters: the sensitivity and throughput of quantitative protein analysis. While proteomic technologies may use different approaches, they face similar challenges, such as quantifying proteins of vastly different abundances, some present in only a few copies and some present in tens of millions of copies per typical mammalian cell. This wide dynamic range poses a substantial challenge for investigating proteome biology.

Mass spectrometry (MS) has powered proteomics from the first demonstration of peptide sequencing using MS in the 1970s^{1,2}. Since then, milestones in MS-based proteomics have included *de novo* sequencing entire proteins in the late 1980s^{3,4}, soft ionization by electrospray⁵, automated spectral interpretation⁶, multiplexing the acquisition of spectra on different peptides using data-independent acquisition⁷, multiplexing the acquisition of different samples using tandem mass tags⁸ and quantifying thousands of proteins in single human cells^{9,10}. Together, the steady growth in the rate of protein identification using MS has been reminiscent of Moore's law, resulting in about 1,250-fold higher throughput: from about 20 protein data points per hour in 2001 (ref. ¹¹) to about 25,000 protein data points per hour achieved by plexDIA¹². This increased throughput has been critical for addressing challenges in biomedical research¹³. It also highlights the power of experimental strategies and technological progress to tackle the immense demands of proteomics in terms of quantity and dynamic range that is required for thorough analysis, given the large number of proteins of widely varying concentrations in a cell.

More recently, non-MS methods have made exciting steps toward identifying and potentially sequencing single polypeptide molecules^{14–16}. Conceptually, these methods aim to adapt flow-cell and

nanopore methods developed for nucleic acid analysis for protein analysis. Flow-cell-based methods include highly parallel single-molecule N-terminal peptide sequencing methods based on either Edman degradation¹⁷ or amino peptidases¹⁸. Another approach aims to use degenerate affinity reagents to recognize individual protein molecules separated spatially in a flow cell^{19,20}. Other groups are working to adapt nanopore sequencing to peptides and proteins^{21,22}. Most of these methods aim to detect a subset of the amino acids within a polypeptide sequence, which provides a fingerprint, or a constraint, on choosing a sequence among the known protein coding gene products from the genome. While these methods have yet to be applied to biologically derived protein mixtures, they have generated enthusiasm within the scientific community as a complement to MS analysis¹⁴.

These developments have motivated renewed interest and investment in advancing proteomics technologies, as reflected in private funding¹⁴ and in recent National Human Genome Research Institute funding opportunities aimed at accelerating the development of technologies for single-molecule sequencing and single-cell proteome analysis. Because there is excitement around emerging single-molecule counting methods for proteomics, we felt it was timely to compare them to strategies used by the current state-of-the-art proteomics methods based on liquid chromatography–tandem mass spectrometry (LC–MS/MS). We hope our opinion will provide benchmarks and directions for the technological breakthroughs that need to be achieved for single-molecule protein or peptide counting to achieve parity with and complement the capabilities of LC–MS/MS-based proteomics methods.

How many molecules need to be counted?

Many of the challenges for accurate and sensitive protein quantification, such as the quantification over a wide dynamic range, are shared by all proteomics methods. Indeed, a typical mammalian cell contains billions of protein molecules but less than half a million RNA molecules²³ (Fig. 1a). Some proteins are present at tens of copies per cell while others (for example, histones) at tens of millions of copies per cell, resulting in about a 10^6 dynamic range²⁴. The range of protein abundances is even larger for body fluids such as plasma, in which protein abundances may differ by 10^{10} , for example, between albumin and interleukin-6 (ref. ²⁵). This presents a fundamental challenge because the presence of abundant proteins makes it rare to count molecules from low-abundance proteins: one would have to count billions of albumin molecules before having a chance to detect a single interleukin-6 molecule. This means that the single-molecule approaches that have been used successfully to quantify the transcriptome, which spans about a 10^3 dynamic range, face major challenges in scaling to quantify the proteome²².

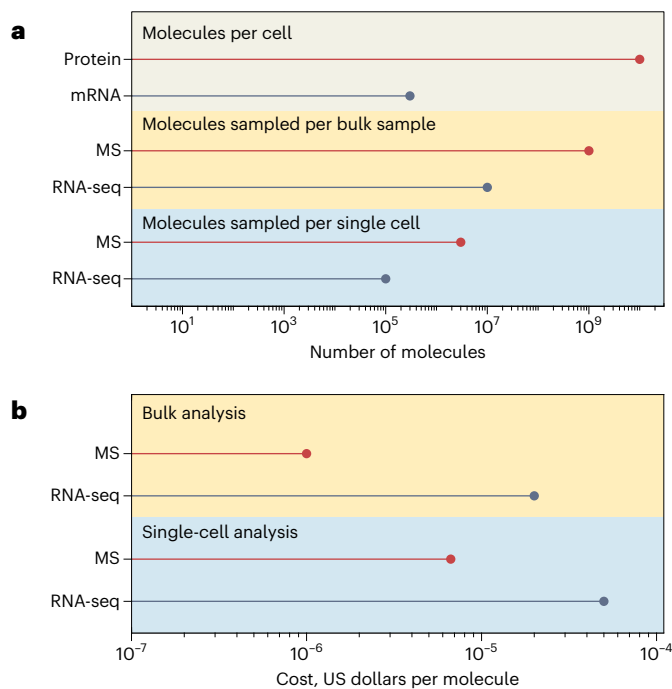


Fig. 1 | Overview of RNA and protein statistics. a, A representative human cell, such as a fibroblast, has billions of protein molecules, as compared to merely hundreds of thousands of RNA molecules²⁶. Accordingly, MS analysis samples more protein molecules per sample than the RNA molecules sampled by RNA-seq. **b**, Estimated cost per molecule for MS and RNA-seq. The single-cell estimates are based on published numbers for unique molecular identifiers per single cell analyzed by Smart-seq3 (ref. ⁶⁷) and number of protein molecules counted by plexDIA¹².

A typical mammalian cell – for example, a HeLa cell with a volume of ~3,000 μm^3 – contains about 300,000 mRNA molecules²⁶ and about 10 billion protein molecules²³ (Fig. 1a). The cell is a crowded mesh of proteins, with a typical density of 3 million protein molecules per cubic micrometer. Even a yeast cell with a volume of ~30 μm^3 contains ~100 million molecules. This protein density estimate has been supported independently using molecular measurement based on MS, as well as fluorescence microscopy using green fluorescent protein. Given these independent measurements, it is estimated that the typical HeLa cell contains at least 5–10 billion proteins per cell and others like macrophages (5,000 μm^3) and cardiomyocytes (15,000 μm^3) will contain substantially more. Because of this range in volume, we used ~10 billion proteins per cell in our calculations.

Given these estimates of the relative abundance ratio of mRNA to protein molecules, we calculate that about 30,000-fold more counts are required to characterize the protein molecules at an analogous coverage to that which has been achieved with the transcriptome (Fig. 1). Given the potential need to count a large number of protein molecules, we next explore the feasibility of achieving the required scale at affordable cost using estimates for cost per molecule. This factor is important, but it must be considered in the context of many other factors, such as the ability to sample large numbers of diverse sequences and to multiplex efficiently.

How much do single-molecule counting methods cost?

While single-molecule protein counting approaches are yet to report the analysis of a complex protein mixtures, we believe that with time and resources the efforts to read peptide sequences in a spatially parallelized format will be successful¹⁴. Without knowing what the capabilities and limitations are for these emerging protein and peptide sequencing methods, we make the optimistic assumption that these methods will be able to achieve sequencing counts of polypeptides on par with what state-of-the-art Illumina sequencing can achieve currently with oligonucleotides. Thus, we use single-molecule RNA sequencing by Illumina as a proxy to represent single-molecule protein counting approaches (Fig. 1b). To estimate the cost for current advanced technologies, we use an estimate of US\$10,000 for sequencing 4 billion reads by Illumina NovaSeq over ~2 days and US\$500 for performing a 2-h quantitative LC-MS/MS analysis. These costs were chosen as conservative estimates based on inquiries from several academic core facilities, and the rates include personnel, sample preparation and basic computational analysis as part of the service. While academic research laboratories may achieve lower costs, these prices represent objective estimates for widely accessible services. The cost per protein molecule analyzed by LC-MS/MS is lower than the cost per DNA molecule sequenced by Illumina (Fig. 1b). This indicates that single-molecule DNA sequencing has not yet achieved a cost that would enable counting of sufficient numbers of molecules to achieve affordable and comprehensive quantification of mammalian proteomes.

Counting ions by LC-MS/MS

Traditionally, the MS proteomics field reports lists of peptides detected and the proteins they are derived from. As peptides elute off the high-performance liquid chromatography column, the instrument counts large numbers of peptide ions based on their mass-to-charge (m/z) ratio, independently of their sequence identification (Fig. 2a). The abundance of each analyte is often determined from a background-subtracted peak area in the extracted ion chromatogram(s). Depending on the method used, the peak area can be obtained from the unfragmented MS1 spectra or from tandem mass spectra (MS/MS or MS2) collected using methods like data-independent acquisition. The peak area is derived from the detector ion current, either from the flow of ions to an electron multiplier²⁷ or the generation of an image current in a Fourier transform mass analyzer²⁸. The current is a measure of the number of ions (charged molecules) counted, normalized by the amount of time spent sampling the signal. The measured signal is proportional to ions per second, and thus it can be converted into a number of counted ions and for direct comparison with single-molecule counting methods^{9,29,30}.

LC-MS/MS methods can improve the sensitivity to low-abundance analytes by changing the time spent sampling the signal (also known as dwell time, integration time or injection time). In some MS instruments, such as ion traps, the time spent sampling ions changes dynamically depending on the signal at that time³¹. This dynamic adjustment of the injection time, known as automatic gain control (AGC), provides an ideal ion population for the MS measurement (Fig. 2b). However, an added benefit of AGC is that it enables the instrument to spend less time on abundant molecular species but scale the current into a larger quantity while maintaining quantitative linearity. Likewise, it enables the instrument to spend more time on less abundant peptides to enable the measurement of the weaker signal. This increases the dynamic range and the total number of ions identified (Fig. 2c). Dividing each spectrum intensity by the time taken to acquire the spectrum gives a

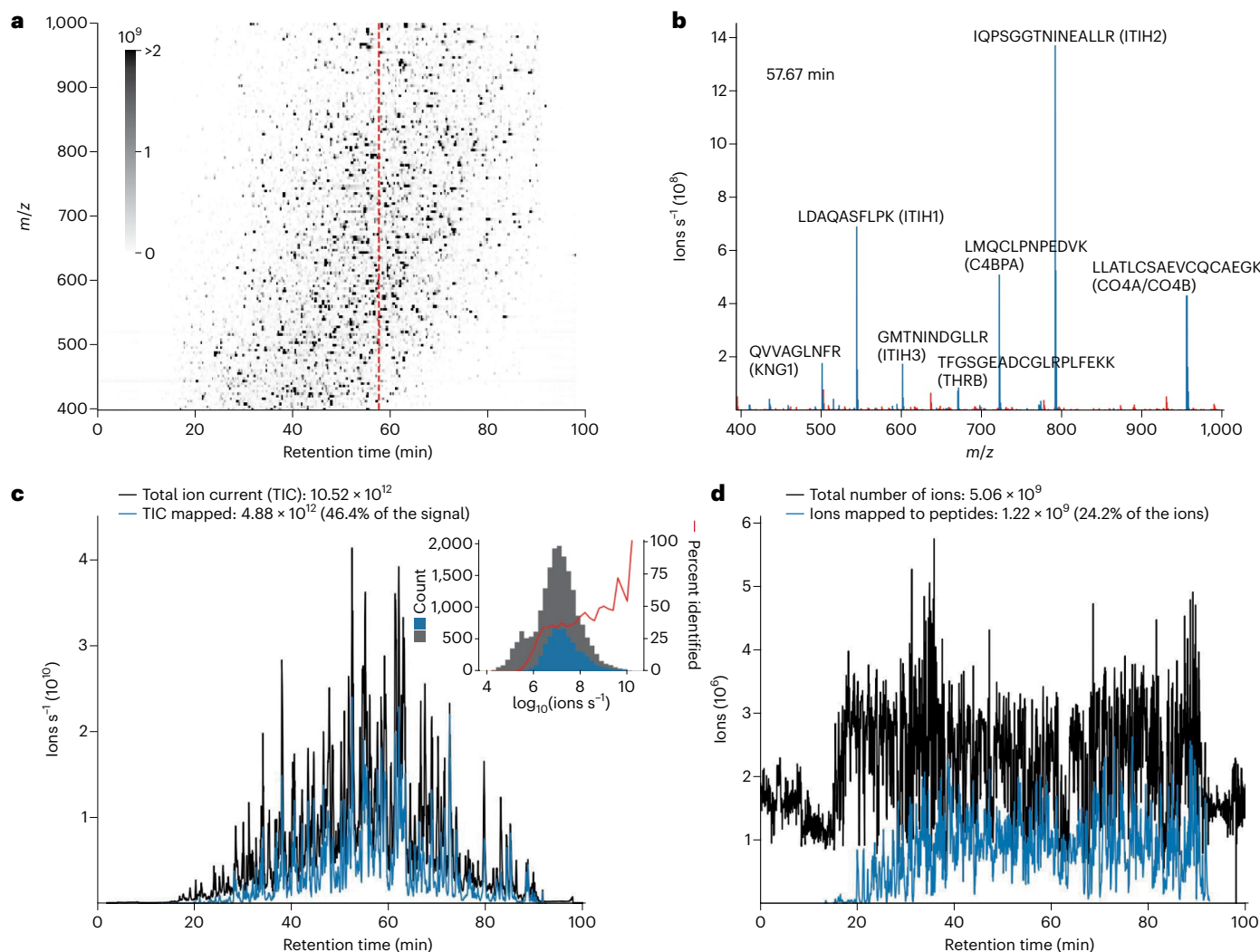


Fig. 2 | A liquid chromatography-mass spectrometry experiment can count billions of peptide ions within 90 min. Signal from the MS1 spectra of an LC-MS run of enriched extracellular vesicles from human plasma using data-independent acquisition on a Thermo Scientific Orbitrap Eclipse. **a**, An ion map of the MS1 peptide signal separated in the retention time and m/z dimensions. The red dashed line indicates the location of the spectrum in **b**. **b**, Selection of a single MS1 spectrum collected at 57.67 min; blue m/z values have been assigned a peptide sequence and red m/z values are unassigned in the analysis. **c**, A total ion current (TIC) plot of the signal intensity from **a** at all time points. The TIC signal is plotted in black, and the blue represents the fraction of the MS1 signal (for example, in **b**) that has been confidently assigned to peptide sequences. The y axis represents an approximation of counts (ions per second). The insert

is a histogram counting distinct molecular entities (features) for different measured intensities. The gray bars of the insert represent all molecular features and the blue represents those assigned a peptide label. The data were analyzed only for unmodified and fully tryptic peptides from the canonical human protein sequences obtained from Uniprot. **d**, Representation of the same data plotted in **c** but with the y axis of each spectrum adjusted to an estimate of ions by multiplying the counts by the Orbitrap fill time. The variable fill times allow peptides with relatively low abundance near 20–30 min to be measured with a similar number of ions as the most abundant peptides in the analysis. The result is billions of peptide ions counted in just 90 min. Data available at https://panoramaweb.org/Single_Molecule_Counting.url and under PXD035637. Code available at https://github.com/uw-maccosslab/single_molecule_counting.

normalized signal for each spectrum that is analogous to normalizing the counts obtained between flow cells in a single-molecule counting experiment³².

LC-MS has a much greater dynamic range than would be expected from simply counting the billions of ions and assigning the counts to peptides. This increase in dynamic range arises because LC-MS first chromatographically separates peptides on the basis of their physical properties so that peptides of the same sequence are measured

together (Fig. 3). This strategy of counting the same peptide sequences together to provide a quantity is effectively a compression scheme for counting molecules. Using gas-phase methods, MS can further improve the dynamic range by measuring the m/z of all peptides and fragments with the same values together. Thus, the effect of highly abundant peptides on the measurement of low-abundance peptides is minimized because they are measured separately and, in some experiment types, in separate trap fills (that is, analogous to measuring

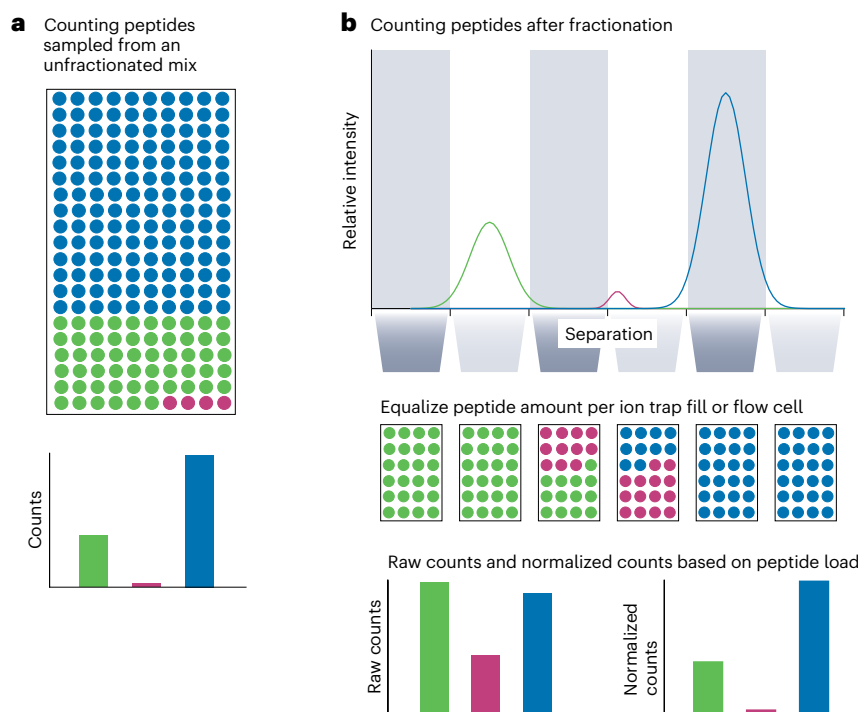


Fig. 3 | Fractionation before counting molecules improves the dynamic range in proteomics. **a**, The dynamic range problem of the human proteome is far more extreme than that of the transcriptome. The enormous dynamic range of peptide abundances requires massive oversampling of the most abundant peptide (blue) to obtain counts for the least abundant peptide (red) **b**, LC-MS separates peptides biochemically, ionizes them and samples the peptides at

different times and with different spectra. While a mass spectrometer works in the gas phase, it is analogous to separating peptides or proteins before counting and then applying normalization to make the quantities comparable between spectra or flow cells. This strategy improves the counting statistics of low-abundance molecules in the presence of high-abundance molecules.

abundant transcripts in different flow cells from low-abundance transcripts).

The mass spectrometry community has capitalized on this strategy to improve the detection and precision of low-abundance molecules^{7,33–35} in the presence of analytes with much greater abundance. Because the timescale of this measurement is fast (sub-second), MS can analyze such compressed groups of ions (~ 10 to 10^6 ion copies at a time) tens of thousands of times per hour. For example, a 90-min LC-MS/MS analysis of peptides in plasma typically measures 3×10^9 ions from just the unfragmented MS1 signal. Yet this frequently represents peptides from only ~ 350 – 450 proteins because the dynamic range of the plasma proteome is notoriously large³⁶. Thus, if plasma is analyzed using a single flow cell with 1 million single-molecule reads, $\sim 950,000$ of those reads will be of the 12 most abundant proteins²⁵, leaving only 50,000 (or 5%) of the remaining reads to quantify the rest of the proteins in the sample. The dynamic range of plasma can be mitigated by depleting the most abundant proteins by immunoaffinity subtraction chromatography³⁷. Such chromatography frequently removes 14 of the most abundant proteins in human plasma (for example, albumin, immunoglobulin G, antitrypsin, immunoglobulin A, transferrin, haptoglobin, fibrinogen, $\alpha 2$ -macroglobulin, $\alpha 1$ -acid glycoprotein, immunoglobulin M, apolipoprotein A1, apolipoprotein A2, complement C3 and transthyretin). Depletion increases the number of detected proteins, but these affinity columns are species specific and thus are largely limited to use with human samples. These

columns also capture the entire complex and binding proteins of the target antigens, removing unintended proteins. For example, patients with cancer make autoantibodies to known cancer biomarkers³⁸ (for example, thyroglobulin, MUC16 (CA125) and prostate-specific antigen), which complicate their analysis using immunoaffinity methods³⁹, and depletion of immunoglobulin G can remove these biomarkers. Depletion of apolipoprotein A1 will also deplete high-density lipoprotein particles⁴⁰, a promising plasma subproteome for the diagnosis of coronary artery disease⁴¹. Such unintentional depletions contribute to biases and complicate the interpretation of the proteomic results.

Figure 2 illustrates the analysis of an extracellular vesicle fraction enriched from plasma, digested using trypsin and measured by data-independent acquisition with an Orbitrap Eclipse instrument. This sample has a lower dynamic range than the whole plasma proteome, making it an interesting avenue for biomarker discovery. The plasma vesicle fraction represents about 1–2% of the plasma proteome, is enriched in tissue-derived proteins, and is depleted in abundant plasma proteins. The total ion current from just the MS1 signal was $>10^{12}$ ions per second, of which 46.4% could be assigned to a peptide sequence using the fragment ion data. This current represented >5 billion ions, of which 1.2 billion ions (24.1%) – not counting the ions measured in the MS/MS spectra – were assigned to peptide sequences. To perform similarly, single-molecule methods like Illumina would analogously need to collect billions of reads from a mixture biochemically separated

into thousands of individual samples (~1 million reads per sample; Fig. 3b). The signal is normalized between flow cells to achieve counts that can be comparable between flow cells, with ~24% of the reads being able to be mapped back to the reference genome. This plasma extracellular vesicle analysis was not sample limited and thus represents an analysis near the upper end of what can be achieved for the analysis of ions per analysis time.

Assuming that emerging polypeptide counting methods can achieve the current throughput of Illumina NovaSeq for DNA (4 billion reads for US\$10,000), their cost for analyzing a mammalian proteome would be much higher than the cost by MS analysis. This also suggests that single-molecule counting approaches must be at least 20-fold cheaper than Illumina sequencing to be cost effective when compared with US\$500 per LC-MS/MS analysis. Stated another way, LC-MS/MS is currently more efficient at counting peptides than next-generation sequencing is at counting oligonucleotides.

Scalability: the elephant in the single-molecule room

The sheer volume of protein molecules in a cell prompts a reality check: will single-molecule methods alone reach the required throughput to sufficiently sample the proteome? For single-molecule counting methods to have the same coverage and breadth of the proteome as they do the transcriptome, they will need to have 10,000 to 30,000 times as many reads of similar quality as currently generated by RNA sequencing (RNA-seq). Thus, protein single-molecule counting-based methods will require technological advancements that greatly exceed the capabilities of nucleotide single-molecule counting methods.

A major factor that limits imaging-based single-molecule sequencing is the density at which the molecules can be spaced and the imaging strategies used to count the spatially resolved reads (we are assuming a 2D imaging plane in this discussion). The limit for the spatial density is constrained by the wavelength of light. Using fluorescence detection, the emission spectrum is in the 250–700 nm range (the actual theoretical resolution limit is about half the wavelength emitted). This provides a practical upper limit on planar molecular density of $\sim 1 \mu\text{m}^2$. Thus, assuming perfect measurement of reads and ideal spatial placement, we can estimate the best-case scenario for the minimal flow cell area versus number of reads: for 1 million reads, 1 mm^2 ; 100 million reads, 1 cm^2 ; 10 billion reads, 10 cm^2 ; 1 trillion reads, 1 m^2 . The area that needs to be imaged is limited by microscopy. These limits can be relaxed by super-resolution imaging, but at the expense of decreased imaging speeds. Even with advances in widefield microscopy, there is a compromise between the field of view and the measured pixel size using a given charge-coupled device (CCD) detector.

These estimates explain why obtaining 10 billion nucleotide reads is time consuming and expensive for single-cell RNA-seq analysis. Thus, the throughput of current nucleotide sequencing methods falls short of achieving the 400 billion reads needed for a full proteome analysis of a bulk sample at a similar coverage to that currently achieved on the transcriptome by RNA-seq.

Methods analyzing intact protein molecules, such as top-down MS⁴² or single-molecule methods that aim to count proteins²⁰, may be able to sample the proteome with fewer total counts. This is in contrast to peptide approaches, which usually count multiple unique peptide sequences per proteoform. The difference between measuring intact proteoforms and peptides from the digestion of complex mixtures is analogous to the differences between short-read RNA-seq and long-read isoform sequencing. Intact protein analysis is further

aided by recent methods for charge detection mass spectrometry, in which individual ion events can be measured^{43,44}.

A look at some alternative advanced single-molecule methods suggests a huge gap in throughput. The Pacific Biosciences *Sequel II* platform for genome sequencing can handle, at best, 10^7 molecules in each sequencing run, which takes a couple of days to complete. The highest throughput Oxford Nanopore Technology (ONT) platform, the PromethION, can run up to 48 flow cells at a time, providing an approximate maximum throughput of 5×10^7 molecules per run, which takes 1–2 days for data acquisition (signal processing time not included). These two examples are the most sophisticated single-molecule analyzers, and yet the throughput offered is significantly short of the required throughput for analyzing protein mixtures on par with LC-MS. The high limit of 5×10^7 for single molecule technologies is no coincidence: these limitations are governed by physical limitations in scaling up device architecture for single-molecule interrogation, limitations in molecular turnover in the devices, as well as limitations in data acquisition and transfer rates.

Taking the ONT pore sequencer and direct RNA sequencing as an example, 500 ng of input RNA contains about 10^{12} mRNA molecules, and only 10^6 of these are sampled in a MinION nanopore-based flow cell. The vast discrepancy between input requirements and actual molecules analyzed (only 1 p.p.m. is sampled) is a testament to the intertwined limitations of single-molecule technologies: 500 ng ensures that molecules arrive to a nanoscale detector with minimal off-times, or else the sensor will be mostly vacant and throughput will be compromised. In addition, the speed at which molecules pass through the pores cannot be too fast (typically 100 nm of polymer contour length per second) because the maximum measurement bandwidth of the electrical signal recording cannot exceed a few kilohertz owing to data transfer speed and signal-to-noise limitations.

These multiple constraints have set natural limits on single-molecule processing, but there is no inherent reason for these to be hard limits. As flow cells are improved to enable analyses from smaller sample volumes and/or strategies to deliver molecules more efficiently to the pores rather than rely on diffusion, one may imagine over 100-fold reductions in input requirements from $>100 \text{ ng}$ to $<1 \text{ ng}$, at similar throughputs. Similarly, if one were to assume that data transfer and bandwidths will increase by ~100 fold over the next 5–7 years, one may expect transitioning from 10^3 pores in a flow cell to 10^5 , which would boost the throughput 100-fold to about 5×10^9 molecules per run (1–2 days). We estimate that these limitations will have to be overcome before single-molecule proteomics can be approached at scale.

What limits LC-MS/MS and can the technology improve to sample the proteome?

Most MS proteomics methods use a bottom-up strategy of digesting proteins to peptides to overcome the enormous physicochemical diversity of proteins in the cell. Overwhelmingly, these methods make use of trypsin, which produces peptides from proteins that have good cleavage specificity, are well suited for reversed phase separations, produce mostly doubly and triply charged peptides, and fragment well because of the localization of a basic C-terminal residue and presence of a mobile proton. That said, not all tryptic peptides are well suited for LC-MS/MS, and because of this, proteins in complex mixtures are mainly identified through partial sequences. The sequence coverage of an identified protein varies between 10% and 100% (on average 30–50%), depending on the protein and the experiment. One approach to mitigate this limitation and maximize protein sequence coverage is to combine the

results from different proteases with different specificities^{45,46}. However, the increased sampling of ions derived from redundant peptides from the same proteins, while useful for improving coverage, comes at the expense of dynamic range as more ions must be sampled from more peptides from abundant proteins before sampling ions from rare molecular species. To overcome the dynamic range problem, alternative methods have been developed to minimize peptide coverage, capturing or depleting a subset of the peptides while maximizing the different proteins sampled. This is analogous to exon capture⁴⁷, ChIP⁴⁸ or similar methods used in genomics before single-molecule sequencing. Thus, there is a balance between maximizing coverage of individual proteins and the dynamic range of the proteins measured.

The main limiting factor in the sensitivity of LC-MS/MS methods is the electrospray process, which turns peptide molecules in solution into gas-phase ions³. If a molecule is not converted to a gas-phase ion, it cannot be quantified with a mass spectrometer. Using electrospray, MS methods can quantify proteins present at 5,000–20,000 copies in the context of complex mammalian proteomes^{9,49}. The number of ions sampled may be increased by using methods like multidimensional chromatography¹¹ or making multiple analyses using different portions of the mass range⁵⁰. These approaches can improve the depth of proteome coverage, but at the expense of increased analysis time and throughput. A sixfold increase in time may increase the number of peptides that can be measured by only twofold because the increased time is at least partially redundant with the peptides measured in previous fractions. Ultimately, this comes at the expense of protein input material and reduces the number of samples that can be measured. Thus, a primary challenge is to achieve deep proteome coverage with smaller samples, such as single cells, and analyze them faster, thus enabling higher throughput^{51,52}.

Another way to improve LC-MS/MS is in the more efficient use of the ions that are generated. In most data-independent acquisition methods, a single wide m/z range is isolated at once and the rest of the ion beam that is not isolated is lost. Data-dependent acquisition methods sample an even smaller fraction of the ion beam. With bulk samples, this means that only $\sim 1/50$ th of the ion beam is being used as only 1 of 50 precursor windows is measured at once⁵³. With single-cell samples, three or four windows are used and thus about one-third of all ions available to the MS instrument are analyzed¹², at the expense of limiting within-spectrum selectivity. Methods like diaPASEF (parallel accumulation–serial fragmentation combined with data-independent acquisition) offer potential to increase the sampling of the peptide ion beam.

Another important way to advance LC-MS/MS is to improve the computational methods that are used to assign peptide sequences to the ion current that is measured. Currently only ~ 15 – 50% of the measured ion current is assigned to peptide sequences⁵⁴. Thus, an improvement in both the physical instrumentation for enhancing the sampling of the ion beam and computational methods for enhanced data interpretation could see a 50- to 75-fold improvement in the number of ions counted before LC-MS/MS becomes limited by the electrospray process. This improvement in ion counts will improve the relative measurement precision of the peptides measured, elevate low abundance species to within the limit of detection, and enable measurements to be made in shorter time and with less material. We expect innovations in data acquisition and interpretation to enable quantification and sequence identification for a large fraction of the tens of thousands of peptide-like features detected in single cells, and thus substantially increase the depth of proteome coverage⁵⁴.

What can emerging single-molecule counting methods adopt from LC-MS/MS?

Peptide quantification using LC-MS has evolved over the last several decades in ways that have improved our analyses of complex protein mixtures. Peptide ions are not counted one at a time but are aggregated, effectively compressing the signal from many peptide ions into a single measurement. This compression reduces time and minimizes the effect of abundant peptides on the counting precision of low-abundance peptides, improving the dynamic range (Fig. 3). However, the emphasis on generating and sorting ions of like character constrains the choice of enzymes to produce peptides ideally suited for the respective method. Because tryptic peptides are ideally suited for LC-MS/MS does not mean they will be ideally suited for other methods. The conundrum is that reducing the bias by adding more distinct enzymes or nonspecific enzymes leads to more peptides with different sequences for each protein, making it even harder to sample low-abundance proteins in the presence of abundant proteins. Put simply, approaches to reduce these biases and increase sequence coverage in proteomics could push the field toward counting more ions from different peptide species – exacerbating the counting problem. Understanding the strengths and weaknesses of LC-MS as it has approached complex proteomes may perhaps constructively guide the emerging field of single-molecule proteomics. As advice to this budding field, consider the following.

Fractionate. It is better to run many smaller counting experiments on fractionated samples than one very large counting experiment (Fig. 2). If peptides or proteins are separated using an analytical method like liquid chromatography, electrophoresis or affinity capture, the less abundant molecules will be enriched in certain fractions, resulting in a better representation of these peptides in the downstream detection and quantification processes. To make optimal use of this separation, methods equivalent to AGC, as done with ion trap instruments⁵⁵, will need to be developed so that uniform fractions are fed into the flow cell for single-molecule readout. For example, each biochemical fraction can be diluted to the same concentration and equal quantities of the fractions loaded into many flow cells.

In addition to improving the dynamic range of the measurement, the use of a separation method based on a physicochemical property can improve the sequence determination of the peptide or protein. In LC-MS/MS, the use of either predicted retention time or previously measured retention time is a powerful feature for the discrimination of correct and incorrect peptide detections^{56–58}. This minimizes the false discovery rate and improves sensitivity. Indeed, nanopore proteomics methods are making first steps in this direction⁵⁹.

The measurement of a signal across many points during a chromatographic separation also enables the integration of a chromatographic peak. Despite the unparalleled selectivity of LC-MS/MS measurements, there is often a background signal that complicates the quantitative linearity of the measurements. By integrating the peak along the separation, it is possible to subtract background, which improves quantitative accuracy.

When there are many molecules to count, you will need to count many at a time. As mentioned above, to measure peptides using mass spectrometry from many billions of ions, it became impractical to count ions one at a time in a realistic timescale. When done in a flow cell, single-molecule counting methods will have to count so many molecules that they will likely either exceed the density of the flow cell or require a flow cell or cells with impractical physical dimensions. We hope to

inspire new methods that are analogous to the switch in mass spectrometry from pulse counting (single molecule) to ion current measurement (each read will contain a variable quantity of many counts).

Overcoming biases. Arguably the most challenging aspect of proteomics is the massive physicochemical diversity of proteins in the cell. To overcome this vast diversity in solubility, size, post-translational modifications, ionization and fragmentation by mass spectrometry, presence of autoantibodies, embedding of domains in membranes, or protein–protein interactions, most proteomics experiments take a bottom-up strategy for the analysis of complex mixtures by digesting proteins to peptides before analysis. Performing analyses at the peptide level greatly simplifies the physicochemical diversity of the analytes. In general, tryptic peptides are well matched for reversed phase chromatography, electrospray ionization and tandem mass spectrometry. Methods for top-down proteomics have advanced enormously and have opened the door to characterizing proteoforms that are often ignored in understanding the function of the cell, but these methods have greater constraints in their ability to analyze proteins with extremes in physicochemical properties⁶⁰.

Over the last two decades there have been massive improvements in nanoflow separations, electrospray ionization, transmission of ions from atmospheric pressure to vacuum, tandem mass spectrometry and pipelined data acquisition that have resulted in sensitivities now approaching 10–50 zmol for peptides. However, a challenge for single-cell and low-input proteomics is the adsorption of proteins and peptides to surfaces. In general, the sensitivity limits of proteomics samples have been because of not LC–MS/MS itself but the loss of sample to surfaces before entering the system. To solve these problems, there have been methods developed specifically to improve the recovery of protein from small numbers of cells using many strategies, including one-pot digestion^{61,62}, massively parallel sample preparation in surface droplets⁶³, addition of carrier proteins⁶⁴, and barcoding and combining of samples using mass tags to spread losses between many samples¹².

Despite the potential sensitivity of emerging single-molecule counting methods, these will need to overcome the same biochemical challenges of analyzing intact proteins, adsorptive losses to surfaces, variable enzyme digestion kinetics, and biases against certain peptide properties. While biases for sequencing peptides and proteins in flow cells and nanopores will almost certainly be different than those of LC–MS/MS, the strategies for improving the recovery of peptides for entry into the instrument will largely be the same.

Sample multiplexing. Peptides from multiple samples can be bar-coded (for example, by covalent chemical labels), subsequently mixed, and analyzed simultaneously. Sample multiplexing has helped increase the throughput of MS proteomics^{8,65}. Analogous multiplexing methods are likely to be implemented by single-molecule methods to increase the number of samples analyzed, as multiplexing is a powerful feature of single-molecule DNA sequencing. Yet multiplexing with single-molecule approaches spreads the counted molecules between many samples and thus reduces the number of molecules counted per sample, which results in shallower depth of proteome coverage and sequence completeness.

Instrument companies focus on the bottom line before science. It is also important for new methods to have a clear fiscal return on investment. A couple of the new single-molecule protein sequencing


methods hope to convert peptide or protein sequences into DNA barcodes that can then be analyzed with traditional next-generation sequencing technology⁶⁶. However, as discussed above, the large number of protein molecules will require sequencing billions of molecules to obtain the coverage of the proteome offered by LC–MS/MS²². Because this coverage can be obtained for -US\$500 per analysis by LC–MS/MS and sequencing billions of DNA reads can cost -US\$10,000, it would require next-generation sequencing companies to reduce their costs to ~5% their current rates to be competitive. Without separation, a proteomic technology needs to count with high specificity about 1 billion intact protein molecules (or 20 billion peptides) for US\$500 (including personnel, sample handling and analysis) to disrupt current LC–MS technologies. This price reduction would be a game changer for DNA sequencing and would further revolutionize genomics. However, it would require DNA sequencing companies to reduce their income from genomics applications to be financially competitive in the proteomics market. If they do this, then they will have done something that is rarely done in the proteomics field: minimize the financial return of existing products to be competitive in new high-risk areas.

Summary

Here we provided a perspective on the potential and challenges of scaling the use of single-molecule counting methods to the analysis of the proteome. We use LC–MS-based proteomics as a comparison by illustrating how many peptide molecules are counted in the gas phase using standard mass spectrometry methods. This comparison will serve as a benchmark for single-molecule counting methods to obtain parity with LC–MS data. Analyzing the proteome by counting single peptide or protein molecules in a spatially resolved flow cell represents significant challenges over counting nucleotides because of both the physicochemical complexity of proteins and the sheer number of proteins in the cell. To support innovation around these emerging methods, we provide some lessons learned by the LC–MS/MS based proteomics community.

Michael J. MacCoss ^{1,8} , **Javier Antonio Alfaro** ^{2,3,4,8} , **Danielle A. Faivre** ¹, **Christine C. Wu**¹, **Meni Wanunu** ⁵ & **Nikolai Slavov** ^{6,7,8} 

¹Department of Genome Sciences, University of Washington, Seattle, WA, USA. ²International Centre for Cancer Vaccine Science, University of Gdańsk, Gdańsk, Poland. ³Department of Biochemistry and Microbiology, University of Victoria, Victoria, British Columbia, Canada. ⁴School of Informatics, University of Edinburgh, Edinburgh, UK. ⁵Department of Physics, Northeastern University, Boston, MA, USA. ⁶Departments of Bioengineering, Biology, Chemistry and Chemical Biology, Single Cell Proteomics Center and Barnett Institute, Northeastern University, Boston, MA, USA. ⁷Parallel Squared Technology Institute, Watertown, MA, USA. ⁸These authors contributed equally: Michael J. MacCoss, Javier Antonio Alfaro, Nikolai Slavov.

 e-mail: maccoss@uw.edu; Javier.Alfaro@proteogenomics.ca; nslavov@northeastern.edu

Published online: 10 March 2023

References

1. Nau, H. & Biemann, K. *Anal. Biochem.* **73**, 139–153 (1976).
2. Hass, G. M. et al. *Biochemistry* **14**, 1334–1342 (1975).

3. Hunt, D. F., Yates, J. R. III, Shabanowitz, J., Winston, S. & Hauer, C. R. *Proc. Natl Acad. Sci. USA* **83**, 6233–6237 (1986).
4. Johnson, R. S. & Biemann, K. *Biochemistry* **26**, 1209–1214 (1987).
5. Yamashita, M. & Fenn, J. B. *J. Phys. Chem.* **88**, 4451–4459 (1984).
6. Eng, J. K., McCormack, A. L. & Yates, J. R. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
7. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. *Nat. Methods* **1**, 39–45 (2004).
8. Ross, P. L. et al. *Mol. Cell Proteomics* **3**, 1154–1169 (2004).
9. Specht, H. et al. *Genome Biol.* **22**, 50 (2021).
10. Petelski, A. A. et al. *Nat. Protoc.* **16**, 5398–5425 (2021).
11. Washburn, M. P., Wolters, D. & Yates, J. R. III *Nat. Biotechnol.* **19**, 242–247 (2001).
12. Derks, J. et al. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-022-01389-w> (2022).
13. Messner, C. B. et al. *Cell Syst.* **11**, 11–24.e4 (2020).
14. Alfaro, J. A. et al. *Nat. Methods* **18**, 604–617 (2021).
15. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. *PLOS Comput. Biol.* **11**, e1004080 (2015).
16. Palmblad, M. *J. Proteome Res.* **20**, 3395–3399 (2021).
17. Swaminathan, J. et al. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.4278> (2018).
18. Reed, B. D. et al. *Science* **378**, 186–192 (2022).
19. Mallick, P. Methods of assaying proteins. US Patent 10948488B2 (2021).
20. Egertson, J. D. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.10.11.463967> (2021).
21. Brinkerhoff, H., Kang, A. S. W., Liu, J., Aksimentiev, A. & Dekker, C. *Science* **374**, 1509–1513 (2021).
22. Slavov, N. *Cell* **185**, 232–234 (2022).
23. Milo, R. *Bioessays* **35**, 1050–1055 (2013).
24. Bekker-Jensen, D. B. et al. *Cell Syst.* **4**, 587–599.e4 (2017).
25. Anderson, N. L. & Anderson, N. G. *Mol. Cell. Proteomics* **1**, 845–867 (2002).
26. Marinov, G. K. et al. *Genome Res.* **24**, 496–510 (2014).
27. Peterson, D. W. & Hayes, J. M. Signal-to-noise ratios in mass spectroscopic ion-current-measurement systems. In *Contemporary Topics in Analytical and Clinical Chemistry* Vol. 3 (eds. Hercules, D. M. et al.) 217–252 (Springer US, 1978).
28. Scigelova, M., Hornshaw, M., Giannakopoulos, A. & Makarov, A. *Mol. Cell. Proteomics* **10**, M111.009431 (2011).
29. Makarov, A. & Denisov, E. *J. Am. Soc. Mass Spectrom.* **20**, 1486–1495 (2009).
30. MacCoss, M. J., Toth, M. J. & Matthews, D. E. *Anal. Chem.* **73**, 2976–2984 (2001).
31. Schwartz, J. C., Zhou, X.-G. & Bier, M. E. Method and apparatus of increasing dynamic range and sensitivity of a mass spectrometer. US Patent 5572022A (1996).
32. Zhao, S., Ye, Z. & Stanton, R. *RNA* **26**, 903–909 (2020).
33. Belov, M. E. et al. *Anal. Chem.* **73**, 5052–5060 (2001).
34. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. *Nat. Methods* **15**, 440–448 (2018).
35. Egertson, J. D. et al. *Nat. Methods* **10**, 744–746 (2013).
36. Anderson, N. L. et al. *Mol. Cell. Proteomics* **3**, 311–326 (2004).
37. Pieper, R. et al. *Proteomics* **3**, 422–432 (2003).
38. Macdonald, I. K., Parsy-Kowalska, C. B. & Chapman, C. *J. Trends Cancer Res.* **3**, 198–213 (2017).
39. Hoofnagle, A. N. & Wener, M. H. *J. Immunol. Methods* **347**, 3–11 (2009).
40. McVicar, J. P., Kunitake, S. T., Hamilton, R. L. & Kane, J. P. *Proc. Natl Acad. Sci. USA* **81**, 1356–1360 (1984).
41. Heinecke, J. W. *J. Lipid Res.* **50** (Suppl.), S167–S171 (2009).
42. Siuti, N. & Kelleher, N. L. *Nat. Methods* **4**, 817–821 (2007).
43. Kafader, J. O. et al. *Nat. Methods* **17**, 391–394 (2020).
44. Wörner, T. P. et al. *Nat. Methods* **17**, 395–398 (2020).
45. Gatlin, C. L., Eng, J. K., Cross, S. T., Detter, J. C. & Yates, J. R. III *Anal. Chem.* **72**, 757–763 (2000).
46. MacCoss, M. J. et al. *Proc. Natl Acad. Sci. USA* **99**, 7900–7905 (2002).
47. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. *Nat. Methods* **6**, 315–316 (2009).
48. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. *Science* **316**, 1497–1502 (2007).
49. Huffman, R. G. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.03.16.484655> (2022).
50. Panchaud, A. et al. *Anal. Chem.* **81**, 6481–6488 (2009).
51. Slavov, N. *Nat. Biotechnol.* **39**, 809–810 (2021).
52. Derks, J. & Slavov, N. *J. Proteome Res.* <https://doi.org/10.1021/acs.jproteome.2c00721> (2023).
53. Pino, L. K., Just, S. C., MacCoss, M. J. & Searle, B. C. *Mol. Cell. Proteomics* **19**, 1088–1103 (2020).
54. Slavov, N. *J. Proteome Res.* **20**, 4915–4918 (2021).
55. Schwartz, J. C. & Kovtoun, V. V. Automatic gain control (AGC) method for an ion trap and a temporally non-uniform ion beam. US Patent 7960690B2 (2011).
56. Klammer, A. A., Yi, X., MacCoss, M. J. & Noble, W. S. *Anal. Chem.* **79**, 6111–6118 (2007).
57. Searle, B. C. et al. *Nat. Commun.* **9**, 5128 (2018).
58. Chen, A. T., Franks, A. & Slavov, N. *PLOS Comput. Biol.* **15**, e1007082 (2019).
59. Zrehen, A., Ohayon, S., Huttner, D. & Meller, A. *Sci. Rep.* **10**, 15313 (2020).
60. Donnelly, D. P. et al. *Nat. Methods* **16**, 587–594 (2019).
61. Zhu, Y. et al. *Nat. Commun.* **9**, 882 (2018).
62. Specht, H. et al. Preprint at *bioRxiv* <https://doi.org/10.1101/399774> (2018).
63. Leduc, A., Huffman, R. G., Cantlon, J., Khan, S. & Slavov, N. *Genome Biol.* **23**, 261 (2022).
64. Budnik, B., Levy, E., Harmange, G. & Slavov, N. *Genome Biol.* **19**, 161 (2018).
65. Slavov, N. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-022-01411-1> (2022).
66. Hong, J. M. et al. *iScience* **25**, 103586 (2021).
67. Hagemann-Jensen, M. et al. *Nat. Biotechnol.* **38**, 708–714 (2020).

Acknowledgements

The authors acknowledge discussions with Edward Marcotte and members of the Alfaro, MacCoss and Slavov labs. M.J.M. appreciates the constructive feedback provided by UW Genome Sciences faculty. This work was supported in part by US National Institutes of Health grants U19 AG065156, R24 GM141156 and F31 AG066318; an Allen Distinguished Investigator award through The Paul G. Allen Frontiers Group to N.S.; a Seed Networks Award from CZI CZF2019-002424 to N.S.; award R01GM144967 from the US National Institute of General Medical Sciences; award R01HG10087 from the US National Human Genome Research Institute to M.W.; and the project 'International Centre for Cancer Vaccine Science' carried out within the International Agendas Programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund. We thank the PL-Grid and CI-TASK Infrastructure, Poland, for providing their hardware and software resources. This work is supported by the Knowledge At the Tip of Your Fingers: Clinical Knowledge for Humanity (KATY) project funded from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101017453.

Author contributions

M.J.M., J.A. and N.S. conceived the project and wrote an initial draft. D.A.F., M.J.M. and N.S. made figures. C.C.W. collected the plasma extracellular vesicle data. All authors contributed significantly to the content, edited, and approved the final manuscript.

Competing interests

The MacCoss laboratory at the University of Washington has a sponsored research agreement with Thermo Fisher Scientific, a manufacturer of mass spectrometry instrumentation. M.J.M. is a paid consultant for Thermo Fisher Scientific. The Slavov laboratory at Northeastern University has a research agreement with Bruker, a manufacturer of mass spectrometry instrumentation.

Additional information

Peer review information *Nature Methods* thanks Tae-Young Yoon and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.