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

Michael J. MacCoss[1,#], Javier Alfaro[2,3,#], Meni Wanunu[4], Danielle A. Faivre[5], & Nikolai Slavov[6,#]

1. Department of Genome Sciences, University of Washington, Seattle, WA 98117, USA 0000-0003-1853-0256 maccoss@uw.edu
2. International Centre for Cancer Vaccine Science, University of Gdańsk, Gdańsk, Poland. javier.alfaro@proteogenomics.ca
3. Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada
4. Department of Physics, Northeastern University, Boston, MA 02115, USA 0000-0002-9837-0004 wanunu@neu.edu
5. Department of Genome Sciences, University of Washington, Seattle, WA 98117, USA dfaivre@uw.edu
6. Bioengineering Department and Barnett Institute of Chemical and Biological Analysis, Northeastern University, Boston, MA 02115, USA 0000-0003-2035-1820 nslavov@alumni.princeton.edu

# Contributed Equally

Abstract

Mammalian cells have about 30,000-fold more protein molecules than mRNA molecules. This larger number of molecules and the associated larger dynamic range have major implications in the application of proteomics technologies. We examine these implications for both liquid chromatography-tandem mass-spectrometry (LC-MS/MS) and single-molecule counting and provide estimates on how many molecules are routinely measured in proteomics experiments by LC-MS/MS. We review strategies that have been helpful for counting billions of protein molecules by LC-MS/MS and suggest that these strategies can benefit single-molecule methods, especially in mitigating the challenges of the wide dynamic range of the proteome. We also examine the theoretical possibilities for scaling up single-molecule and mass-spectrometry proteomics approaches to quantifying the billions of protein molecules that make up the proteomes of our cells.

Introduction

The ubiquitous roles of proteins in biomedicine are well appreciated, and have motivated technologies seeking to advance 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[1–3]. Since then, milestones in MS-based proteomics have included *de novo* sequencing entire proteins in the late 1980s[4–7], soft ionization by electrospray[8], automated spectral interpretation[9], multiplexing the acquisition of spectra on different peptides using data independent acquisition[10], multiplexing the acquisition of different samples using tandem mass tags[11], and quantifying thousands of proteins in single human cells[12,13]. 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[14] to about 25,000 protein data points per hour[15]. This increased throughput has been critical for addressing challenges in biomedical research[16]. 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 towards identifying and potentially sequencing single polypeptide molecules[17–19]. 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 degradation[20] or amino peptidases[21]. Another approach aims to use degenerate affinity reagents to recognize individual protein molecules separated spatially in a flow cell[22,23]. Other groups are working to adapt nanopore sequencing to peptides and proteins[24,25]. 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 significant enthusiasm within the scientific community as a complement to MS analysis[17].

These developments have motivated renewed interest and investment in advancing proteomics technologies, as reflected in private funding[17] and in recent National Human Genome Research Institute (NHGRI) funding opportunities aimed at accelerating the development of technologies for single-molecule sequencing and single-cell proteome analysis. Because there is excitement for new emerging single-molecule counting methods for proteomics, we felt it was timely to provide a perspective that compares strategies used by the current state-of-the-art proteomics methods based on liquid chromatography-tandem mass spectrometry (LC-MS/MS) to the new and emerging counting-based-methods that enabled complete and accurate transcriptome sequencing. We hope our opinion will provide benchmarks and directions for the technological breakthroughs that need to be achieved for single molecule protein/peptide counting to achieve parity and complement the capabilities of LC-MS/MS based proteomics methods. How many molecules need to be counted? How extreme is the dynamic range problem? Do current solutions for handling the dynamic range problem limit the sequence coverage of the proteome? What will new technologies need to accomplish to reach parity with LC-MS/MS, and how will these technologies complement one another? What can these emerging technologies learn from LC-MS/MS based proteomics? These are the questions we aim to address below.

How many molecules need to be counted?

Many of the challenges for accurate and sensitive protein quantification are shared by all proteomics methods, such as the quantification over a wide dynamic range. Indeed, a typical mammalian cell contains billions of protein molecules but less than half a million RNA molecules[26], Figure 1a. Some proteins are present at hundreds of copies per cell while others (e.g., histones) at tens of millions of copies per cell, resulting in about 10^6 dynamic range[27]. The range of protein abundances is even larger for body fluids, such as plasma where protein abundances may differ by 10^{10}, e.g., between albumin and IL-6[28]. This wide range of abundance is a fundamental challenge for proteomics because the presence of abundant proteins make it rare to count molecules from low abundant proteins, such as having to count billions of albumin molecules before having a chance to detect a single IL-6 molecule. This means that an emphasis on highly-sensitive single-molecule approaches that have been used successfully to quantify the transcriptome, which spans about 10^3 dynamic range, face major challenges in scaling to quantifying the proteome[25].

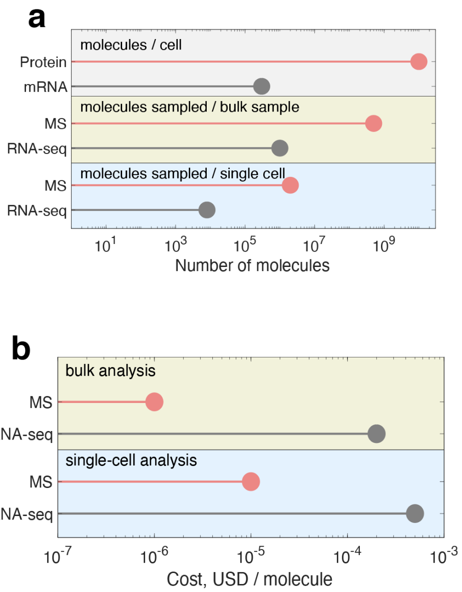

Figure 1 | Overview of RNA and protein statistics. a, A representative human cell, such a fibroblast, has billions of protein molecules compared to merely hundreds of thousands of RNA molecules[29]. Accordingly, MS analysis samples more protein molecules per sample than the RNA molecules sampled by RNA-seq. **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[30] and number of protein molecules counted by plexDIA[15].

A typical mammalian cell, i.e., a HeLa cell with a volume of ~3,000 μm^3, contains about 300,000 mRNA molecules[29] and about 10,000,000,000 protein molecules[26], Figure 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 different independent measurements, it is estimated that the typical HeLa cell contains at least ~3-5 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. These estimates of the relative abundance ratio of mRNA to protein molecules has the direct consequence of requiring about 30,000-fold more counts to characterize the protein molecules at an analogous coverage of what has been achieved with the transcriptome, Figure 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 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 of reading peptide sequences in a spatially parallelized format will be successful[17]. 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 the estimate of \$10,000 for sequencing 4 billion reads by Illumina NovoSeq over ~2-days and \$500 for performing a 2 hour quantitative LC-MS/MS analysis. These costs were chosen as conservative estimates from several academic core facilities. While academic research laboratories may achieve lower costs, these prices represent objective estimates for widely accessible services. The results in Fig. 1b indicate that the cost per protein molecule analyzed by LC-MS/MS is lower than the cost of DNA molecule sequenced by Illumina. 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. The abundance of each analyte is often determined from a background subtracted peak area of 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 multiplier[31] or the generation of an image current in a Fourier transform mass analyzer[32]. The current is a measure of the number of ions (charged molecules) counted normalized for the amount of time spent sampling the signal. The measured signal is proportional to ions/second, and thus, it can be converted into a number of counted ions and compared directly with single molecule counting methods (Figure 2)[12,33,34].

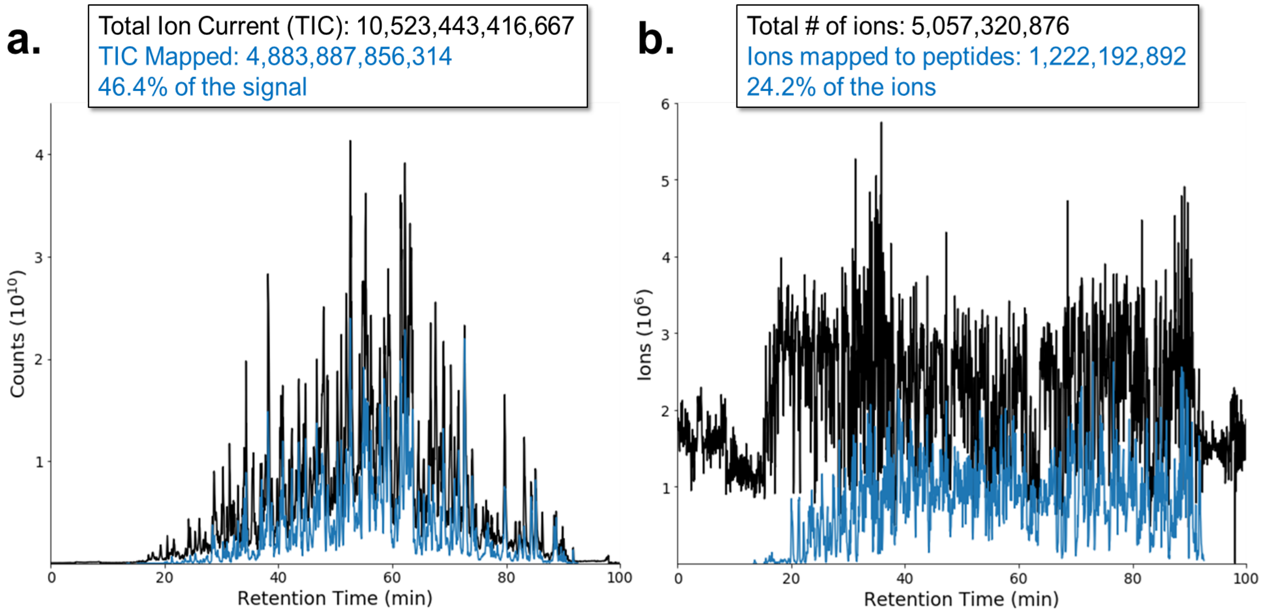

Figure 2 | Signal from the MS1 spectra of an LC-MS run of enriched extracellular vesicles from human plasma using data independent acquisition of an ThermoFisher Eclipse Tribrid. **a**, Represents the total ion current (TIC) in black and the blue represents the fraction of the MS1 signal that can be assigned to peptide sequences using a combination of the MS1, MS2 and chromatographic retention time information. The y-axis represents an approximation of counts (ions per second). **b**, Represents the same data plotted in (A) but with the y-axis of each spectrum adjusted to an estimate of ions by multiplying the counts by the Orbitrap fill time. Even a simple experiment can count billions of peptide ions within 90 min. The variable fill times allow peptides with relatively low abundance near 20-30 min to be measured with similar ion counts as the most abundant peptides in the analysis. Data available at: https://panoramaweb.org/Single_Molecule_Counting.url under PXD035637.

LC-MS/MS methods can improve the sensitivity of low abundant analytes by changing the time spent sampling the signal (aka 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 time.[35]. This dynamic adjustment of the injection time, known as automatic gain control (AGC), provides an ideal ion population for the mass spectrometry measurement. 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 spending more time on less abundant peptides to enable a precise measurement of the weaker signal – increasing the dynamic range (Figure 2). The normalization of each spectrum intensity by the “time” results in a current measurement that is analogous to normalizing the total number of reads in a single molecule counting experiment to compare measurements between flow-cells in genomics[36,37].

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 based on their physical properties so that peptides of the same sequence are measured together (Figure 3). This strategy of counting the same peptide sequences together to provide a quantity is effectively a compression scheme for counting molecules. Additionally, 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 lowly abundant peptides is minimized because they are measured separately and in some experiment types, separate trap fills (i.e. analogous to measuring abundant transcripts in different flow cells from low abundance transcripts). In fact, over the years the mass spectrometry community has capitalized on this strategy to improve the detection and precision of low abundant molecules[10,38–40] in the presence of analytes with much greater abundance. Because the timescale of this measurement is fast (sub 1 second) MS can analyze such compressed groups of ions (~10 to $1x10^6$ ion copies at a time) tens of thousands of times per hour.

Summing up with an example, a 90 min LC-MS/MS analysis of peptides in plasma frequently measures $3x10^9$ ions from just the unfragmented MS1 signal. Yet, this frequently represents only peptides from ~350-450 proteins because the dynamic range of the plasma proteome is notoriously large[41]. Thus if plasma is analyzed using a single flow cell with 1 million single molecule "reads", ~950,000 of those reads would be of the 12 most abundant proteins[28] 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 chromatography[42]. Such chromatography frequently removes 14 of the most abundant proteins in human plasma (e.g. albumin, IgG, antitrypsin, IgA, transferrin, haptoglobin, fibrinogen, alpha2-macroglobulin, alpha1-acid glycoprotein, IgM, apolipoprotein A1, apolipoprotein A2, complement C3 and transthyretin). Depletion substantially increases the number of detected proteins, but unfortunately these affinity columns are species specific and thus are largely limited for use with human samples. These columns also capture the entire complex and binding proteins of the target antigens – removing unintended proteins. For example, albumin binds as many as 35 proteins and the albuminome itself has been proposed as a plausible plasma subfraction of biomedical interest[43]. It is well known that patients with cancer make autoantibodies to known cancer biomarkers[44] (e.g. thyroglobulin, MUC16 (CA125), and PSA) complicating their analysis using immunoaffinity methods[45] and further, depletion of IgGs can remove these biomarkers. Depletion of apolipoprotein A1, will also deplete HDL particles[46] , a promising plasma sub proteome for the diagnosis of coronary artery disease[47]. Such unintentional depletions contribute to biases and complicate the interpretation of the proteomics results.

Figure 2 illustrates the analysis of an extracellular vesicle (EV) fraction enriched from plasma, digested using trypsin and measured by data independent acquisition with an Orbitrap Eclipse. The plasma vesicle fraction represents about 1-2% of the plasma proteome, is enriched in tissue derived proteins, and depleted in abundant plasma proteins. The total ion current (ions per second) from just the MS1 signal was $>10^{10}$, of which 46.4% of the current 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%) were assigned to peptide sequences – not counting the ions measured in the MS/MS spectra. To perform similarly, single-molecule methods like illumina would analogously need to collect billions of reads from a mixture biochemically separated into 1000s of individual flow cells (~1 million reads per flow cell; see Figure 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 EV 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 NovoSeq for DNA (4 billion reads for $10,000), their cost for analyzing a mammalian proteome would be much higher than the cost for MS analysis. This also suggests that single-molecule counting approaches must be at least 20x cheaper than Illumina sequencing to be cost effective when compared with $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.

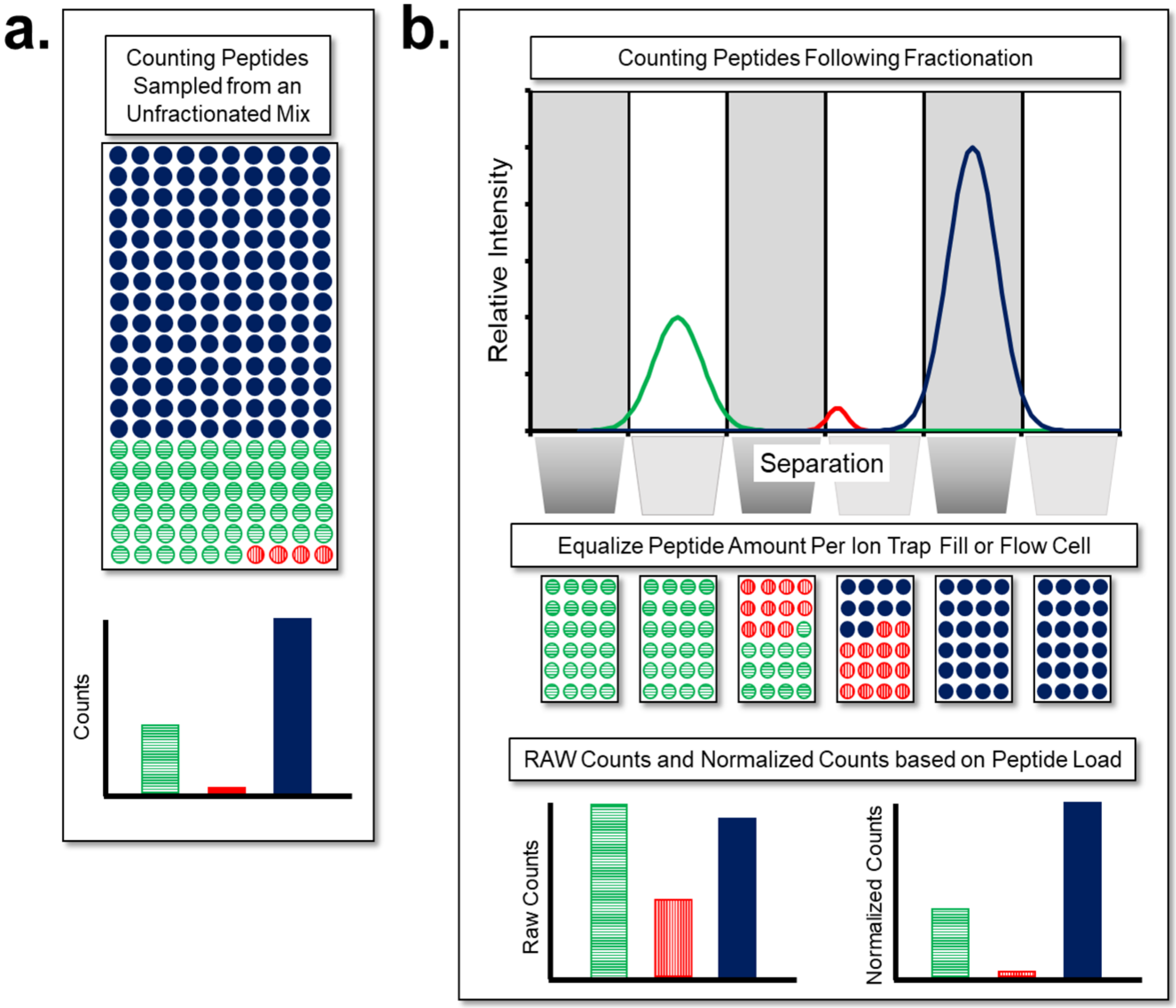
a.
Counting Peptides Sampled from an Unfractionated Mix
Counts
b.
Counting Peptides Following Fractionation
Relative Intensity
Separation
Equalize Peptide Amount Per Ion Trap Fill or Flow Cell
RAW Counts and Normalized Counts based on Peptide Load
Raw Counts
Normalized Counts

Figure 3 | Fractionation prior to counting molecules improves the dynamic range in proteomics. a, The dynamic range problem of the human proteome is far more extreme than that of transcriptomics. The enormous dynamic range of peptide abundances requires massive sampling of the most abundant peptide (solid blue) to obtain counts for the least abundant peptide (vertical stripe red) **b,** Instead**,** 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/proteins prior to counting and applying normalization to make the quantities comparable between spectra or flow cells. This strategy significantly improves the counting statistics of low abundance molecules in the presence of high abundance molecules. In ion trap mass spectrometry, the normalization approach to optimize ions in each spectrum and adjust the signal by the variable fill time is known as automatic gain control (AGC).

Scalability: the big elephant in the room with single-molecule methods

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 10,000-30,000x more reads of similar quality as currently performed by RNA-seq. Currently nucleotide sequencing methods are approaching their limits for single cell-RNA-seq where ~1000x more reads are needed for 1000 single cells than a bulk analysis. Thus, single molecule counting based methods will require technological advancements in nanopores, flow-cells and fluorescent detection methods that significantly exceed the capabilities of nucleotide single molecule sequencing 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 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 ~1 μm^2 . Thus, assuming perfect measurement of reads and ideal spatial placement we can estimate the best case scenario for the minimal flow cell area vs. number of reads: 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 largely 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 10 billion nucleotide reads is time consuming and expensive for sc-RNAseq 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 the coverage currently achieved on the transcriptome by RNA-seq.

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 a given sequencing run, which takes a couple of days to complete[48]. 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 5x10^7 molecules per run, which takes 1-2 days for data acquisition (signal processing time not included)[49]. 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 5x10^7 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 contain 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 (10^6) 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, otherwise the sensor will be mostly vacated 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 kHz due to data transfer speed and signal-to-noise limitations. These multiple constraints have set natural limits for 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 can imagine over 100-fold reductions in input requirements from >100 ng to <1 ng, at similar throughputs. Similarly, if one were to assume that data transfer and bandwidths would increase by ~100 fold over the next 5-7 years, one can expect transitioning from 10^3 pores in a flowcell to 10^5, which would boost the throughput 100-fold to about $5x10^9$ molecules per run (1-2 days). It seems likely that these limitations will have to be overcome first for genomics/transcriptomics over the next 5-7 years, before single-molecule proteomics can be approached at scale using single-molecule tools.

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 physiochemical 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 both 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 10-100% (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 specificity[50,51]. 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 additional 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 capture[52], ChIP[53], or similar methods using in genomics prior to single molecule sequencing. Thus, there is a balance between maximizing coverage of individual proteins and the dynamic range of the proteins measured.

The major limiting factor in the sensitivity of LC-MS/MS methods is the electrospray process. Electrospray is the Nobel prize winning invention that is most commonly used for turning peptide molecules in solution into gas-phase ions[8]. However, tif a molecule isn't 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[12,54].

The number of ions sampled may be increased by using methods like multidimensional chromatography[14] or making multiple analyses using different portions of the mass range[55,56]. These approaches can significantly improve the depth of proteome coverage at the expense of increased analysis time. Thus, these gains in proteome coverage come at the cost of lower throughput and don’t increase linearly with the time spent. For example, a 6x increase in time often only increases the number of peptides that can be measured by 2x – because the increased time is at least partially redundant with the peptides measured in prior fractions. Ultimately this comes at the expense of protein input material and significantly 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 faster, thus enabling higher throughput[57].

Another way to improve LC-MS/MS is in the more efficient use of the ions that are generated. Currently, in most data independent acquisition methods, a single wide m/z range is isolated at once and the rest of the ion beam that isn’t isolated is lost. Data dependent acquisition methods sample an even smaller fraction of the ion beam. With bulk samples, this means that only ~1/50th of the ion beam is currently being used as only one of 50 precursor windows is measured at once[58]. With single-cell samples, 3-4 windows are used and thus about ⅓ of all ions available to the MS instrument are analyzed[15] at the expense of limiting within spectrum selectivity. Methods like diaPASEF (parallel accumulation-serial fragmentation combined with data independent acquisition) offer potential to significantly 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 sequences[59]. 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 50-75x 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 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 coverage[59].

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 abundant peptides – improving the dynamic range (Figure 2). However, the emphasis on generating sorting 'like' ions constrains the choice of enzymes to produce peptides ideally suited for the respective method. Because tryptic peptides are ideally suited for LC-MS/MS doesn’t mean it 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 towards 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 can perhaps constructively guide the emerging field of single-molecule proteomics. As advice to this budding field, consider the following.

Fractionate: Better to run many smaller counting experiments on fractionated samples than one very large counting experiment (Figure 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/quantification processes. To make optimal use of this separation, methods equivalent to automatic gain control (AGC), as done with ion trap instruments[60], 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 be used to 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[61–63]. This minimizes the FDR and improves the sensitivity. Indeed, nanopore proteomics methods are making first steps in this direction[64].

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 perform a background subtraction, 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 1) exceed the density of the flow cell or 2) require a flow cell(s) 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). Single molecule counting works great for transcripts because there are so few transcript molecules to count but the density can't scale easily 30,000x to extend to the dynamic range of the proteome.

Overcoming biases. Arguably the most challenging aspect of proteomics is the massive physiochemical diversity of proteins in the cell. To overcome this vast diversity in solubility, embedded transmembrane domain containing proteins, size, combinatorial post-translational modification, ionization and fragmentation by mass spectrometry, presence of autoantibodies, or protein-protein interactions, most proteomics experiments take a bottom-up strategy for the analysis of complex mixtures by digesting proteins to peptides prior to analysis. Performing analyses on the peptide level greatly simplifies the physiochemical 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 properties[65].

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, one of the greatest challenges for single cell and low-input proteomics is the absorption of proteins and peptides to surfaces. In general, the sensitivity limits of proteomics samples have not been because of LC-MS/MS itself but the loss of sample to surfaces prior to 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[66,67], nanodroplet sample preparation[68], addition of carrier proteins[69], and barcoding and combining samples using mass tags to spread losses between many samples[15].

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 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 barcoded (e.g., by covalent chemical labels), subsequently mixed, and analyzed simultaneously. Such sample multiplexing has helped increase the throughput of MS proteomics[11,70] likely to be efficiently implemented by single-molecule methods to increase the number of samples analyzed. Indeed, multiplexing is a powerful feature of single molecule DNA sequencing. Yet, multiplexing with single-molecule approaches spreads the counted molecules across 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 historically 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 technology[71]. However, as discussed above, the large number of protein molecules will require sequencing billions of molecules to obtain coverage of the proteome that can be obtained by LC-MS/MS[25]. Because this coverage can be obtained for ~$500 per analysis by LC-MS/MS and sequencing billions of DNA reads can cost ~$10,000, it would require Next Generation Sequencing companies to reduce their costs to ~5% their current rates. 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 be useful for single molecule counting methods to use as a benchmark to obtain parity with LC-MS data. The challenges of 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 physiochemical complexity of proteins and the sheer greater number of proteins in the cell. To support innovation around these emerging methods we provide some suggestions learned by the LC-MS/MS based proteomics community.

Acknowledgements

This work was supported in part by National Institutes of Health grants U19 AG065156, R24 GM141156, 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., the project 'International Centre for Cancer Vaccine Science' that is 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. The authors would like to acknowledge the helpful discussions with Edward Marcotte, Christine Wu, and members of the Alfaro, MacCoss, and Slavov labs. MJM appreciates the constructive feedback provided by UW Genome Sciences faculty.

1. Nau, H. & Biemann, K. Amino acid sequencing by gas chromatography--mass spectrometry using perfluoro-dideuteroalkylated peptide derivatives. A. Gas chromatographic retention indices. *Anal. Biochem.* **73**, 139–153 (1976).
2. Gerber, G. E. *et al.* Partial primary structure of bacteriorhodopsin: sequencing methods for membrane proteins. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 227–231 (1979).
3. Hass, G. M. *et al.* The amino acid sequence of a carboxypeptidase inhibitor from potatoes. *Biochemistry* **14**, 1334–1342 (1975).
4. Hunt, D. F., Yates, J. R., 3rd, Shabanowitz, J., Winston, S. & Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 6233–6237 (1986).
5. Hunt, D. F. *et al.* Tandem quadrupole Fourier-transform mass spectrometry of oligopeptides and small proteins. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 620–623 (1987).
6. Gibson, B. W. & Biemann, K. Strategy for the mass spectrometric verification and correction of the primary structures of proteins deduced from their DNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 1956–1960 (1984).
7. Johnson, R. S. & Biemann, K. The primary structure of thioredoxin from Chromatium vinosum determined by high-performance tandem mass spectrometry. *Biochemistry* **26**, 1209–1214 (1987).
8. Yamashita, M. & Fenn, J. B. Electrospray ion source. Another variation on the free-jet theme. *J. Phys. Chem.* **88**, 4451–4459 (1984).

9. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).

10. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1**, 39–45 (2004).

11. Ross, P. L. *et al.* Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3**, 1154–1169 (2004).

12. Specht, H. *et al.* Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome Biol.* **22**, 50 (2021).

13. Petelski, A. A. *et al.* Multiplexed single-cell proteomics using SCoPE2. *Nat. Protoc.* **16**, 5398–5425 (2021).

14. Washburn, M. P., Wolters, D. & Yates, J. R., 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247 (2001).

15. Derks, J. *et al.* Increasing the throughput of sensitive proteomics by plexDIA. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01389-w.

16. Messner, C. B. *et al.* Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection. *Cell Systems* vol. 11 11–24.e4 (2020).

17. Alfaro, J. A. *et al.* The emerging landscape of single-molecule protein sequencing technologies. *Nat. Methods* **18**, 604–617 (2021).

18. Swaminathan, J., Boulgakov, A. A. & Marcotte, E. M. A theoretical justification for single molecule peptide sequencing. *PLoS Comput. Biol.* **11**, e1004080 (2015).

19. Palmblad, M. Theoretical Considerations for Next-Generation Proteomics. *J. Proteome Res.* **20**, 3395–3399 (2021).

20. Swaminathan, J. *et al.* Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4278.

21. Reed, B. D. *et al.* Real-time dynamic single-molecule protein sequencing on an integrated semiconductor device. *bioRxiv* 2022.01.04.475002 (2022) doi:10.1101/2022.01.04.475002.

22. Mallick, P. Methods of assaying proteins. *US Patent* (2021).

23. Egertson, J. D. *et al.* A theoretical framework for proteome-scale single-molecule protein identification using multi-affinity protein binding reagents. *bioRxiv* 2021.10.11.463967 (2021) doi:10.1101/2021.10.11.463967.

24. Brinkerhoff, H., Kang, A. S. W., Liu, J., Aksimentiev, A. & Dekker, C. Multiple rereads of single proteins at single-amino acid resolution using nanopores. *Science* **374**, 1509–1513 (2021).

25. Slavov, N. Counting protein molecules for single-cell proteomics. *Cell* (2022).

26. Milo, R. What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays* **35**, 1050–1055 (2013).

27. Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst* **4**, 587–599.e4 (2017).

28. Anderson, N. L. & Anderson, N. G. The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1**, 845–867 (2002).

29. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* **24**, 496–510 (2014).

30. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).

31. 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: Volume 3* (eds. Hercules, D. M., Hieftje, G. M., Snyder, L. R. & Evenson, M. A.) 217–252 (Springer US, 1978).

32. Scigelova, M., Hornshaw, M., Giannakopulos, A. & Makarov, A. Fourier transform mass spectrometry. *Mol. Cell. Proteomics* **10**, M111.009431 (2011).

33. Makarov, A. & Denisov, E. Dynamics of ions of intact proteins in the Orbitrap mass analyzer. *J. Am. Soc. Mass Spectrom.* **20**, 1486–1495 (2009).

34. MacCoss, M. J., Toth, M. J. & Matthews, D. E. Evaluation and optimization of ion-current ratio measurements by selected-ion-monitoring mass spectrometry. *Anal. Chem.* **73**, 2976–2984 (2001).

35. Schwartz, J. C., Zhou, X.-G. & Bier, M. E. Method and apparatus of increasing dynamic range and sensitivity of a mass spectrometer. *US Patent* (1996).

36. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome

annotation and quantification using RNA-seq. *Nat. Methods* **8**, 469–477 (2011).

37. Zhao, S., Ye, Z. & Stanton, R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* **26**, 903–909 (2020).
38. Belov, M. E. *et al.* Dynamic range expansion applied to mass spectrometry based on data-dependent selective ion ejection in capillary liquid chromatography fourier transform ion cyclotron resonance for enhanced proteome characterization. *Anal. Chem.* **73**, 5052–5060 (2001).
39. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **15**, 440–448 (2018).
40. Egertson, J. D. *et al.* Multiplexed MS/MS for improved data-independent acquisition. *Nat. Methods* **10**, 744–746 (2013).
41. Anderson, N. L. *et al.* The human plasma proteome: a nonredundant list developed by combination of four separate sources. *Mol. Cell. Proteomics* **3**, 311–326 (2004).
42. Pieper, R. *et al.* Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome. *Proteomics* **3**, 422–432 (2003).
43. Gundry, R. L., Fu, Q., Jelinek, C. A., Van Eyk, J. E. & Cotter, R. J. Investigation of an albumin-enriched fraction of human serum and its albuminome. *Proteomics Clin. Appl.* **1**, 73–88 (2007).
44. Macdonald, I. K., Parsy-Kowalska, C. B. & Chapman, C. J. Autoantibodies: Opportunities for Early Cancer Detection. *Trends Cancer Res.* **3**, 198–213 (2017).
45. Hoofnagle, A. N. & Wener, M. H. The fundamental flaws of immunoassays and potential solutions using tandem mass spectrometry. *J. Immunol. Methods* **347**, 3–11 (2009).
46. McVicar, J. P., Kunitake, S. T., Hamilton, R. L. & Kane, J. P. Characteristics of human lipoproteins isolated by selected-affinity immunosorption of apolipoprotein A-I. *Proc. Natl. Acad. Sci. U. S. A.* **81**, 1356–1360 (1984).
47. Heinecke, J. W. The HDL proteome: a marker--and perhaps mediator--of coronary artery disease. *J. Lipid Res.* **50 Suppl**, S167–71 (2009).
48. PacBio sequel systems. *PacBio* https://www.pacb.com/technology/hifi-sequencing/sequel-system/ (2018).
49. PromethION. *Oxford Nanopore Technologies* https://nanoporetech.com/products/promethion.
50. Gatlin, C. L., Eng, J. K., Cross, S. T., Detter, J. C. & Yates, J. R., 3rd. Automated identification of amino

acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Anal. Chem.* **72**, 757–763 (2000).

51. MacCoss, M. J. *et al.* Shotgun identification of protein modifications from protein complexes and lens tissue. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 7900–7905 (2002).
52. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* **6**, 315–316 (2009).
53. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
54. Gray Huffman, R. *et al.* Prioritized single-cell proteomics reveals molecular and functional polarization across primary macrophages. *bioRxiv* 2022.03.16.484655 (2022) doi:10.1101/2022.03.16.484655.
55. Scherl, A. *et al.* Genome-specific gas-phase fractionation strategy for improved shotgun proteomic profiling of proteotypic peptides. *Anal. Chem.* **80**, 1182–1191 (2008).
56. Panchaud, A. *et al.* Precursor acquisition independent from ion count: how to dive deeper into the proteomics ocean. *Anal. Chem.* **81**, 6481–6488 (2009).
57. Slavov, N. Increasing proteomics throughput. *Nature biotechnology* vol. 39 809–810 (2021).
58. Pino, L. K., Just, S. C., MacCoss, M. J. & Searle, B. C. Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Mol. Cell. Proteomics* **19**, 1088–1103 (2020).
59. Slavov, N. Driving Single Cell Proteomics Forward with Innovation. *J. Proteome Res.* **20**, 4915–4918 (2021).
60. Schwartz, J. C. & Kovtoun, V. V. Automatic gain control (AGC) method for an ion trap and a temporally non-uniform ion beam. *US Patent* (2011).
61. Klammer, A. A., Yi, X., MacCoss, M. J. & Noble, W. S. Improving tandem mass spectrum identification using peptide retention time prediction across diverse chromatography conditions. *Anal. Chem.* **79**, 6111–6118 (2007).
62. Searle, B. C. *et al.* Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**, 5128 (2018).
63. Chen, A. T., Franks, A. & Slavov, N. DART-ID increases single-cell proteome coverage. *PLoS Comput.*

Biol. **15**, e1007082 (2019).

64. Zrehen, A., Ohayon, S., Huttner, D. & Meller, A. On-chip protein separation with single-molecule resolution. *Sci. Rep.* **10**, 15313 (2020).

65. Donnelly, D. P. *et al.* Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nat. Methods* **16**, 587–594 (2019).

66. Zhu, Y. *et al.* Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nat. Commun.* **9**, 1–10 (2018).

67. Specht, H., Harmange, G., Perlman, D. H. & Emmott, E. Automated sample preparation for high-throughput single-cell proteomics. *BioRxiv* (2018) doi:10.1101/399774.

68. Leduc, A., Huffman, R. G. & Slavov, N. Droplet sample preparation for single-cell proteomics applied to the cell cycle. *bioRxiv* (2021) doi:10.1101/2021.04.24.441211.

69. Budnik, B., Levy, E., Harmange, G. & Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19**, 161 (2018).

70. Framework for multiplicative scaling of single-cell proteomics. *Nat. Biotechnol.* 10.1038/s41587–022–01411–1 (2022).

71. Hong, J. M. *et al.* ProtSeq: Toward high-throughput, single-molecule protein sequencing via amino acid conversion into DNA barcodes. *iScience* **25**, 103586 (2022).