Why proteomics is not the new genomics and the future of mass spectrometry in cell biology

Simone Sidoli,[1] Katarzyna Kulej,[1,2] and Benjamin A. Garcia[1]

[1]Epigenetics Program, Department of Biochemistry and Biophysics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104
[2]Division of Cancer Pathobiology, Children's Hospital of Philadelphia, Philadelphia, PA 19104

Mass spectrometry (MS) is an essential part of the cell biologist's proteomics toolkit, allowing analyses at molecular and system-wide scales. However, proteomics still lag behind genomics in popularity and ease of use. We discuss key differences between MS-based -omics and other booming -omics technologies and highlight what we view as the future of MS and its role in our increasingly deep understanding of cell biology.

In the mid 1990s, MS fiercely entered the cell biology field, as its potential in identifying and quantifying entire proteomes became clear. In 2014, *Nature* released an issue titled "The human proteome," containing studies analyzing the proteome of human tissues and cell lines using MS and generating a wealth of data (Kim et al., 2014; Wilhelm et al., 2014). That same year, Hebert et al. (2014) published "The one hour yeast proteome," where they showed that about one hour of chromatographic separation coupled with high-performance MS is sufficient to achieve extensive proteome coverage for a simple organism like yeast. This was a milestone for MS-based proteomics, attesting to the high throughput of the technique. Today, the number of proteomics core facilities and services are growing, indicating that the technique reached a level of robustness and reproducibility that can be detached from specific research laboratories holding the technical expertise required for MS. One might think that the proteomics field is where the genomics field was ∼15 yr ago. However, an interesting statistic emerges from counting scientific publications: In the last 10 yr or so, genomics studies have been growing at a faster rate than proteomics (Fig. 1 A). How is proteomics not the new genomics yet? What is MS missing from becoming the ideal tool for a comprehensive characterization of biological systems? In this viewpoint, we highlight the common obstacles that prevent successful data interpretation in the MS field and contrast them with the rapid progress seen in the genomics field. We also discuss how the cell biology community, by overcoming these hurdles in data interpretation and material sharing, can use MS to reach deeper levels of analyses at single-cell and system-wide levels.

Why are MS applications and results still so hard to interpret?

By definition, a mass spectrometer determines the mass-to-charge ratio of a signal ionized in gas phase, which can be converted into the mass of the molecule. Countless experiments can thus be performed where a mass or a mass shift is used as readout. Proteomics is most used for (a) the identification of peptides, proteins, and posttranslational modifications; (b) the measure of protein amounts or turnover, by combining labeling techniques to MS; (c) the characterization of protein structure; and (d) the identification of protein interactions with proteins or nucleic acids (e.g., He et al., 2016). This plethora of applications requires focused efforts, and thus MS laboratories have specialized to optimize the methods for an application of interest. Nowadays, it is common to refer to one proteomics laboratory dedicated to protein structure analyses and to a different laboratory for protein–protein interactions, as the entire instrumental setup is likely different. This specialization of laboratories has not occurred in genomics, as experiments like chromatin immunoprecipitation sequencing, RNA sequencing, assay for transposase-accessible chromatin with high-throughput sequencing, and deep sequencing require similar types of knowledge in operating the instrument and in data analysis.

In addition, the MS field faces different difficulties from the genomics field. Part of the issue is sensitivity; nucleotide sequences can be amplified, allowing analyses up to the single-cell level, which is currently impossible for proteins and metabolites besides rare exceptions. Another difficulty relates to the free distribution of software for data analysis; in genomics, bioinformatics tools are almost never proprietary, whereas it is a much more common practice for proteomics and metabolomics. Finally, MS results are usually more complex to interpret, as the output depends on the method of acquisition and data analysis platform. Further, a considerable challenge in the MS field is defining how much in the output is "real." Every MS scientist who has dealt with collaboration knows the feeling of looking a biologist in the eyes while he/she is asking, "Is this metabolite present in my sample? And is it gone after the treatment?" Sometimes, the answer is not a simple "yes" or "no." Mass spectrometers are very sensitive instruments (limit of detection < attomoles), but defining the threshold between signal and noise is difficult. You might be able to select a signal that, once fragmented, produces a pattern very similar to the metabolite of interest. Moreover, mass spectrometers are quantitative instruments; the intensity of the signal can be correlated with its abundance in the sample. However, the signal of the analyte of interest could be mixed with background noise, as other metabolites might have isobaric masses, i.e., the same atomic

Correspondence to Benjamin A. Garcia: bgarci@mail.med.upenn.edu

The Rockefeller University Press $30.00
J. Cell Biol. Vol. 216 No. 1 21–24
https://doi.org/10.1083/jcb.201612010

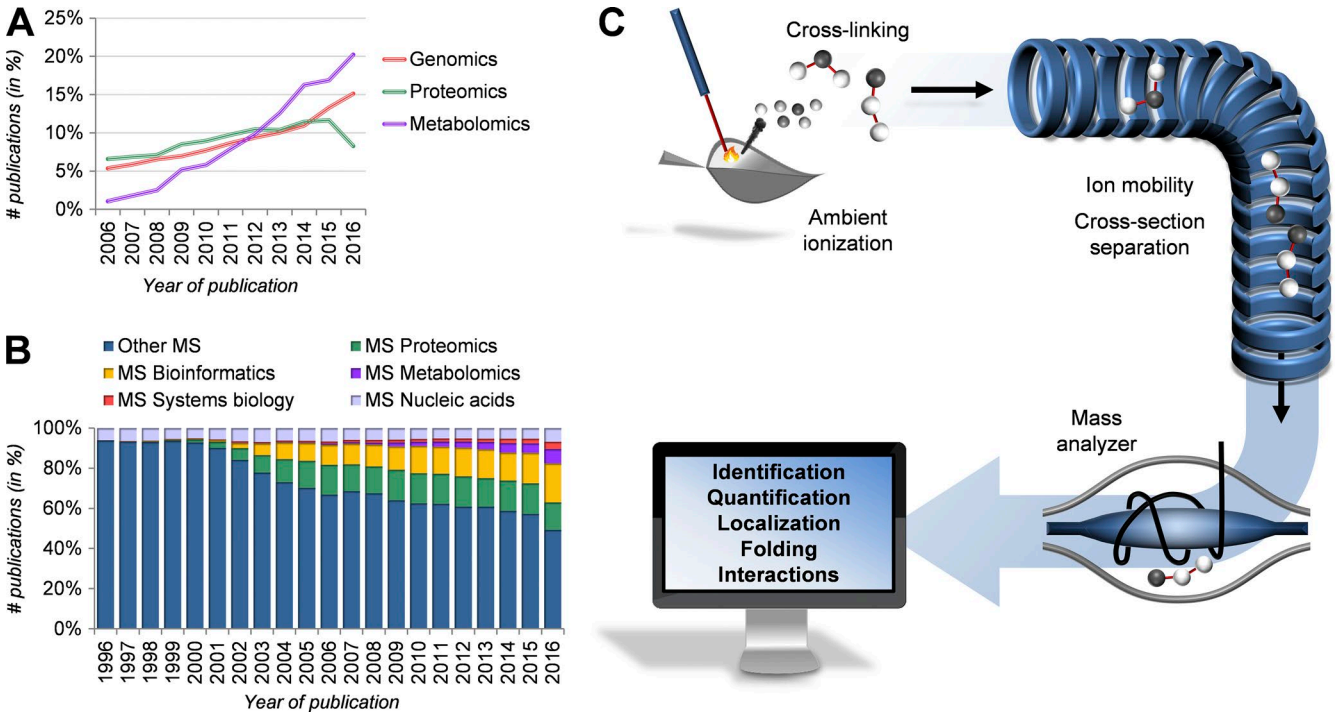

Figure 1. **MS past, present, and future.** (A) Number of publications containing the terms genomics, proteomics, or metabolomics in title or abstract (based on PubMed). Each value per year was normalized by the total across all years analyzed. (B) Same representation dating back to 1996. Papers were counted if they contained the term "mass spectrometry" plus the term listed in the legend. (C) Representation of applications of MS-based proteomics studies. Ambient ionization allows for site-specific identification of analytes; cross-linking preserves interactions; ion mobility allows for separation of same-mass analytes based on their cross section; at the end of the pipeline, the mass analyzer determines mass and intensity of analytes.

composition but different structure. Usually, the answer to the collaborator is, "I do see a molecule with the same mass as your metabolite of interest, and I cannot detect a signal in the treatment!" You can tell from his/her eyes that a clearer answer is desirable. This is more a problem of communication rather than of unreliable or unclear results. Scientists are used to analyses at the level of the genome that can be performed from a single cell with unambiguous determination of a specific genomic location or of the presence of a mutation. A mass spectrometer detects everything that ionizes, not just the biomolecule of interest. Indeed, mass spectrometers offer high resolution (>400,000 mass/Δmass), high mass accuracy (<1 ppm), high sensitivity (<attomol), and high speed (12–20 Hz), meaning that they can easily generate 12–20 mass spectra within a second of analysis. Proteomics, metabolomics, and lipidomics deal with highly complex samples with a wide, dynamic range in analyte abundance. Sample complexity leads to mixed spectra of difficult interpretation, and sample dynamic range harms linear quantification for low abundant signals. The great sensitivity of MS is a risk for false positives, so expert analysts tend to be conservative when providing results.

Given the challenges in data interpretation, how can we be confident in MS results?

To help reduce the risk of false positives, technical efforts have focused on gas and liquid chromatography to optimize the online coupling to MS, as chromatography reduces the complexity of the signal within each MS scan and increases the confidence in spectra identification. Although MS technology has grown faster and more sensitive, more and better bioinformatics tools for MS have also been developed (Fig. 1 B). About 20 different database-searching engines now exist for proteomics analyses, several of them freeware, reflecting the effort for confident identification of spectra and protein mixtures. A similar trend is observed for metabolomics, even though it is, in a sense, a younger academic field than proteomics and fewer tools are available. Thanks to recent computational advances, protein quantification is more accurate. Spectral counting was literally based on counting spectra that identified a certain protein, using a similar approach to genomic analyses that count the number of reads covering a specific sequence. However, MS cannot provide the depth of coverage provided by high-throughput sequencing; low abundance peptides are detected within an MS run by one to five spectra, which are insufficient values for accurate quantification. Nowadays, software automatically extract the intensity or the area of ion chromatograms and associate those signals with identifications of peptides or metabolites generated within the same MS run. Special care is thus taken for chromatography, which should provide defined and Gaussian-shaped peaks. Recently, quantification improved in accuracy thanks to data-independent acquisition methods (Gillet et al., 2012), which allow extraction of both precursor and fragment ion chromatograms of an analyte, increasing the confidence in selecting the proper signal. Proteomics and metabolomics are now reliable quantitative disciplines, whereas they were labeled "semi-quantitative" a few years ago. However, reproducibility is still lower than in genomics, as ion chromatograms are more complex to extract with high confidence as compared with reads of DNA libraries. Thus, stringent statistics, such as correction for multiple testing, are almost never applied on quantitative values of MS studies, and the sole fold change is sometimes erroneously used as a threshold to filter relevant results (e.g., Najm et al., 2015).

What does MS bring to cell biology?

Despite important limitations, MS has undoubtable advantages. Although it is obvious, much information about cell phenotype cannot be detected otherwise, including protein posttranslational modifications, protein interactions, metabolite abundance, and, not less importantly, protein stoichiometry (Schwanhäusser et al., 2011). The concept of protein hindrance is usually ignored by experiments that compare sample A to sample B. However, it is important to consider that proteins occupy the vast majority of the cell and that the number of such large biomolecules affects cell shape and behavior. Likely, a protein up-regulated from 20 to 40 copies has a different effect than one changing from 10 to 20 million copies, even though results display the same twofold change. MS results are inevitably affected by this issue. For instance, fewer proteins are identified in a muscle fiber than its respective stem cell. This is not only because of difficulties in homogenizing a differentiated tissue but also because a specialized cell expresses more copies of specific proteins, suppressing the signal of others. Keeping in mind a quantitative understanding of protein abundance and relative protein proportions in cells is not only necessary to gain a better understanding of the system studied but also helps with proper data normalization and facilitates experimental design.

Where is MS heading?

Even with many specialized adaptations of MS protocols and equipment, certain information is hard to extract based on the mass of the analyte of interest, revealing important technical limits to what cell biological questions can be addressed with MS. For instance, studying protein folding is challenging, as two proteins with the same sequence but different folding produce a single signal in MS. This technical limit has been addressed in part by the development of ion mobility (Kanu et al., 2008), which consists in a tube inside the mass spectrometer filled with gas that generates a friction on the molecules flying toward it (Fig. 1 C). Intuitively, an unfolded protein has a larger cross section than a folded one, so it will be retained for a longer period and the mixture will generate two distinct signals. This is basically a new dimension of separation, which can be used to separately analyze two species despite having the same mass. Interestingly, ion mobility has gained more popularity in the metabolomics community than the proteomics one, which may be because metabolites include a large number of molecules with the same intact mass. MS fragmentation, also called tandem MS or MS/MS, is used to increase confidence in analyte identification, as the fragmentation pattern of a peptide or a metabolite is more unique than its intact mass. MS/MS can thus be used to discriminate between multiple species with the same intact mass, but it is sometimes insufficient, and, thus, ion mobility has helped increase confidence by characterizing an additional feature to the analyte named "drift time."

Beyond technical improvements to better interrogate the same cell biological topics, the field is also under pressure to broaden the level of analyses. MS laboratories are now routinely asked, "Can we do single-cell analysis?" This question is typically important because a sample can be a mix of heterogeneous cell types that cannot easily be sorted. Proteomics and metabolomics live in the eternal conflict between simplifying the workflow while maintaining the depth of coverage, determining the localization of specific analytes, and adopting more and more complex purification strategies. Unfortunately, it is not possible to determine the organelle localization of proteins or metabolites from a tissue lysate if the tissue sample is homogenized before MS analysis. A new discipline in MS is thus gaining popularity—imaging (Fig. 1 C). Any molecule that flies and is ionized potentially enters the mass spectrometer. Under this assumption, a few ion sources have been developed to charge and sublimate analytes, generating a field named "ambient MS" (Cooks et al., 2006). The ion source is normally a laser, but it could also be a desorption electrospray ionization (Takáts et al., 2005). Ultimately, the sample is placed on a support and it is scanned, generating a pixelated image. Each of those pixels corresponds to a spectrum, so it becomes possible to monitor the localization of a specific analyte by extracting the spectra containing such ion in a manner similar to extracted ion chromatography. Today, imaging MS has reached translational applications (e.g., iKnife [Balog et al., 2013]). iKnife exploits the characteristics of electrosurgical scalpels used during dissections and creates a small aerosol of the tissue during cutting. Such an aerosol can be driven into a mass spectrometer and the spectra generated are monitored in real time. By having software instructed to identify spectra fingerprinting, e.g., healthy and tumor tissues, it is possible to determine on-the-fly the nature of the tissue being excised.

Toward an integrated view of MS and other -omics analyses

Imaging MS and ion mobility have quickly pushed the field forward by improving data interpretation and resolution, and these developments have been met with strong interest and excitement. Attesting to a shift in the field, at the 2016 American Society for Mass Spectrometry annual conference, three oral sessions were solely dedicated to ion mobility, and four to imaging, whereas only one session combined older MS-based approaches including electrospray, matrix-assisted laser desorption/ionization, and mass analyzers. The recent focus on imaging MS shows that the future of MS goes beyond just mass detection. We reached a stage where high-throughput sequencing and MS cover all the large-scale disciplines we can define: genomics, transcriptomics, proteomics, metabolomics, and lipidomics. With this amount of information, how we do not already know everything? Most agree that combining the results of all those techniques is the future of cell biology (Gomez-Cabrero et al., 2014). However, the challenges nested into the computational infrastructure needed are significant. First, systems are highly dynamic. Epigenetic mechanisms are regulated throughout the entire lifespan; metabolites vary within a single day because of the circadian clock (Minami et al., 2009); protein phosphorylation status changes within milliseconds in, e.g., synaptosomes (Craft et al., 2008). Time should then be a variable when merging the different -omics in the study of a biological system, but this requires multiple measurements and clustering, which is not always feasible. Another problem is communication across fields. We highlighted some misunderstanding between scientists with sequencing- or MS-based backgrounds but more are present. The solution is not simple, but more interdisciplinary collaboration would be a good place to start. In addition, we speculate that not only genomics but also proteomics and metabolomics classes will soon be taught in biochemistry and molecular biology courses.

In conclusion, we foresee that in the coming 10 yr at least one mass spectrometer will be present in every biology department. The flexibility and potential of MS remains to be fully exploited, and the creativity of MS researchers is only limited

by the questions they are asking. Applications like proteomics are becoming a routine technical service that no longer requires highly specialized researchers, but, in our view, we are not there yet. It will take a few more years for other applications such as metabolomics to display their full potential, and highly trained scientists focused on advancing the technology are needed to push the field forward. Ion mobility and imaging have proven their relevance, but considering the amount of ongoing research on methodological developments, it is safe to assume that the best has yet to come. Finally, emerging applications like microfluidics, i.e., spraying entire cells into MS, still have to go beyond proofs-of-concept, and we recommend keeping an open mind about them, as MS is capable of this and much more. MS still has a lot of optimization ahead but also a lot to offer, and we are excited for future collaborative MS-based research to provide insights into previously inaccessible corners of cell biology.

Acknowledgments

We are grateful to Amber K. Weiner for her expert opinion and advice on the genomics aspect of the manuscript.

We would also like to thank funding from the National Institutes of Health (GM110104, CA196539, and AI118891).

The authors declare no competing financial interests.

Submitted: 2 December 2016
Accepted: 2 December 2016

References

Balog, J., L. Sasi-Szabó, J. Kinross, M.R. Lewis, L.J. Muirhead, K. Veselkov, R. Mirnezami, B. Dezső, L. Damjanovich, A. Darzi, et al. 2013. Intraoperative tissue identification using rapid evaporative ionization mass spectrometry. *Sci. Transl. Med.* 5:194ra93. http://dx.doi.org/10.1126/scitranslmed.3005623

Cooks, R.G., Z. Ouyang, Z. Takats, and J.M. Wiseman. 2006. Detection Technologies. Ambient mass spectrometry. *Science.* 311:1566–1570. http://dx.doi.org/10.1126/science.1119426

Craft, G.E., M.E. Graham, N. Bache, M.R. Larsen, and P.J. Robinson. 2008. The in vivo phosphorylation sites in multiple isoforms of amphiphysin I from rat brain nerve terminals. *Mol. Cell. Proteomics.* 7:1146–1161. http://dx.doi.org/10.1074/mcp.M700351-MCP200

Gillet, L.C., P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold. 2012. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Mol. Cell. Proteomics.* 11:O111.016717. http://dx.doi.org/10.1074/mcp.O111.016717

Gomez-Cabrero, D., I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér. 2014. Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8:I1. http://dx.doi.org/10.1186/1752-0509-8-S2-I1

He, C., S. Sidoli, R. Warneford-Thomson, D.C. Tatomer, J.E. Wilusz, B.A. Garcia, and R. Bonasio. 2016. High-resolution mapping of RNA-binding regions in the nuclear proteome of embryonic stem cells. *Mol. Cell.* 64:416–430. http://dx.doi.org/10.1016/j.molcel.2016.09.034

Hebert, A.S., A.L. Richards, D.J. Bailey, A. Ulbrich, E.E. Coughlin, M.S. Westphall, and J.J. Coon. 2014. The one hour yeast proteome. *Mol. Cell. Proteomics.* 13:339–347. http://dx.doi.org/10.1074/mcp.M113.034769

Kanu, A.B., P. Dwivedi, M. Tam, L. Matz, and H.H. Hill Jr. 2008. Ion mobility-mass spectrometry. *J. Mass Spectrom.* 43:1–22. http://dx.doi.org/10.1002/jms.1383

Kim, M.S., S.M. Pinto, D. Getnet, R.S. Nirujogi, S.S. Manda, R. Chaerkady, A.K. Madugundu, D.S. Kelkar, R. Isserlin, S. Jain, et al. 2014. A draft map of the human proteome. *Nature.* 509:575–581. http://dx.doi.org/10.1038/nature13302

Minami, Y., T. Kasukawa, Y. Kakazu, M. Iigo, M. Sugimoto, S. Ikeda, A. Yasui, G.T. van der Horst, T. Soga, and H.R. Ueda. 2009. Measurement of internal body time by blood metabolomics. *Proc. Natl. Acad. Sci. USA.* 106:9890–9895. http://dx.doi.org/10.1073/pnas.0900617106

Najm, F.J., M. Madhavan, A. Zaremba, E. Shick, R.T. Karl, D.C. Factor, T.E. Miller, Z.S. Nevin, C. Kantor, A. Sargent, et al. 2015. Drug-based modulation of endogenous stem cells promotes functional remyelination in vivo. *Nature.* 522:216–220. http://dx.doi.org/10.1038/nature14335

Schwanhäusser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. 2011. Global quantification of mammalian gene expression control. *Nature.* 473:337–342. http://dx.doi.org/10.1038/nature10098

Takáts, Z., J.M. Wiseman, and R.G. Cooks. 2005. Ambient mass spectrometry using desorption electrospray ionization (DESI): Instrumentation, mechanisms and applications in forensics, chemistry, and biology. *J. Mass Spectrom.* 40:1261–1275. http://dx.doi.org/10.1002/jms.922

Wilhelm, M., J. Schlegl, H. Hahne, A.M. Gholami, M. Lieberenz, M.M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature.* 509:582–587. http://dx.doi.org/10.1038/nature13319