Proteomics

DOI: 10.1002/anie.201406545

From the Human Genome to the Human Proteome**

Javier Muñoz and Albert J. R. Heck**

human genome · human proteome · mass spectrometry · proteomics

An ultimate goal in biology is to fully understand how a cell, or even a whole organism, works. Ideally, such knowledge might be used to develop models that predict the responses of cells to specific cues or diseases. A first step towards this goal is to identify and characterize all the molecular players present in cells. The Human Genome Project[1] was probably one of the most ambitious scientific endeavors so far and provided the first essential pieces in this puzzle. The availability of the 3 billion base pairs that make up our DNA generated worldwide excitement as this information might lead to understanding the molecular mechanisms of human pathologies.

However, soon thereafter the complexity embedded in our genetic code was already realized. A surprising finding was the low percentage of DNA (less than 2% of the genome) coding for proteins—roughly 20.000 human genes. However, recent analysis indicates that 80% of the human genome is functional and either transcribed, binding to regulatory proteins, or associated with other biochemical functions.[2] Although genomic information is vital, it does not touch upon proteins, the main molecular effectors of cells. Every researcher will agree that the analysis of the proteome is of more relevance, but still this has been less exploited due to technical hurdles and by the fact that the proteome is inherently several magnitudes more complex. Whereas the genome is nearly identical in every cell of the human body and also relatively constant over the lifetime of an organism, the proteome of every cell is very different and changes dramatically over time (Figure 1).

[*] Prof. Dr. A. J. R. Heck
Biomolecular Mass Spectrometry and Proteomics, Bijvoet Center for Biomolecular Research and Utrecht Institute for Pharmaceutical Sciences, Utrecht University
Padualaan 8. 3584 CH Utrecht (The Netherlands)
E-mail: a.j.r.heck@uu.nl
Dr. J. Muñoz
Proteomics Unit
Spanish National Cancer Research Centre (CNIO, ProteoRed-ISCIII)
Melchor Férnandez Almagro, 3, 28029 Madrid (Spain)
E-mail: jmunozpe@cnio.es

[**] A.J.R.H. acknowledges continuous support to the Netherlands Proteomics Center by the Netherlands Organization for Scientific Research (NWO) supported large-scale proteomics facility Proteins@Work (project 184.032.201) and the PRIME-XS Project (Grant Agreement 262067) funded by the European Union Seventh Framework Program. J.M. is supported by Ramon y Cajal Programme (MINECO) RYC-2012-10651. The CNIO Proteomics Unit belongs to ProteoRed, PRB2-ISCIII, supported grant PT13/0001.

Notwithstanding these challenges, the field of proteomics has witnessed tremendous developments over the last decade, primarily through advances in mass spectrometry and bioinformatics, and is now somewhat coming up to par with genomics and transcriptomics technologies. This is evidenced by two recent reports in *Nature* from a German team led by Bernard Küster[3] and a USA/India-based collaboration headed by Akhilesh Pandey,[4] who independently initiated an unprecedented effort with the aim of identifying all the human proteins encoded in the genome. To this end, both laboratories performed extensive proteomic analyses on more than 70 human tissues and body fluids and more than 150 cell lines. Although the two teams used a very similar MS-centric workflow, some differences exist between these two studies, especially in the depth of the analyses. While Pandey et al. performed around 2000 mass spectrometric (LC-MS) runs, Küster et al. carried out more than 6000 analyses and made use of another 10000 measurements publicly available in proteomic repositories. Assuming an average of two hours per run, the instrument time used to acquire these data would reach an astonishing number of 34000 h (4.3 years if only one mass spectrometer had been used). The analysis of all the data resulted in the identification of 946000 and 293000 non-redundant unique peptide sequences in Küster's and Pandey's studies, respectively. Strikingly, and despite the significant difference in depth, the two studies found evidence for a nearly identical number of protein-coding genes: 18097 (Küster) and 17294 (Pandey). Although a careful comparison of the two studies is still needed, a first conclusion can be drawn: the unequivocal existence of protein translation for 90–95% of the human genes. This is a highly relevant finding, as previously almost one-third of the human genes had been barely annotated, and there was no experimental evidence that they could lead to proteins. Another relevant discovery derived from these studies concerns the extent of alternative splicing in the generation of protein isoforms. It is clear that the number of genes does not correlate with the complexity of an organism (*C. elegans* for instance has 20500 genes) and it has been suggested that alternative splicing might increase the repertoire of functional proteins. However, these proteomic studies could only identify as many as 9000 of the 67000 isoforms annotated in Uniprot. Although some of these isoforms may produce only one unique peptide, decreasing the likelihood of observation by proteomics, these data could also support the idea that there is a dominant isoform per gene.[5] Both studies confirmed the existence of a core proteome present in all tissues, made up of "housekeeping

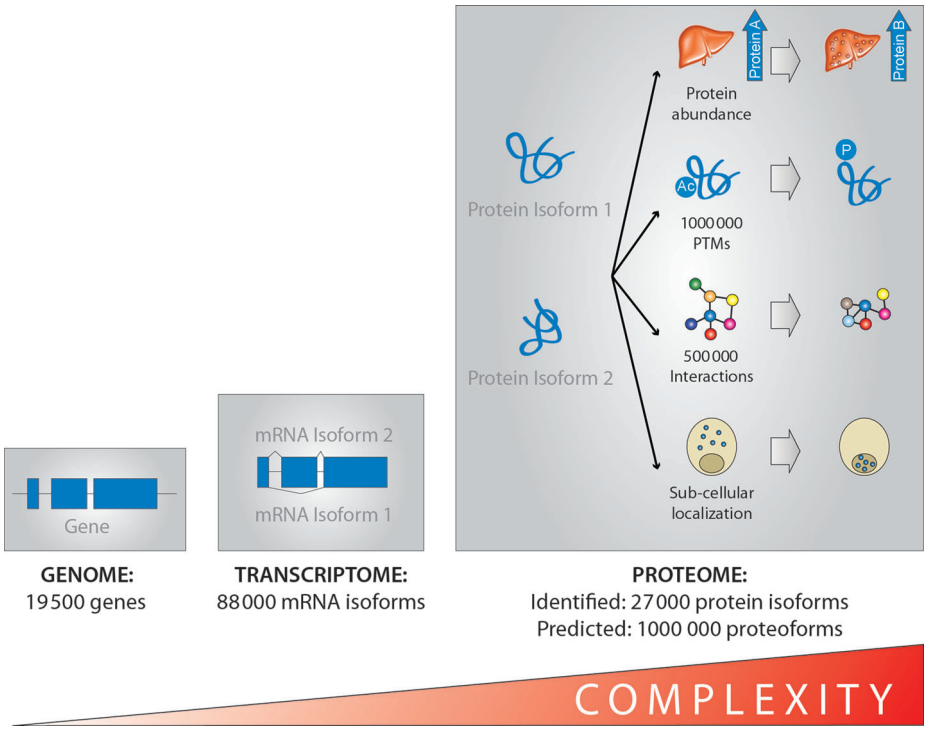

Figure 1. One gene, one protein? The complexity of the human proteome. The number of human protein-coding genes is estimated to be about 20000. Due to alternative splicing, the number of mRNAs transcribed is already significantly larger. Two recent studies have now provided the first drafts of the human proteome, demonstrating the translation of 18000 proteins (including more than 27000 isoforms). The complexity of the proteome is further amplified by tissue-specific expression, posttranslational modifications (PTMs), protein–protein interactions, and specific subcellular localization. These profiles are dynamically modified upon biological/pathological perturbations. It is estimated that more than 1000000 proteoforms (resulting from genetic variation, alternative splicing, and PTMs) may coexist in the human body.

proteins" (e.g. histones, ribosomal proteins, metabolic enzymes, and cytoskeletal proteins), representing often up to 75% of the total protein mass. On the other hand, expression of many tissue-specific proteins was observed, defining tissue proteomic signatures. Importantly, all these data have been made publicly accessible in two user-friendly portals (https://www.proteomicsdb.org und www.humanproteomemap.org), allowing the scientific community to navigate through these extended drafts of human proteomes.

With the development of high-throughput DNA and RNA sequencing technologies, the analysis of genomes and transcriptomes has been made possible at a reasonable speed and cost. This is in part due to the nature of DNA and RNA, consisting of molecules made up of only four nucleotides that are easily amplifiable by polymerases. Proteomes, on the other hand, are much more challenging from an analytical point of view. For instance, the 20 amino acids, building blocks of proteins, possess quite different physicochemical properties. Moreover, protein abundance can differ by ten orders of magnitude in cells and protein post-translational modifications (e.g. phosphorylation, glycosylation) and protein–protein interactions are highly dynamic in space and time. However, on par with gene and transcript sequencing, a "next-generation proteomics" is emerging mainly due to substantial improvements in analytical chemistry and informatics (sample preparation and separation), mass spectrometry, and data analysis; now the analysis of complex proteomes is possible at relatively low cost and high speed.[6] The most common proteomic workflow consists typically of 1) protein extraction, 2) digestion, 3) peptide separation, 4) mass spectrometric measurement, and 5) data analysis (Figure 2). Efficient peptide separation is key to achieving deep proteome coverage, as 1–2 million peptides are generated upon digestion of a human proteome. Multidimensional approaches coupling orthogonal separation techniques are essential. HPLC systems now work at very high pressure (15000 psi) to separate these peptides, taking advantage of long columns (>50 cm) with small internal diameters (<50 μm) packed with small particles (1.7 μm). This dramatically reduces sample complexity and addresses the dynamic-range problems that complicate subsequent peptide sequencing by mass spectrometry. Mass spectrometers have also evolved dramatically. Modern instruments have high resolving powers of 240000 FWHM (FWHM = full width at half maximum) to measure precursor and fragment ions with high accuracy and faster MS/MS peptide-sequencing speeds of 20 Hz. Different fragmentation techniques (e.g. CID, HCD, and ETD) can be used for specific peptide sequencing, maximizing identification rates. Not to be ignored are the parallel development in informatics, with a plethora of new software suites available that can convert the gigabytes of raw MS data into thousands of peptide identifications. The "yeast proteome race" illustrates this progress: 144 hours of MS analysis were needed in 2008 to identify yeast's 4000 proteins,[7] which could be shortened to just one hour in 2013.[8] Cumulatively, these parallel developments in enabling

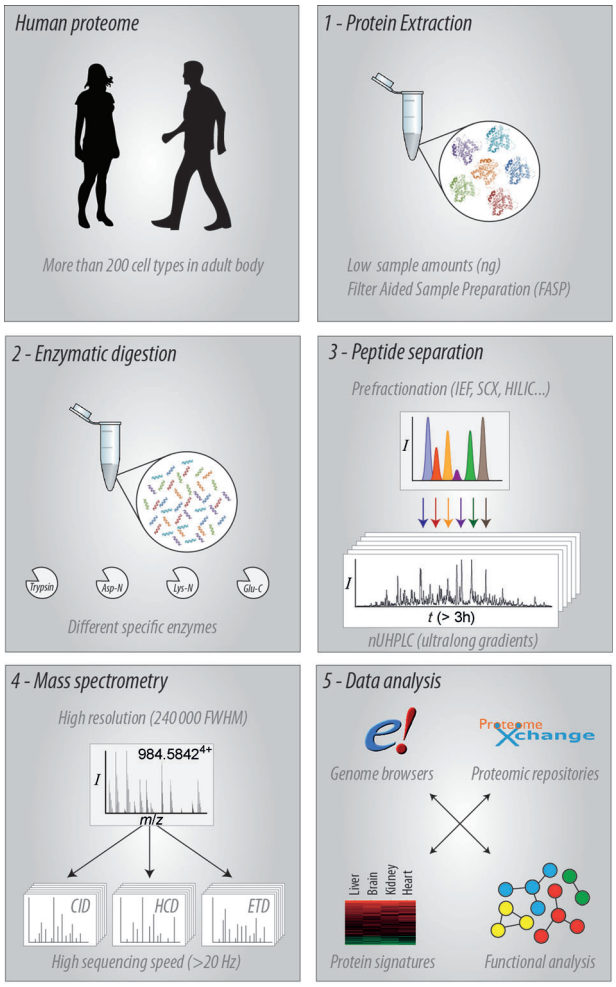

Figure 2. Typical proteomic workflow for the analysis of highly complex human proteomes. Some recent technological developments are indicated in gray. CID = collision-induced dissociation, ETD = electron transfer dissociation, HCD = higher-energy collision dissociation, IEF = isoelectric focusing, HILIC = hydrophilic interaction liquid chromatography, SCX = strong cation exchange.

technologies have facilitated these first drafts of the human proteome.

The presented maps represent quite extensive reference resources of the human proteome, but there still is a long way to go to understand the proteome in its full glory. Knowing that nearly each gene exists at the protein level does not tell us the function of each of these proteins, and how this function is regulated by the posttranslational modifications the protein harbors and the dynamic interactions of the protein with other molecular entities in the cell. A logical next step is to identify all posttranslational modifications (more than 200 types are known[9]) that decorate especially human proteins in order to start understanding how they regulate their functions, activities, and/or localizations. Similarly, reconstruction of all protein–protein interactions in a tissue-specific manner would generate protein networks that would help to identify new functional complexes and pathways. Alternative proteomic approaches will still be necessary in this venture. Top-down approaches analyze by MS intact proteins without prior digestion so more accurate knowledge about proteoforms and the co-occurrence of PTMs will be gathered.[10] Furthermore, antibody-based profiling can be used to decipher the localized expression of proteins in different cell types.[11] Finally, the analysis of dynamic proteome remodeling upon biological perturbations will provide essential information on the biology of the human cells. The availability of highly relevant "proteomic big data" on par with other "-omics" data will be the key to success for personalized or precision medicine.

Received: June 24, 2014
Published online: July 30, 2014

[1] a) E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al., *Nature* **2001**, *409*, 860–921; b) J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al., *Science* **2001**, *291*, 1304–1351.
[2] E. Pennisi, *Science* **2012**, *337*, 1159–1161.
[3] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, et al., *Nature* **2014**, *509*, 582–587.
[4] M.-S. Kim, 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., *Nature* **2014**, *509*, 575–581.
[5] M. Gonzàlez-Porta, A. Frankish, J. Rung, J. Harrow, A. Brazma, *Genome Biol.* **2013**, *14*, R70.
[6] a) A. F. M. Altelaar, J. Munoz, A. J. R. Heck, *Nat. Rev. Genet.* **2013**, *14*, 35–48; b) A. Bensimon, A. J. Heck, R. Aebersold, *Annu. Rev. Biochem.* **2012**, *81*, 379–405; c) J. Cox, M. Mann, *Annu. Rev. Biochem.* **2011**, *80*, 273–299; d) J. R. Yates, C. I. Ruse, A. Nakorchevsky, *Annu. Rev. Biomed. Eng.* **2009**, *11*, 49–79.
[7] L. M. F. de Godoy, J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Fröhlich, T. C. Walther, M. Mann, *Nature* **2008**, *455*, 1251–1254.
[8] A. S. Hebert, A. L. Richards, D. J. Bailey, A. Ulbrich, E. E. Coughlin, M. S. Westphall, J. J. Coon, *Mol. Cell. Proteomics* **2014**, *13*, 339–347.
[9] C. T. Walsh, S. Garneau-Tsodikova, G. J. Gatto, *Angew. Chem.* **2005**, *117*, 7508–7539; *Angew. Chem. Int. Ed.* **2005**, *44*, 7342–7372.
[10] J. C. Tran, L. Zamdborg, D. R. Ahlf, J. E. Lee, A. D. Catherman, K. R. Durbin, J. D. Tipton, A. Vellaichamy, J. F. Kellie, M. Li, et al., *Nature* **2011**, *480*, 254–258.
[11] M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester, S. Hober, et al., *Nat. Biotechnol.* **2010**, *28*, 1248–1250.

