Biofísica Magazine

COOL BIOPHYSICS

The race towards the human proteome

Jesús Vázquez, *CNIC (Madrid)*

While current genomics approaches allow the eventual analysis of all human genes, the ability to analyze all their protein products has traditionally been considered a utopian dream. Proteins cannot be amplified like genes, and mass spectrometry (MS) —the most powerful approach for protein analysis— is hampered by the large dynamic range of human protein concentrations. However, these limitations are already being overturned by the remarkable advances made in MS in the past few years, in terms of sensitivity, mass accuracy and, above all, speed.

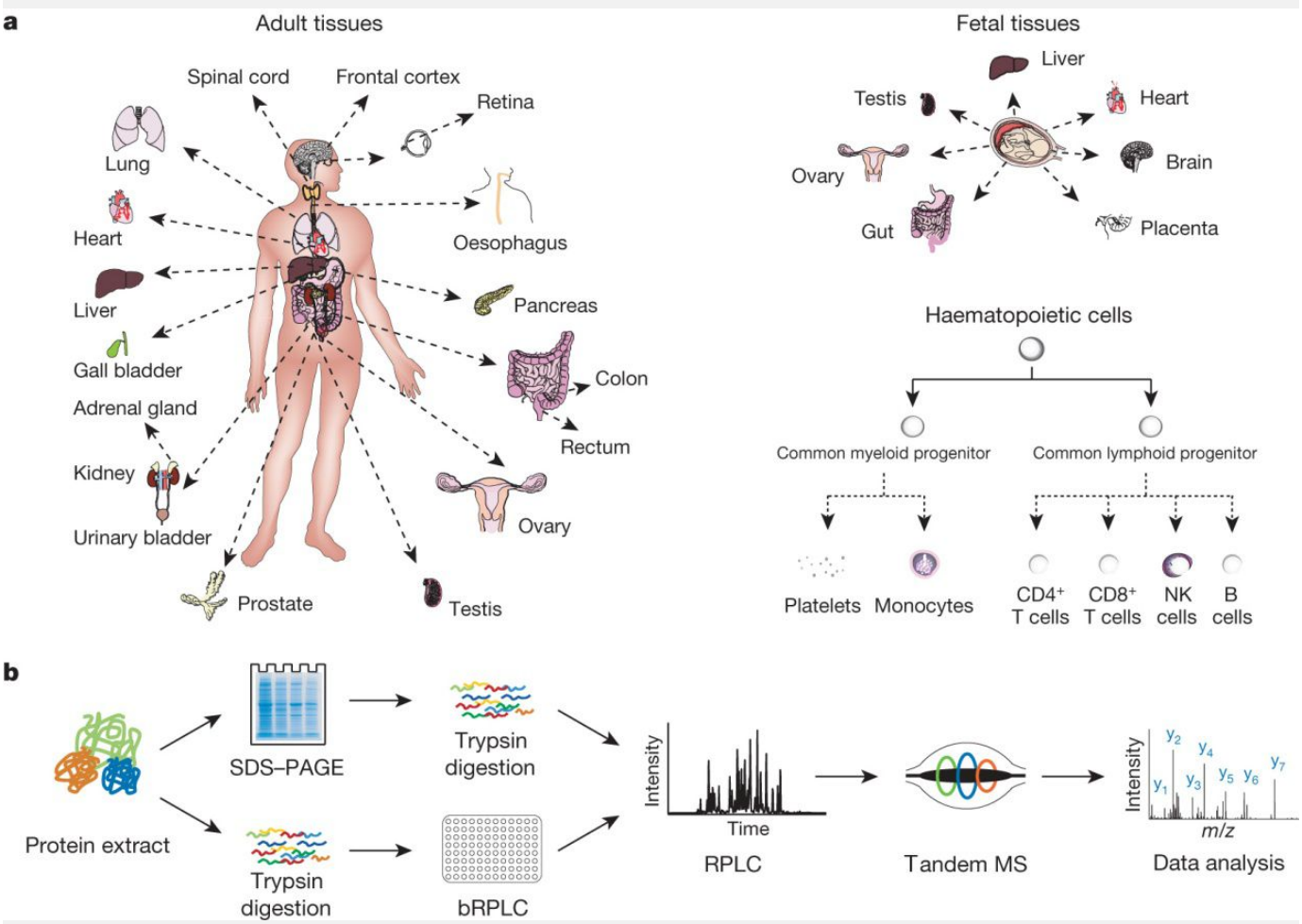
Early efforts: The proteome of specific cell lines

The first comprehensive analyses of what was at the time considered the *complete* proteome of two human cell lines (~10.000 distinct proteins) were published simultaneously in **2011** by the groups of Mann (HeLa cells, Nagaraj, et al. 2011) and Aebersold (U2OS cells, Beck, et al. 2011). This was possible by subjecting the peptidome (the result of digesting the proteins with a protease, making the products suitable to MS analysis) to considerable fractionation, and analyzing all the fractions using state-of-the art MS. Although these studies overthrew the dogma that proteins present in very few copies per cell would never be accessible to MS, subsequent experience demonstrated that surpassing the four-digit barrier was a task beyond the reach of most proteomics laboratories.

Drafts of the human proteome

The first analyses of what we could call the *human proteome* were published simultaneously in **2014** by two large laboratory consortia. One of them, led by Pandey (Kim, et al. 2014), was composed mostly of North American and Indian laboratories, while the other was composed of German groups and was directed by Kuster (Wilhelm, et al. 2014). The Pandey study used protein extracts from 30 human tissues (of normal, adult and fetal origin) and 6 hematopoietic cell lines. Each of the extracts was subjected to extensive fractionation at both the protein and peptide levels; the effort required more than 2,000 chromatography runs, which is about 20-fold more than the number

required previously to map a single human cell type, and identified ca. 17,300 gene products. In the Kuster study, 60% of the information was taken from existing MS repositories and a further 40% was new data generated by the authors. Analysis of the equivalent of 17,000 runs allowed the identification of 18,000 proteins. By that time, in a joint bioinformatics effort by two Spanish groups from the CNIO and the CNIC, in collaboration with two US groups, we were able to narrow down the number of potentially coding human genes to 19,000 (Ezkurdia, et al. 2014). The Pandey and Kuster studies were thus able to map approximately 95% of the human proteome.



Workflow and comparison of human proteome data with public repositories Used with permisson from the nature publishing group, M.S. Kim, et al. Nature 509: 575-581 (2014) doi:10.1038/nature13302. © 2014 by the nature publishing group.

These two reports provided the proteomics community with a rich source of information and their preliminary analysis yielded some interesting insights. For instance, the comparison of protein abundance profiles across all tissues and cell lines allowed the construction of the ***"housekeeping proteome"***: a list of 2,300 proteins ubiquitously and constitutively expressed in all situations, the primary function of which is the general control and maintenance of cells. Interestingly, this proteome is composed mainly of histones, ribosomal proteins, metabolic enzymes and cytoskeletal proteins, and constitutes approximately 75% of total protein mass. However, these proteins show across-tissue expression differences of up to five orders of magnitude, and the number of proteins

that are exclusively or preferentially detected in a particular organ is surprisingly small. This suggests that differences in biological function are achieved by a slightly varying set of proteins whose relative proportions are adjusted in a site-specific manner to achieve tailored biological roles. Similarly, studying the correlation of protein abundance across tissues provided important insights into the composition of complexes and the dynamic nature of protein-protein interactions. The huge amounts of protein and MS data gathered was also used for proteogenomics studies, in which, for instance, extensive analysis was conducted of the mechanisms through which protein abundance is regulated from the corresponding mRNA transcript and of the presence of isoforms, novel protein-coding regions (pseudogenes, non-coding RNAs, upstream ORFs and alike) and protein N-termini. An interesting outcome, which seems to confirm recent viewpoints, was that while mRNA abundance was a very poor predictor of protein abundance, the ratio of mRNA to protein tended to be protein-specific and was remarkably conserved across tissues, suggesting that the translation rate is characteristic of each transcript and that protein abundance in the cell is controlled predominantly at the transcriptional level.

Concerns and refinement

Very soon after the publication of these reports, criticisms were raised in the proteomics community. The first report was, again, a joint collaboration between CNIO and CNIC (Ezkurdia, et al. 2014), in which we demonstrated a lack of rigor in protein identification in the Pandey and Kuster studies. We concentrated our analysis on the olfactory receptor family, an interesting group of vertebrate-specific genes whose transcription levels are very low and which are confined to nasal tissue, not included in the earlier studies. Unexpectedly, we found 108 olfactory receptors in the Kuster report and 200 in the Pandey report. Most of the identified OR peptides shared their sequence with other proteins, and a great proportion of peptide spectra were of surprisingly low quality. Our results highlighted serious problems in the control of peptide-to-protein redundancy and the false discovery rate (FDR) of protein identification, the latter probably a consequence of the accumulation of large amounts of data and the use of very narrow precursor mass windows, as we pointed out recently (Bonzon-Kulichenko, et al. 2015).

Some proverbs say that humans are the only animals that make the same errors twice, and this story is reminiscent of past errors committed when the human genome was first published, and of the advent of high-throughput, MS-based peptide identification at the beginning of this century. The numbers of identified proteins were growing so fast that in the rush to beat records, proteomicians ignored the fact that the protein identification criteria were not extrapolatable from one situation to another. The result was that a large proportion of identifications reported at that time were incorrect. This situation was resolved by the introduction of decoy databases and the use of FDR at the peptide level, providing a robust and reproducible means of controlling identification quality that is still a must today.

Future directions

The current accumulation of proteins from large repositories and MS studies has produced what is called the *protein buildup FDR problem*, which, incidentally, was acknowledged in the Kuster report to be an *unsolved issue* (Wilhelm, et al. 2014). The conclusion is that there is an urgent need to resolve the protein FDR problem before we begin to trust the huge amounts of protein information produced in these collaborative efforts. Fortunately, the Human Proteome Organization is fully aware of the problem and is dedicating resources in this direction. This time round the Spanish proteomics community is fully implicated through the ProteoRed bioinformatics network, where we are working on the development of a novel integrative identification framework.

Jesús Vázquez
Cardiovascular Proteomics Laboratory and Proteomics Unit
Department of Vascular Biology and Inflammation
Centro Nacional de Investigaciones Cardiovasculares – CNIC
C/ Melchor Fernández Almagro, N. 3, 28029 Madrid
E-mail: jesus.vazquez@cnic.es

References

Beck M, Schmidt A, Malmstroem J, Claassen M, Ori A, Szymborska A, Herzog F, Rinner O, Ellenberg J and Aebersold R. "The quantitative proteome of a human cell line".*Mol Syst Biol*, **2011**, 7: 549. DOI:10.1038/msb.2011.82.

Bonzon-Kulichenko E, Garcia-Marques F, Trevisan-Herraz M and Vázquez J "Revisiting Peptide Identification by High-Accuracy Mass Spectrometry: Problems Associated with the Use of Narrow Mass Precursor Windows". *J Proteome Res*, **2015**, 14: 700. DOI:10.1021/pr5007284.

Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vázquez J, Valencia A and Tress ML"Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes". *Hum Mol Genet*, **2014**, 23: 5866. DOI: 10.1093/hmg/ddu309.

Ezkurdia I, Vázquez J, Valencia A and Tress M "Analyzing the First Drafts of the Human Proteome".*J Proteome Res*, **2014**, 13: 3854. DOI: 10.1021/pr500572z.

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS ... and Pandey GA"A draft map of the human proteome". *Nature*, **2014**, 509: 575. DOI:10.1038/nature13302.

Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Pääbo S and Mann M"Deep proteome and transcriptome mapping of a human cancer cell line". *Mol Syst Biol*, **2011**, 7: 548. DOI:10.1038/msb.2011.81.

Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M ... and Kuster B"Mass-spectrometry-based draft of the human proteome". *Nature*, **2014**, 509: 582. DOI:10.1038/nature13319.