SCIENCE CHINA
Life Sciences

• INSIGHT •

October 2014 Vol.57 No.10: 1039–1040
doi: 10.1007/s11427-014-4738-7

Is the mission to identify all the human proteins achievable? —Commenting on the human proteome draft maps

CHANG ZengYi[1*] & GU LiangCai[2]

[1]*Center for Protein Science, State Key Laboratory of Protein and Plant Gene Studies, School of Life Sciences, Center for History and Philosophy of Science, Peking University, Beijing 100871, China;*
[2]*Department of Genetics, Harvard Medical School, Boston 02115, USA*

Received August 10, 2014; accepted August 12, 2014; published online August 22, 2014

Citation: Chang ZY, Gu LC. Is the mission to identify all the human proteins achievable?—Commenting on the human proteome draft maps. Sci China Life Sci, 2014, 57: 1039−1040, doi: 10.1007/s11427-014-4738-7

Unveiling the structure and function of all the proteins a human body produces is undoubtedly a key task for us to better understand ourselves. Although a mission difficult to achieve, this apparently becomes more feasible upon the determination of the DNA sequence of the whole human genome, which was predicted to encode a total of 20,000–25,000 proteins [1]. Recent efforts of direct (*de novo*) protein isolation and peptide sequencing, mainly using liquid chromatography-tandem mass spectrometry (LC-MS/MS) and in-depth bioinformatics analyses, led to the apparent identification of over 17,000 proteins from large sets of human tissues, cell lines and body fluids [2,3].

Remarkably, 2,535 (~15%) of the 17,294 total human proteins detected by MS were found to be encoded by DNA sequences previously defined as non-protein coding ones, e.g., pseudogenes, non-coding RNAs, 5′ or 3′ untranslated regions, thus escaped any kind of previous notification [2]. Given that the direct protein isolation and identification of human proteins by Kim et al. [2] and Wilhelm et al. [3] was intrinsically incomplete, more novel proteins will be undoubtedly discovered in the future with further improvement of methodologies.

One outstanding unresolved issue here is that, although these protein profiling efforts provided direct evidences to confirm the presence of ~84% of the approximately 20,000 previously annotated protein-coding genes [2], the rest 16% of them were left unconfirmed. They may have escaped the detection due to their absence or low abundance in the tested samples, or being inaccessible to mass spectrometry analysis which relies on the protein digestion with limited choices of proteases, or none-existence (i.e., of incorrect prediction). The direct confirmation of these putative proteins that failed the detection by these first rounds of attempts is certainly highly desired but greatly challenging.

More noteworthy is the fact that approximately 16 million out of the 25 million peptide mass spectra obtained failed to match any of the human proteins hitherto annotated [2]. After an expanded search, although a small portion of these 16 million peptides were matched to the human genome sequence, leading to the discovery of 2,535 novel protein-coding genes. Nevertheless, still a tremendous number of these detected peptides, assuming to be real, await to be matched to proteins encoded by the human genome. The detection of such a great number of ‘orphan’ peptides apparently indicates that many more polypeptides might be encoded, transcribed, translated or modified in the human body via currently unknown mechanisms. It follows that the grammar we so far learned about the DNA language of the human genome is far from complete and we have to find new ways to read it, as attempted by Vanderperre et al. [4].

Without doubt, to work out the full picture of the human proteome we still have a long way to go. The parallel Human Protein Altas project seeks to generate antibodies for

*Corresponding author (email: changzy@pku.edu.cn)

each of the predicted putative human proteins using the information of their predicted amino acid sequences, and then apply the antibodies to detect whether the putative proteins actually exist in human samples [5]. This antibody-based approach, though serving as a parallel and likely more definitive way to demonstrate whether those annotated human proteins are genuine, would be hardly effective in finding out the missing proteins to which this large number of unmatched 'orphan' peptides might be derived from. In this regard, one potential way to demonstrate from what human protein these unmatched peptides are derived is to make antibodies against these putative peptides and use them to examine the human samples. But this approach would at best only be partially effective, because usually only one peptide sequence is available for antibody production for each putative protein (multiple of these peptides might be derived from one single protein, but this information is not known beforehand), whereas at least two antibodies each of which binds to a different epitope would be desired for detecting one putative protein [5].

In contrast to the Human Genome, which is largely identical in each of our cells and thus a clear finish line can be drawn for its complete sequencing, we are not clear where the end point is for gaining a full picture of the human proteome, which differs in each different type of cells and tissues, and during different developmental stages or under various life conditions. To make things worse, many human proteins are produced in extremely low abundance, and/or only in a rather dynamic time- and space-dependent manner. Without doubt, the identification and characterization of all the proteins that the human body produces is certainly a mission difficult to achieve, if achievable at all [6].

1 The International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, 2004, 431: 931–945

2 Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, Thomas JK, Muthusamy B, Leal-Rojas P, Kumar P, Sahasrabuddhe NA, Balakrishnan L, Advani J, George B, Renuse S, Selvan LD, Patil AH, Nanjappa V, Radhakrishnan A, Prasad S, Subbannayya T, Raju R, Kumar M, Sreenivasamurthy SK, Marimuthu A, Sathe GJ, Chavan S, Datta KK, Subbannayya Y, Sahu A, Yelamanchi SD, Jayaram S, Rajagopalan P, Sharma J, Murthy KR, Syed N, Goel R, Khan AA, Ahmad S, Dey G, Mudgal K, Chatterjee A, Huang TC, Zhong J, Wu X, Shaw PG, Freed D, Zahari MS, Mukherjee KK, Shankar S, Mahadevan A, Lam H, Mitchell CJ, Shankar SK, Satishchandra P, Schroeder JT, Sirdeshmukh R, Maitra A, Leach SD, Drake CG, Halushka MK, Prasad TS, Hruban RH, Kerr CL, Bader GD, Iacobuzio-Donahue CA, Gowda H, Pandey A. A draft map of the human proteome. Nature, 2014, 509: 575–581

3 Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, Mathieson T, Lemeer S, Schnatbaum K, Reimer U, Wenschuh H, Mollenhauer M, Slotta-Huspenina J, Boese JH, Bantscheff M, Gerstmair A, Faerber F, Kuster B. Mass-spectrometry-based draft of the human proteome. Nature, 2014, 509: 582–587

4 Vanderperre B, Lucier JF, Bissonnette C, Motard J, Tremblay G, Vanderperre S, Wisztorski M, Salzet M, Boisvert FM, Roucou X. Direct detection of alternative open reading frames translation products in human significantly expands the proteome. PLoS One, 2013, 8: e70698

5 Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, Zwahlen M, Kampf C, Wester K, Hober S, Wernerus H, Björling L, Ponten F. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol, 2010, 28: 1248–1250

6 Leng FW. Opportunity and challenge: ten years of proteomics in China. Sci China Life Sci, 2012, 55: 837–839