SCIENCE CHINA
Life Sciences

• INSIGHT •

March 2014 Vol.57 No.3: 358–360
doi: 10.1007/s11427-014-4618-1

How to discover new proteins—translatome profiling

ZHANG Gong*, WANG Tong* & HE QingYu*

College of Life Science and Technology, Jinan University, Guangzhou 510632, China

Received November 27, 2013; accepted December 22, 2013; published online February 14, 2014

Citation: Zhang G, Wang T, He QY. How to discover new proteins—translatome profiling. Sci China Life Sci, 2014, 57: 358−360, doi: 10.1007/s11427-014-4618-1

According to the Central Dogma, genes in genome can be transcribed and translated into proteins to perform a variety of biological functions. Since proteins are the actual functional molecules in almost all physiological processes, discovering new proteins is essential to understand the nature of various biological phenomena. So far, the Human Genome Project and many transcriptomic analyses have defined about 20300 genes that can be transcribed into mRNA. These genes, comprising only 2% of human genome, are considered as coding genes that may be translated into proteins.

Are the other 98% human genome really non-coding? Or there may be numerous annotation errors in human genome? The definite answer to these questions is to provide evidence of the presence or absence of those genes at the protein level. However, the expression of proteins varies temporally and spatially. Some proteins are expressed specifically in one tissue but not in another; other proteins are expressed only under specific physiological conditions. Thus, many proteins have not been ever detected.

The Human Proteome Project (HPP) was therefore proposed by the Human Proteome Organization (HUPO), aiming at identifying all proteins encoded by coding genes and determining their distribution in tissues, cells, sub-cellular organelles, with various physiological and pathological phenotypes [1–3]. One important ambition of HPP is to identify "missing proteins" [1]. Generally, missing proteins refer to two classes of proteins, namely "undetected proteins" and "new (novel) proteins". Undetected proteins are those encoded by the coding genes but their protein evidence has not been detected by current experimental technologies although they may actually exist in samples. New (novel) proteins are those encoded by erroneously annotated "non-coding genes".

Conventional methods to directly identify proteins include mass spectrometry and antibody-based immunological detection. Using mass spectrometry, peptide fragments from a digested protein can be detected as mass spectral signals, which are matched to the theoretical peptide mass spectrum based on the protein sequence database to identify the protein. Currently, advanced mass spectrometry can identify approximately 12000 proteins in single human cells; the detection sensitivity can even be improved by using SRM/MRM technology [4,5]. However, mass spectrometry has its inherent limitations in protein identification, including the requirement for the large amount of samples, low identification rates for low-abundance proteins and low reproducibility from different laboratories [6,7]. Moreover, the physical and chemical properties of proteins also affect their detectability in mass spectrometry. Some proteins, including hydrophobic membrane proteins, structurally compact proteins that are resistant to enzyme digestion, and easily degradable proteins, are difficult to be detected. These problems lead to the low peptide coverage: only a few peptides of a certain protein are detected, although the protein may have been digested into many peptides.

Complementarily, HPP proposed antibody-based detection as an alternative tool, which indeed made great progress in protein identification [1,2,8]. Nevertheless, such a method may encounter low-staining situation, puzzling the

*Corresponding author (email: zhanggong@jnu.edu.cn; tongwang@jnu.edu.cn; tqyhe@jnu.edu.cn)

detection of target proteins. In addition, an antibody typically recognizes a specific epitope of a protein, covering a limited part of the protein sequence. More importantly, a pre-requirement for both mass spectrometry and antibody-based verification is the establishment of protein sequence database. For those new proteins and mutants resulting from single nucleotide variation (SNV) and RNA editing, their sequences either are absent in or differ from those in the database. Thus, both detection methods have limitation for the identification of new proteins.

Transcriptome sequencing can provide indirect clue of possible coding mRNAs in cells under steady-state. It can also detect mRNA sequence variations and splice variants [4,9–14]. However, transcriptome sequencing still has crucial drawbacks even with substantially developed next-generation sequencing techniques (NGS). Numerous studies have found untranslated mRNA species in various human cell lines [15–18], and these species comprise approximately 5% of the transcriptome [2,14]. Transcriptome sequencing does not distinguish the translating and untranslated RNA. Therefore, this strategy cannot give information to guide the discovery of new proteins.

To correct the possible annotation errors in genome and to discover new proteins, we proposed a new strategy that analyzes the translating mRNA (translatome sequencing) to find the translating evidence of known and unknown coding genes [14,19]. It is known that proteins are synthesized via translation, therefore the translating evidence of proteins, obtained by translatome sequencing and sequence analysis, can be a sensitive method to find new proteins. This method purifies ribosome-nascent-chain complex (RNC) and sequences the mRNA in this fraction. The RNC purification is the key step that can be performed efficiently using sucrose gradient ultracentrifugation [14]. Under steady-state, protein species can exist only via translation, i.e., translating mRNAs most likely correspond to proteins. This realizes using sequencing technology to study proteins.

Full-length translating mRNA sequencing (RNC-seq) exhibits unique advantages against traditional protein identification methods. The high sensitivity of NGS ensures high identification efficiency. Under normal sequencing throughput (10–20 million reads), it is possible to identify and quantify more than 14000 RNC-mRNA in RefSeq database from single human cell lines, by using highly accurate FANSe series mapping algorithm [14,19,20]. The power of quantification and identification exceeds the mass spectrometry considerably at low cost. NGS is independent of physical and chemical properties of proteins, therefore, translatome sequencing is easier to provide translation evidence to those proteins that are difficult to be identified by mass spectrometry. Another advantage of translatome sequencing is to identify enough number of protein coding genes at lower starting quantity of samples and lower throughput. For example, the down sampling of translatome sequencing dataset of human hepatocarcinoma cells Hep3B revealed that two million reads can identify more than 11000 genes [21]. Translatome sequencing can provide definitive translation evidence of proteins and avoids the "low staining" problem of antibody detection. Therefore, it is effective to detect cell/tissue-specific expressed proteins. For example, we identified 750 and 2105 genes that are specifically expressed in human bronchial epithelial cells HBE and human colorectal cancer cell Caco-2 [19].

Taking the advantage of the high-throughput NGS, translatome sequencing achieves high sequence coverage. For mid-to-low abundance RNC-mRNAs (10–50 rpkM), almost 100% sequence coverage can be typically reached [19]. The third-generation sequencing technology is theoretically possible to directly sequence the full-length mRNA. This covering power consolidates the identification. It is also possible to detect any sequence variation and alternative spliced transcripts by translatome sequencing, thus detecting the missing proteins effectively.

Translatome sequencing does not rely on the database annotation. Therefore, conventional non-coding RNA (ncRNA) may be identified in the RNC-mRNA fraction. For example, we found 1397 genes that were annotated as ncRNA in RefSeq database in Caco-2 translating mRNA, indicating that these genes are being translated [19]. Among them, *HMGB3P1* and *ESRG* are traditional ncRNAs but were found in the RNC-mRNA fraction in multiple human cell lines including HBE, A549, H1299 and Caco-2 [14,19]. As an example of new protein discovery, *ESRG* has been found to code HESRG protein [22,23].

In the HPP, HUPO suggested three resource pillars for human proteome studies and missing protein identification: mass spectrometry, bioinformatics and antibody [1]. Currently, we have proposed that translatome sequencing can serve as the fourth resource pillar for HPP [19]. This strategy has already been applied in the collaboration work carried out by Chinese HPP Consortium (JPR special issue, 2014). We believe that translatome sequencing is an important complementary method of traditional proteomics, and can independently lead to discovery of new proteins functioning in various physiological and pathological conditions.

1 Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee HJ, Na K, Choi EY, Yan F, Zhang F, Zhang Y, Snyder M, Cheng Y, Chen R, Marko-Varga G, Deutsch EW, Kim H, Kwon JY, Aebersold R, Bairoch A, Taylor AD, Kim KY, Lee EY, Hochstrasser D, Legrain P, Hancock WS. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotechnol, 2012, 30: 221–223

2 Paik YK, Omenn GS, Uhlen M, Hanash S, Marko-Varga G, Aebersold R, Bairoch A, Yamamoto T, Legrain P, Lee HJ, Na K, Jeong SK, He F, Binz PA, Nishimura T, Keown P, Baker MS, Yoo JS, Garin J, Archakov A, Bergeron J, Salekdeh GH, Hancock WS. Standard guidelines for the chromosome-centric human proteome project. J Proteome Res, 2012, 11: 2005–2013

3 Hühmer AF, Paulus A, Martin LB, Millis K, Agreste T, Saba J, Lill JR, Fischer SM, Dracup W, Lavery P. The chromosome-centric human proteome project: a call to action. J Proteome Res, 2013, 12:

28–32
4 Nagaraj N, Wisniewski JR, Geiger T, Cox J, Kircher M, Kelso J, Pääbo S, Mann M. Deep proteome and transcriptome mapping of a human cancer cell line. Mol Syst Biol, 2011, 7: 548
5 Wu S, Li N, Ma J, Shen H, Jiang D, Chang C, Zhang C, Li L, Zhang H, Jiang J, Xu Z, Ping L, Chen T, Zhang W, Zhang T, Xing X, Yi T, Li Y, Fan F, Li X, Zhong F, Wang Q, Zhang Y, Wen B, Yan G, Lin L, Yao J, Lin Z, Wu F, Xie L, Yu H, Liu M, Lu H, Mu H, Li D, Zhu W, Zhen B, Qian X, Qin J, Liu S, Yang P, Zhu Y, Xu P, He F. First proteomic exploration of protein-encoding genes on chromosome 1 in human liver, stomach, and colon. J Proteome Res, 2013, 12: 67–80
6 Thompson AJ, Abu M, Hanger DP. Key issues in the acquisition and analysis of qualitative and quantitative mass spectrometry data for peptide-centric proteomic experiments. Amino Acids, 2012, 43: 1075–1085
7 Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, Sechi S, Nilsson T, Bergeron JJ; HUPO Test Sample Working Group. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. Nat Methods, 2009, 6: 423–430
8 Fagerberg L, Oksvold P, Skogs M, Algenäs C, Lundberg E, Pontén F, Sivertsson A, Odeberg J, Klevebring D, Kampf C, Asplund A, Sjöstedt E, Al-Khalili Szigyarto C, Edqvist PH, Olsson I, Rydberg U, Hudson P, Ottosson Takanen J, Berling H, Björling L, Tegel H, Rockberg J, Nilsson P, Navani S, Jirström K, Mulder J, Schwenk JM, Zwahlen M, Hober S, Forsberg M, von Feilitzen K, Uhlén M. Contribution of antibody-based protein profiling to the human Chromosome-centric Proteome Project (C-HPP). J Proteome Res, 2013, 12: 2439–2448
9 Chen G, Gharib TG, Huang CC, Taylor JM, Misek DE, Kardia SL, Giordano TJ, Iannettoni MD, Orringer MB, Hanash SM, Beer DG. Discordant protein and mRNA expression in lung adenocarcinomas. Mol Cell Proteomics, 2002, 1: 304–313
10 Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F. Annotating genomes with massive-scale RNA sequencing. Genome Biol, 2008, 9: R175
11 Maier T, Guell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS Lett, 2009, 583: 3966–3973
12 Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, Algenäs C, Lundeberg J, Mann M, Uhlen M. Defining the transcriptome and proteome in three functionally different human cell lines. Mol Syst Biol, 2010, 6: 450
13 Akan P, Alexeyenko A, Costea PI, Hedberg L, Solnestam BW, Lundin S, Hällman J, Lundberg E, Uhlén M, Lundeberg J. Comprehensive analysis of the genome transcriptome and proteome landscapes of three tumor cell lines. Genome Med, 2012, 4: 86
14 Wang T, Cui Y, Jin J, Guo J, Wang G, Yin X, He QY, Zhang G. Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. Nucleic Acids Res, 2013, 41: 4743–4754
15 Thireos G, Griffin-Shea R, Kafatos FC. Untranslated mRNA for a chorion protein of *Drosophila melanogaster* accumulates transiently at the onset of specific gene amplification. Proc Natl Acad Sci USA, 1980, 77: 5789–5793
16 Standart N, Hunt T, Ruderman JV. Differential accumulation of ribonucleotide reductase subunits in clam oocytes: the large subunit is stored as a polypeptide, the small subunit as untranslated mRNA. J Cell Biol, 1986, 103: 2129–2136
17 Nielsen FC, Ostergaard L, Nielsen J, Christiansen J. Growth-dependent translation of IGF-II mRNA by a rapamycin-sensitive pathway. Nature, 1995, 377: 358–362
18 Khan D, Sharathchandra A, Ponnuswamy A, Grover R, Das S. Effect of a natural mutation in the 5' untranslated region on the translational control of p53 mRNA. Oncogene, 2013, 32: 4148–4159
19 Zhong J, Cui Y, Guo J, Chen Z, Yang L, He QY, Zhang G, Wang T. Resolving chromosome-centric human proteome with translating mRNA analysis: a strategic demonstration. J Proteome Res, 2014, 13: 50–59
20 Zhang G, Fedyunin I, Kirchner S, Xiao C, Valleriani A, Ignatova Z. FANSe: an accurate algorithm for quantitative mapping of large scale sequencing reads. Nucleic Acids Res, 2012, 40: e83
21 Zhang C, Li N, Zhai L, Xu S, Liu X, Cui Y, Ma J, Han M, Jiang J, Yang C, Fan F, Li L, Qin P, Yu Q, Chang C, Su N, Zheng J, Zhang T, Wen B, Zhou R, Lin L, Lin Z, Zhou B, Zhang Y, Yan G, Liu Y, Yang P, Guo K, Gu W, Chen Y, Zhang G, He QY, Wu S, Wang T, Shen H, Wang Q, Zhu Y, He F, Xu P. Systematic analysis of missing proteins provides clues to help define all of the protein-coding genes on human chromosome 1. J Proteome Res, 2014, 13: 114–125
22 Zhao M, Ren C, Yang H, Feng X, Jiang X, Zhu B, Zhou W, Wang L, Zeng Y, Yao K. Transcriptional profiling of human embryonic stem cells and embryoid bodies identifies HESRG, a novel stem cell gene. Biochem Biophys Res Commun, 2007, 362: 916–922
23 Wanggou S, Jiang X, Li Q, Zhang L, Liu D, Li G, Feng X, Liu W, Zhu B, Huang W, Shi J, Yuan X, Ren C. HESRG: a novel biomarker for intracranial germinoma and embryonal carcinoma. J Neurooncol, 2012, 106: 251–259