Proteoforms scouting in HepG2 cell line

Olga Kiseleva, Ekaterina Poverennaya

Institute of biomedical chemistry, 10/8 Pogodinskaya str., Moscow, Russia, olly.kiseleva@gmail.com

Biochemical processes occurring in living organisms are to a large extent conditioned by genomic laws but are not limited to them. Investigation of genes and transcripts often provides only insufficient information of a probabilistic nature. Thus, the results of the studies require validation at the proteome level [1]. Alternative splicing, single amino acid polymorphisms, as well as a number of spontaneous and natural post-translational events, lead to changes in the structure and functions of aberrant protein variants, thereby enabling the heterogeneity of the proteome. Sketchy data on proteome complexity and heterogeneity provide information only about some «average» protein sequence. Critical analysis of heterogeneous proteome can be a basis for understanding the functioning of complex biochemical systems. Considerable progress of multiomics approaches opened up new avenues for accurate, specific and high-performance protein analysis.

To study the diversity of human proteins species (proteoforms) we carried out a comprehensive research of gene expression for HepG2 cell line by transcriptomic and proteomic methods.

We used two independent NGS platforms (Illumina and SOLiD) to obtain transcriptomic data. Strong correlation (R^2=0.9) between gene expression levels, produced separately by each sequencing platform, allowed us to consolidate the results [2]. We referred to translated Ensembl proteins (GRCh38 release 80) and applied PPLine, which integrates a set of popular tools: Trimmomatic, Tophat2, samtools, GATK, Cufflinks, and Annovar [3].

Transcriptomic profiling revealed totally 75 thousand aberrant hits (substitutions, deletions and insertions), in which 1008 hits correspond to human chromosome 18 - object chromosome of Russian part of international «Human Proteome Project». Third part of hits was annotated in UniProtKB. More than 130 alternative splicing events were revealed for 89 protein coding genes, 37 of these events were detected for the first time ever. Transcriptomic analysis of HepG2 cell line resulted in customized database, containing 52 thousand amino acid sequences, encoded by 12 thousand genes.

Proteome profiling by advanced 2DE (separating gel into 96 cells containing corresponding protein spots) and further MS analysis of HepG2 sample was performed. Herewith only 18 thousand different proteins are available to be visualized by means of gel electrophoresis.

For scouting of proteoforms using MS we analyzed obtained MS/MS data (192 raw files of total volume 114 Gb) with variable computing machines, settings, databases, and combinations of search algorithms to overcome difficulties caused by low protein concentration in some cells of 2DE gel. To select an effective solution for customized search strategy, realized in an open-source graphical user interface SearchGUI [4], we optimized set of search engines. We selected combination of X!Tandem, MS-GF+ and OMMSA as the most time-efficient and productive combination of search engines and compared it with Mascot results using test-kit UPS-2, containing 48 human proteins. We also added homemade java-script to automatize our pipeline from files' picking to reports' generation. All these settings resulted in rise of the efficiency of our customized pipeline unobtainable by manual scouting: the analysis of 192 files searched against customized database took 11 hours.

The detail investigation of proteomic data with customized database and information about protein molecular weight and pI allowed us to identify and describe 1302 canonical forms, 426 splice forms, 257 forms with SAPs and approximately 734 proteins with PTM from 29485 protein hits obtained for whole genome. For chromosome 18 we identified 28 of 228 detected proteoforms (16 splice variants, 5 forms with SAPs and 7 PTMs) encoded by 50 genes of this chromosome.

Such multi-omics approach, involving analysis of transcriptome and proteome profiling of the same biologically unique sample, allows focusing on target search of certain proteoform, validated at transcriptomic level. Effective tandem of 2DE and mass spectrometry made it possible to forecast modifications, which can change physical-chemical parameters (and the location of protein spot on the gel, consequently).

The synergy of transcriptomic and proteomic data complemented by bioinformatics as well as whole heterogeneous proteome monitoring provides an opportunity to come closer to understanding of mechanisms of complex biological systems.

This work was supported by RSF grant #15-15-30041.

1. J. Munoz, A.J. Heck (2014) From human genome to human proteome, *Angewandte Chemoe (International edition in English)*, **53**:10864–10866.

2. E.V. Poverennaya, A.T. Kopylov, E.A. Ponomarenko et al. (2016) State of the art of chromosome 18-centric HPP in 2016: transcriptome and proteome profiling of liver tissue and HepG2 cells, *Journal of Proteome Research*, **15**:4030–4038.

3. G.S. Krasnov, A.A. Dmitriev, A.V. Kudryavtseva et al. (2015) PPLine: an automated pipeline for SNP, SAP, and splice variant detection in the context of proteogenomics, *Journal of Proteome Research*, **14**:3729–3737.

4. M. Vaudel, H. Barsnes, F.S. Berven et al. (2011) SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches, *Proteomics*, **11**:996–999.