Exome-based proteogenomics of human cancer cell lines

Ksenia G. Kuznetsova,

Institute of Biomedical Chemistry, 10 Pogodinskaya St., Moscow, 119121, Russia,
kuznetsova.ks@gmail.com

Dmitry S. Karpov,

Institute of Biomedical Chemistry, 10 Pogodinskaya St., Moscow, 119121, Russia,

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, 119991, Russia
aleom@yandex.ru

Mark V. Ivanov

Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, 119334, Moscow, Russia,

Moscow Institute of Physics and Technology (State University), 141707, Dolgoprudny, Moscow region, Russia

markmipt@gmail.com

Irina Y. Ilina

Institute of Biomedical Chemistry, 10 Pogodinskaya St., Moscow, 119121, Russia,
ne-murka@list.ru

Maria A. Karpova

Institute of Biomedical Chemistry, 10 Pogodinskaya St., Moscow, 119121, Russia,
moushutic@mail.ru

Lev I. Levitsky

Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, 119334, Moscow, Russia,

Moscow Institute of Physics and Technology (State University), 141707, Dolgoprudny, Moscow region, Russia
lev.levitsky@phystech.ru

Mikhail V. Gorshkov

Institute for Energy Problems of Chemical Physics, Russian Academy of Sciences, 119334, Moscow, Russia,

Moscow Institute of Physics and Technology (State University), 141707, Dolgoprudny, Moscow region, Russia
mike.gorshkov@gmail.com

Sergei A. Moshkovskii

Institute of Biomedical Chemistry, 10 Pogodinskaya St., Moscow, 119121, Russia,

Pirogov Russian National Research Medical University (RNRMU), 117997, Moscow, Russia
smosh@mail.ru

A degree of genetic variance deciphered by recent efforts in genomics made proteome researchers to revise their approach to database search. Indeed, in most cases shotgun proteomics uses genomic database to identify and quantify proteins. Then, one would use customized genomic database to analyze each individual proteome more precisely. Use of personalized or otherwise customized genome databases to search proteins has formed one of the branches of so called proteogenomics [1]. It is of special interest to search protein variants expressed by cancer genome where encoded changes of amino acid sequence may play driver role in carcinogenesis.

We implemented proteogenomics approach combining publicly available high-throughput exome data [2] and shotgun proteomics analysis [3] for cancer cell lines from NCI-60 panel to demonstrate further that the cell lines can be effectively recognized using identified variant peptides. A database was generated containing mutant protein sequences of NCI-60 panel of cell lines. The proteome data was searched using Mascot and X!Tandem search engines against databases of both reference and variant protein sequences. The identification quality was further controlled by calculating a fraction of variant peptides encoded by own exome sequence for each cell line. We found that up to 92.2% peptides identified by both search engines are encoded by the own exome. The data has illustrated the validity of proteogenomics search methods. Further, we used the identified variant peptides for cell line recognition. Notably, it was shown that, depending of the cell line, we have found that 0.66-1.31% of predicted variant peptides and 5.53-8.11% of predicted wild-type peptides have been identified, respectively, in deep proteomes from [3]. This fact indicates a decreased expression of variant proteins in cancer cell lines.

Using NCI-60 cell line data, the own HEK-293 cell line deep proteome and publicly available proteomes of TCGA colon cancer samples we have shown that one-stage method of false-discovery rate (FDR) calculation yields more variant peptides identified than two-stage FDR method, in the cases where customized genome or RNA database is used for proteomic search.

As it was shown elsewhere [1], a level of false positive peptide identification in proteogenomics using shotgun proteomics data may be increased when chemical native and

artifact modifications of amino acid residues mimic genetically encoded variant. Using own mass-spectra and available data [3] we characterized methionine-to-isothreonine artifact conversion in proteins, the reaction simulating methionine-to-threonine genetically encoded mutation.

Thus, we have implemented a workflow for correct search and characterization of expressed protein variants in cancer samples using customized exome databases.

The work was supported by the Russian Scientific Fund, grant # 14-15-00395 to S.A. Moshkovskii.

References

1. A.I. Nesvizhskii (2014) Proteogenomics: concepts, applications and computational strategies, *Nat. Methods*, **11**: 1114-1125.

2. O.D. Abaan et al (2013) The exomes of the NCI-60 panel: a genomic resource for cancer biology and systems pharmacology, *Cancer Res.,* **2013,** 73: 4372-4382.

3. A. Moghaddas Gholami et al. (2013) Global proteome analysis of the NCI-60 cell line panel, *Cell Rep.***, 4**: 609-620.

4. B. Zhang et al (2014) Proteogenomic characterization of human colon and rectal cancer, *Nature*, 513: 382-387.