



Gene-centric view on the human proteome project: The example of the Russian roadmap for chromosome 18

Alexander Archakov^{1,2}, Alexander Aseev³, Victor Bykov⁴, Anatoly Grigoriev⁵, Vadim Govorun⁶, Vadim Ivanov^{6,7}, Alexander Khlunov⁸, Andrey Lisitsa¹, Sergey Mazurenko⁹, Alexander A. Makarov¹⁰, Elena Ponomarenko¹, Renad Sagdeev¹¹ and Konstantin Skryabin¹²

¹ Orekhovich Institute of Biomedical Chemistry, Russian Academy of Medical Sciences (RAMS), Moscow, Russia

² Russian Proteome Society (RHUPO), Moscow, Russia

³ Institute of Semiconductor Physics, Siberian Branch (SB) of Russian Academy of Sciences (RAS), Novosibirsk, Russia

⁴ Joint-Stock Company "NT-MTD", Zelenograd, Russia

⁵ Institute for Biomedical Problems, RAS, Moscow, Russia

⁶ Research Institute of Physical-Chemical Medicine of the Ministry of Public Health of the Russian Federation, Moscow, Russia

⁷ Shemjakin and Orekhovich Institute of Bioorganic Chemistry, RAS, Moscow, Russia

⁸ Department of Science, High Technologies and Education at the Government of the Russian Federation, Moscow, Russia

⁹ The Ministry of Education and Science of the Russian Federation, Moscow, Russia

¹⁰ Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia

¹¹ International Tomography Center SB, RAS, Moscow, Russia

¹² Center "Bioengineering", RAS, Moscow, Russia

During the 2010 Human Proteome Organization Congress in Sydney, a gene-centric approach emerged as a feasible and tractable scaffold for assemblage of the Human Proteome Project. Bringing the gene-centric principle into practice, a roadmap for the 18th chromosome was drafted, postulating the limited sensitivity of analytical methods, as a serious bottleneck in proteomics. In the context of the sensitivity problem, we refer to the "copy number of protein molecules" as a measurable assessment of protein abundance. The roadmap is focused on the development of technology to attain the low- and ultralow "copied" portion of the proteome. Roadmap merges the genomic, transcriptomic and proteomic levels to identify the majority of 285 proteins from 18th chromosome – master proteins. Master protein is the primary translation of the coding sequence and resembling at least one of the known isoforms, coded by the gene. The executive phase of the roadmap includes the expansion of the study of the master proteins with alternate splicing, single amino acid polymorphisms (SAPs) and post-translational modifications. In implementing the roadmap, Russian scientists are expecting to establish proteomic technologies for integrating MS and atomic force microscopy (AFM). These technologies are anticipated to unlock the value of new biomarkers at a detection limit of 10^{-18} M, i.e. 1 protein copy per 1 μ L of plasma. The roadmap plan is posted at www.proteome.ru/en/roadmap/ and a forum for discussion of the document is supported.

Received: September 8, 2010

Revised: January 27, 2011

Accepted: February 11, 2011

Keywords:


Detection limit / Human Proteome Project / Roadmap / Technology

Correspondence: Alexander Archakov, 119121, Orekhovich Institute of Biomedical Chemistry, RAMS, Pogodinskaya str., 10, Moscow, Russia

E-mail: inst@ibmc.msk.ru

Fax: +74992460857

Abbreviation: HPP, human proteome project



Correspondence concerning this and other Viewpoint articles can be accessed on the journals' home page at: <http://viewpoint.proteomics-journal.de>

Correspondence for posting on these pages is welcome and can also be submitted at this site.

The Human Proteome Project (HPP) was proposed to identify and characterize the proteins encoded by the human genome [1]. This is expected to be a greater challenge than deciphering the genome since the proteome differs depending on the cell type or biological fluids and varies over time [2]. Second, transcriptional, translational and post-translational modifications create a wide diversity of various protein forms originating from a single gene. Finally, due to the absence of a PCR analog for proteins, a technological problem exists in determining low and ultralow protein copy numbers [3].

Proteomics was established as a genome-scale protein science. The gene-centric approach was proposed to narrow the problem of proteome investigation [1, 4]: instead of mining the products of the whole genome, the gene-centric principle is focused on exhaustive information of the pre-selected subset of genes [5], e.g. the genes located on a single chromosome. The gene-centric approach, unlike the currently dominating organ-based and disease-based proteomic approaches, enables the creation of a transparent roadmap for the inventory of protein species.

There is an inability to detect low and ultralow protein copy numbers, which may hamper implementation of the HPP. To define the scope of the HPP, it is necessary to determine a limit to which the proteome can be explored. Relative to this limit, it will be possible to compare the technical progress of proteomic technologies. Decades ago, during the initial days of the Human Genome Project, the competition began in which the number of sequenced nucleotides per unit of time served as a general performance

measure. In a similar way, the HPP could assess its overall progress by the number of identifiable (Intentionally, we are refraining from considering the technical definition of protein identification.) protein copies per volume of biomaterial. Currently, proteomics is in an initial stage where proteins can be detected in concentrations down to the 10^{-14} M range [6]. The lowermost boundary for the ultralow protein copy numbers is 10^{-18} M, shifting four orders of magnitude from the current level. This concentration corresponds to about 1 protein copy per 1 μ L of plasma. As for liver cells, which have a volume of ~ 10 pL, the detection limit of 10^{-18} M enables the identification of 1 protein copy per 10^3 cells.

The general considerations above enable the creation of a chromosome-based roadmap for implementation of the HPP. The following criteria were suggested for selection the chromosome for the Russian part of HPP: (i) minimum of protein-coding genes; (ii) abundance of genes relevant to the diseases according to the available literary data; (iii) lack of immunoglobulin-coding genes. Chromosome 18, selected as the Russian contribution to the HPP, consists of 76 M bases [7] containing 513 genes, of which 285 genes are encoding the potentially expressed proteins (<http://www.ensembl.org>, release 60). This figure exactly corresponds to the UniProt data: it reports 285 proteins, coded on 18th chromosome, and existence of 194 of them is evidenced at protein level (<http://www.uniprot.org>, release 2011_1). According to the Human Protein Atlas (www.proteinatlas.org, v.7.0, [8]), for the 285 proteins encoded on chromosome 18, the antibodies are known only for 134.

By analyzing protein identification in proteomic repositories, we obtained rough estimation of the distribution of proteins from a given chromosome among the abundance categories. Sifting through 187 experiments with human plasma deposited in PRIDE, we observed that 42% of master proteins of 18th chromosome origin were never ever observed in plasma. On the contrary, transthyretin was observed in 57% of plasma MS experiments. Transthyretin is the only protein from the 18th chromosome listed in the

Table 1. Roadmap initial guidelines for classifying proteins according to their abundance

Abundance range	Concentration in blood plasma	18th Chromosome Estimations for Plasma	
		Initial assumptions on the number of master proteins ^{a)}	Comment
High	$\geq 10^{-6}$	1	Transthyretin [9]
Medium	10^{-6} to 10^{-11}	119	Including proteins identified at least in 2 plasma experiments in PRIDE ^{b)}
Low	10^{-11} to 10^{-14}	165	Including proteins identified once in PRIDE
Ultralow	10^{-14} to 10^{-18} ^{c)}	285	Including proteins not detectable in plasma by MS

a) Cumulative values.

b) Concentration (C) was estimated by formula $\log_{10}(N) = 0.365 * \log_{10}(C) - 0.711$ provided in [10], where N is the number of peptides detected for a given protein.

c) Corresponds to the utmost reasonable level of 1 protein copy per 1 μ L of plasma.

hit parade of 150 plasma proteins, spanning the range from 1.5×10^{-5} M to 3×10^{-5} M [9]. Besides this single highly abundant protein, there were 119 others identified at least twice in different PRIDE experiments and 31 out of these – at least in five experiments. Correlating the number of identified unique peptides with their plasma abundance according to [10], we obtained that most of the plasma proteins frequently occurring in PRIDE could be expected to be present at the concentration from 10^{-8} to 10^{-11} M. The rest of the proteins can be preliminary split onto two fractions: those having low number of copies and thus occurring just in a single experiment, and ultralow – with no identifications at all.

Table 1 illustrates how preliminary assumptions on the protein copy number can be mapped onto the chromosome-relevant data. In our view, gene-centric roadmap has to justify its activities basing on the iteratively updated picture of protein distribution among the abundance categories in a given biological specimen. Planning the amount of efforts during the chromosome survey, it is necessary to reference to the concentration ranges, like exemplified by Table 1.

The roadmap proposes to identify 18th chromosome coded proteins that are expressed in liver/HEPG2 cells or are available in plasma. During the pilot phase, which will be implemented in 3 years, only those proteins that are present at the level above 1 protein copy per μ L of plasma or 1 copy per 10^3 of liver/HEPG2 cells will be considered. The executive phase of the roadmap will be completed in 5 years. It will cover experimental revelation of the modifications of previously identified master proteins, including single amino acid polymorphisms (SAPs), products of alternative splicing and PTMs.

It is expected that the roadmap elaboration process will include several tasks that will span both phases of the project. These tasks fall into the following categories:

- (i) genome/transcriptome analysis using the next generation sequencing technologies to perform deep sequencing of putative coding regions of chromosome 18, to elucidate the alternatively spliced transcripts and further trace them in the proteome [11];
- (ii) detection of medium, low and ultralow protein copy numbers and peptides by combining multidimensional separation with multiple reactions monitoring (MRM) technology with a special emphasis on protocols of irreversible binding of proteins on microbeads;
- (iii) proteotyping and proteogenomic profiling [12]; protein affinity capturing [13] to collect the 18th chromosome-centered portion of the full-cell interactome – partner proteins, interacting with “baits”, coded by 18th chromosome might be further used to compile a complete human interactome, as an integrative part of the HPP.
- (iv) development of advanced technologies for identification of ultralow copy proteins using atomic force microscopy (AFM) in combination with MS [14].

Development of analytical technologies is the primary focus of the Russian roadmap activity. In our opinion, the top target in this direction is use of nanotechnological approaches, which are already gaining strength in genomics [15]. If high-speed single molecule detection is applicable for proteins [16], then the problem of concentration sensitivity in proteomics might be resolved [17, 18]. In case these new methods will be applicable for complex biological specimens, the opportunity to analyze cells and biological fluids at the resolution of few protein copies per sample will emerge.

Within the national framework of the HPP, Russia is planning to establish proteomic technologies that integrate MS with AFM (AFM-MS). The method of irreversible binding of proteins to the microbeads will be utilized in combination with multiple reactions monitoring to detect low and ultralow protein copy numbers. With such technologies, one may expect to attain sensitivity at the level of 10^{-18} M in blood plasma (~ 1 protein copy per 1μ L) or, correspondingly, 1 protein molecule per 10^3 of cells. The application of irreversible binding in mining the proteome will make it possible to reveal a principally new group of early biomarkers directly relating to the onset of disease.

Additionally, single amino acid polymorphisms, alternative splicing and chemical post-translation modifications may also serve as a source for biomarker discovery [19–21]. These points are discussed in the Medical addendum to Roadmap, which bridge a gap between the disease-centric and gene-centric views.

The roadmap was prepared under the supervision of the Ministry of Education and Science of the Russian Federation in collaboration with the Department of Science, High Technologies and Education of the Government of the Russian Federation.

The authors have declared no conflict of interest.

References

- [1] Pearson, H., Biologists initiate plan to map human proteome. *Nature* 2008, 452, 920–921.
- [2] Frenkel-Morgenstern, M., Cohen, A. A., Geva-Zatorsky, N., Eden, E. et al., Dynamic Proteomics: a database for dynamics and localizations of endogenous fluorescently-tagged proteins in living human cells. *Nucleic acids Res.* 2010, 38, D508–D512.
- [3] Archakov, A., Ivanov, Y., Lisitsa, A., Zgoda, V., Biospecific irreversible fishing coupled with atomic force microscopy for detection of extremely low-abundant proteins. *Proteomics* 2009, 9, 1326–1343.
- [4] Pearson, H., Biologists initiate plan to map human proteome. *Nature* 2008, 452, 920–921.

- [5] Rabilloud, T., Hochstrasser, D., Simpson, R. J., Is a gene-centric human proteome project the best way for proteomics to serve biology? *Proteomics* 2010, 10, 3067–3072.
- [6] Archakov, A. I., Ivanov, Y. D., Lisitsa, A. V., Zgoda, V. G., AFM fishing nanotechnology is the way to reverse the Avogadro number in proteomics. *Proteomics* 2007, 7, 4–9.
- [7] Nusbaum, C., Zody, M. C., Borowsky, M. L., Kamal, M. et al., DNA sequence and analysis of human chromosome 18. *Nature* 2005, 437, 551–555.
- [8] Pontén, F., Jirström, K., Uhlen, M., The Human Protein Atlas—a tool for pathology. *J. Pathol.* 2008, 216, 387–393.
- [9] Hortin, G. L., Sviridov, D., Anderson, N. L., High-abundance polypeptides of the human plasma proteome comprising the top 4 logs of polypeptide abundance. *Clin. Chem.* 2008, 54, 1608–1616.
- [10] States, D. J., Omenn, G. S., Blackwell, T. W., Fermin, D. et al., Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat. Biotechnol.* 2006, 24, 333–338.
- [11] Menon, R., Omenn, G. S., Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. *Cancer Res.* 2010, 70, 3440–3449.
- [12] Armengaud, J., Proteogenomics and systems biology: quest for the ultimate missing parts. *Expert Rev. Proteomics* 2010, 7, 65–77.
- [13] Buneeva, O., Gnedenko, O., Zgoda, V., Kopylov, A. et al., Isatin-binding proteins of rat and mouse brain: proteomic identification and optical biosensor validation. *Proteomics* 2010, 10, 23–37.
- [14] Kaysheva, A. L., Ivanov, Y. D., Zgoda, V. G., Frantsuzov, P. A. et al., Visualization and identification of hepatitis C viral particles by atomic force microscopy combined with MS/MS analysis. *Biochemistry (Mosc.)* 2010, 4, 15–24.
- [15] Tsutsui, M., Taniguchi, M., Yokota, K., Kawai, T., Identifying single nucleotides by tunnelling current. *Nat. Nanotechnol.* 2010, 5, 286–290.
- [16] Shibata, M., Yamashita, H., Uchihashi, T., Kandori, H., Ando, T., High-speed atomic force microscopy shows dynamic molecular processes in photoactivated bacteriorhodopsin. *Nat. Nanotechnol.* 2010, 5, 208–212.
- [17] Lee, M., Lee, D., Jung, S., Lee, K. et al., Measurements of serum C-reactive protein levels in patients with gastric cancer and quantification using silicon nanowire arrays. *Nanomed. Nanotechnol. Biol. Med.* 2010, 6, 78–83.
- [18] Ivanov, Y. D., Govorun, V. M., Bykov, V. A., Archakov, A. I., Nanotechnologies in proteomics. *Proteomics* 2006, 6, 1399–1414.
- [19] Godai, T. I., Suda, T., Sugano, N., Tsuchida, K. et al., Identification of colorectal cancer patients with tumors carrying the TP53 mutation on the codon 72 proline allele that benefited most from 5-fluorouracil (5-FU) based postoperative chemotherapy. *BMC Cancer* 2009, 9, 420.
- [20] Menon, R., Omenn, G. S., Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers. *Cancer Res.* 2010, 70, 3440–3449.
- [21] Hart, G. W., Copeland, R. J., Glycomics hits the big time. *Cell.* 2010, 143, 672–676.