ISSN 0026-8933, Molecular Biology, 2018, Vol. 52, No. 3, pp. 335–349. © Pleiades Publishing, Inc., 2018.
Original Russian Text © O.I. Kiseleva, A.V. Lisitsa, E.V. Poverennaya, 2018, published in Molekulyarnaya Biologiya, 2018, Vol. 52, No. 3, pp. 394–410.

REVIEWS

UDC 577.1, 57.088

Proteoforms: Methods of Analysis and Clinical Prospects

O. I. Kiseleva[a], *, A. V. Lisitsa[a], and E. V. Poverennaya[a]

[a] *Orekhovich Institute of Biomedical Chemistry, Moscow, 119121 Russia*

**e-mail: olly.kiseleva@gmail.com*

Received July 4, 2017; in final form, September 19, 2017

Abstract—A critical analysis of proteomes provides a basis for understanding the operation of complex biochemical systems. A personalized approach to therapy takes into account biological uniqueness of each patient at genome, transcriptome, and proteome levels, and is a priority area in molecular medicine. The identification of proteoforms, which have dramatic impact on the phenotype of a disease, is a fundamental task of personal molecular profiling. Considerable progress of proteomic approaches presented new avenues for accurate, specific, and high-performance protein analysis. Thus, the identification of new efficient biomarkers can be expected based on studies of aberrant proteoforms associated with various diseases.

Keywords: proteoforms, proteome heterogeneity, human proteome, biomarkers

DOI: 10.1134/S0026893318030068

INTRODUCTION

The completion of the Human Genome Project resulted in some 20000 protein-coding genes, which is approximately 50000 times smaller than the initial estimates [1]. The coding space is rather limited and the complexity of living systems is achieved through the regulation of transcription, translation, and various modifications, which is responsible for the diversity of proteoforms. Due to alternative splicing, a protein-coding gene produces four splice-mediated transcripts on average [2–4]; this is further multiplied by the number of possible single amino acid polymorphisms (SAPs) and post-translational modifications (PTM), which results in a theoretical total number of protein variants or proteoforms encoded by a single gene exceeding a hundred [5].

Today, proteomic studies are focused on protein forms represented in the organism the most, that is, canonical (unmodified) amino acid sequences. The sequence is considered canonical if it is encountered the most frequent among several alternative transcripts, is most similar to orthologous forms of protein, or its amino acid composition or length allow for the most complete description of domains, polymorphisms, etc. (protein amino acid sequence database http://www.uniprot.org/help/canonical_and_isoforms). The concept of canonical protein products encoded by a single gene is mainly used in the context of alternative splicing and implies that one of the translated splice forms is selected as the reference; the rate of SAP is taken into account when the reference sequence is appointed.

Aberrant (noncanonical) proteoforms are products of amino acid sequence editing that differs from that of the canonical proteoform. The concept of a canonical proteoform only makes sense when indicating the organ or tissue under consideration, since different proteoforms that are encoded by a single gene frequently prevail in different types of biomaterials, even under normal conditions [6].

The emergence and progression of pathology in an organ or tissue gives rise to the accumulation of multiple proteoforms that are normally absent or expressed in insignificant amounts in the tissue [7, 8]. Building up with the progression of the disease, these variants affect the regulation of the cell cycle, apoptosis, or DNA reparation either directly or indirectly. The molecular profiling of diseases has become a modern trend in the development of fundamental medicine. Proteomic profiles of malignant tumors of the highest heterogeneity compared to other similar tumors or normal tissues are of particular interest. In tumor tissues, the number of mutant protein variants per gene can exceed that in normal tissue by several orders of magnitude [8].

In 2014, the results of two large-scale projects that present draft human proteome and report the identification of over 90% of protein-coding genes [9, 10] were published; nevertheless, the total number of proteoforms that comprise an organism's proteome or its individual organ or tissue is not reliably known. This lack of data is largely due to the methodological difficulties associated with the experimental confirmation

Abbreviations: SAP, single amino acid polymorphism; PTM, post-translational modification; SNP, single nucleotide polymorphism; ELISA, enzyme-linked immunosorbent assay; SRM, selected reaction monitoring.

of proteoform translation. The difficulties are caused by dynamic nature of the proteome and sensitivity limitations of the existing analytical techniques [11]. Taking into account various scenarios of aberration generation, human proteome may contain hundred thousand to several billion proteoforms [1, 12, 13].

SOURCES OF PROTEOME HETEROGENEITY

Single Amino Acid Polymorphism

Single nucleotide polymorphism (SNP) is a major instrument that generates heterogeneity at the level of genes [14, 15]. The abundance of the aberration is perfectly illustrated by the fact that genomes of two healthy humans differ by 4 million nucleotides on average [15, 16]. Today, over 150 million SNPs (NCBI dbSNP Build 146) have been discovered in the human genome.

Nonsynonymous nucleotide substitutions at the level of genes generate not only start and stop codons, but also SAPs in the protein. Changes in the protein function and scenarios of its interaction with other proteins are caused by changes in stability and/or conformation of the protein product [17, 18].

Biochemical reactions that occur in living systems are sensitive to stereometric parameters of the interacting substances. Even if products of the reaction that involve a modified protein remain the same, the kinetics of the process might change; for example, single amino acid substitutions in the NIS protein that are responsible for the active transport of sodium iodide, lead to impairment of the electrostatic interactions in its transmembrane domains and decrease in the activity of iodide ion transport resulting in inherent hypothyreosis [19]. Substitution of an amino acid residue (for example, a bulky side chain with a compact one) can influence the stability of a protein molecule [20]. For example, among the disease-associated SAPs, only 20% does not change the protein stability [21].

Not every aberration leads to noticeable change in properties of the protein product. Aberrations can be arbitrarily divided into neutral (passenger) and pathology-associated (driver) aberrations. The former are typical of individual populations; they occur spontaneously in somatic cells through their whole life and play no significant role in formation of a pathology phenotype. The latter occur in disease-associated genes and lead to significant changes in the level of protein expression, as well as the activation or suppression of its functional activity. In the case of oncology diseases, mutations play a key role in the malignant transformation and progression of tumors [22].

Pathology-associated mutations and aberrant protein products they encode are of the most interest in the search for prognostic and diagnostic biomarkers, as well as therapeutic targets. However, even in known oncogenes and oncosuppressors, as little as 40% amino acid substitutions are considered meaningful without a single mutation, but a pattern of several substitutions is critical [23].

Single mutations can considerably modify the protein profile [24–26]. Replacement of asparagine with valine at position 6 of the hemoglobin β-subunit provides stability to the malaria parasite due to the difference in its tertiary structure and the properties of the protein it creates [27].

Cases of a significant modification of protein characteristics due to substitutions that do not provide advantages in survival, but on the contrary are associated with the emergence and development of pathology are more frequent. The so-called butterfly effect when a single amino acid substitution induces the transformation of the gene-expression profile of a healthy cell into a profile typical of malignant tumor, was exemplified by the H1047R mutation in the PIK3CA protein, which is responsible for cell growth, proliferation, and survival in healthy mammary gland cell line MCF-10A [28].

Alternative Splicing

The annotation of transcriptomes of various organisms has revealed that almost every multiexon gene is subject to alternative splicing, i.e., the excision of silent introns and the ligation of information-rich exons of protein-encoding genes, which results in a range of unique mRNAs [29].

The contribution of alternative splicing to an increase in proteome heterogeneity is the subject of many discussions [4, 30]. A popular hypothesis defines the role of alternative splicing as a key instrument in the tissue-specific commutation of protein–protein interactions [31, 32]. The opposite hypothesis formulated in a number of works on comparing the levels of heterogeneity of transcriptome and proteome is based on the supposition that alternative splicing is of a stochastic nature and far from all alternative transcripts encode proteins [33]. In addition to the evaluation of alternative splicing contribution to the generation of proteome diversity, the timely importance of these studies is associated with the fact that spliceoforms produced by a single gene can equally rest unnoticed by the organism or turn out to be critical, either vital or fatal. At least 15% of point mutations associated with genetic diseases are harmful due to the impairment of constitutive splicing sites or the inactivation of their regulators [34, 35].

The maintenance of the balance between various proteoforms is principal for the functioning of a healthy cell: e.g., the impairment of the ratio between isoforms 3R and 4R of the τ-protein, which are normally present in equal amounts, is associated with the progression of Alzheimer's disease [36].

The misbalanced expression of aberrant splice variants provides grounds for the progression of oncological diseases at both the initial stages of tumor genera-

tion and subsequent tumor progression [37]. The *TP53* oncosuppressive gene illustrates the oncology-associated alternative splicing. TP53 is known to generate 12 spliceoforms; although they are detected in normal tissues as well, impaired balance between canonical and aberrant forms is stably observed in tumor cells and cell lines, which correlates with decreased survival [38–40].

The importance to take into account alternative splicing when developing a treatment strategy was demonstrated by the example of vemurafenib, an inhibitor of the BRAF protein (V600E), which is responsible for cell proliferation. Clinical studies of vemurafenib efficacy in metastasizing melanoma treatment revealed that approximately 80% patients responded positively to therapy with further habituation and, consequently, a decrease in the drug efficiency [41, 42]. Acquired resistance is explained by the increased expression of the p61 splice variant of the BRAF protein prone to dimerization. The splice variant lacks several exons that surround the domain interacting with G proteins. Remarkably, patients who displayed no BRAF (V600E) variants prone to dimerization retained susceptibility to vemurafenib [43].

Post-Translational Modifications

Another major source of the proteome heterogeneity are PTMs of proteins. At this stage of proteome formation, protein functional diversity is enlarged by covalent binding of functional groups (phospho, glyco, etc.) or other proteins (e.g., ubiquitin). The synthesized protein may be covalently modified at any stage of the cell's life cycle, and any PTM leads to changes in the spatial structure of protein and its properties and functions. PTMs of proteins provide the basis for the control of a number of key events in cell, from signal transduction into the nucleus and regulation of the gene transcription to triggering the degradation of the deregulated protein [44, 45].

Approximately 5% of the proteome is represented by enzymes that produce over 400 types of PTMs [44]. In normal cell, most PTMs are reversible and are used as a signal to trigger or terminate proliferation processes. Irreversible reactions of aberrant PTMs turn off normal regulation processes in a cell. Chemical modifications influence the level of gene expression, the activation of signal pathways, increased proliferation, and a number of other processes; thus, knowledge on modified proteins is used in the diagnostics of the disease [46]. Approximately 10% of clinical biomarkers in blood plasma and serum approved by the Food and Drug Administration (FDA) in the United States make use of PTM analysis.

Ubiquitination, which is a system for recognizing and utilizing damaged proteins, is one of the most studied PTMs [48]. Abnormal ubiquitination leads to the impairment of the mechanisms of cleavage of misfolded proteins, the accumulation of damaged components, and the slowdown of the renewal of intracellular proteins. A loss of the functional integrity of the proteome is considered to be the major factor in the aging and risk of major human pathologies, including oncological, cardiovascular, and neurodegenerative diseases [49, 50]. Mass spectrometry analysis of the level of modification of τ-protein with paired helical filaments isolated from brain tissues of an Alzheimer's patient has revealed 30 phosphorylation sites and three ubiquitinated lysine residues (K254, K311, and K353). The hyperphosphorylated and ubiquitinated protein not subjected to proteasome degradation has been proposed for use as a marker of Alzheimer's disease [51].

Profile of aberrant modifications rising from PTMs can also be used as a biomarker [52]. For example, in various types of malignant tumors accumulation of inactive hyperphosphorylated form of a major negative regulator of cell cycle RB is often observed in cell nucleus in the absence of mutations in the *RB* gene and maintenance of its expression level [53]. Moreover, the level of the protein modification can be used to personify treatment strategy and predict tumor response to treatment. For example, high level of phosphorylation of protein kinase PKB/Akt in glioblastoma patients decreases the efficiency of erlotinib, an inhibitor of receptor tyrosine kinase acting on epidermal growth factor receptors (EGFR) [54].

Other Sources of Proteoforms

RNA editing. In 1986, changes in nucleotides of mRNA sequence that modify the genetically determined transcript have been registered for the first time in mitochondria of a single-cell organism. Enzymatic modification of the information space of transcripts was termed RNA editing [55].

RNA editing includes a number of ribonucleotide PTMs diving rise to a sequence different from that encoded by genes, with modified functions, stability, or coding ability [56]. According to the preliminary evaluation, at least 1.5% of human mRNA is subjected to editing [57]. Classical examples of RNA editing include modifications of cytosine and adenine.

RNA editing in epithelial cells of small intestine in mammals is studied the most. These cells produce a proteoform of apolipoprotein B APOB-48, which is integrated into intestinal lipoproteins. The same gene is expressed in human liver cells, but in hepatocytes its protein product APOB100, governing the capture and degradation of low-density lipoproteins, is twice longer than the intestinal variant ApoB-48. The reason for this difference is the C → U substitution replacing the glutamine codon CAA with a stop codon UAA. The modification plays an important role in the metabolism of lipoproteins and determines the uniqueness of mechanisms of fat transport in the intestines and liver [58].

Hydrolytic deamination of adenosine to inosin (A → I editing) catalyzed by the ADAR family of enzymes is widely spread. A → I editing is mostly present in cells of the central nervous system; the diversity and frequency of RNA editing in these cells can be explained by the need for the diversity of proteoforms needed for the flexibility of system functioning [59].

The impairment of A → I editing is also associated with oncological pathologies [60, 61]. One of the reasons for acute myeloid leukemia is the impairment of PTPN6 oncosuppressive functions through its mRNA editing. The study of the biomaterial obtained from leukemia patients revealed that the decrease in the rate of RNA editing correlates with an increased duration of the remission period in the patients [62].

Chimerism. Hybrid (chimeric) proteins are proteins encoded by several fused genes. Many of these genes that rise from complicated chromosomal rearrangements are associated with various pathologies [63]. One of the first chromosomal rearrangements detected in human tumor cells was the Philadelphian chromosome typical of myeloid leukemia [64]. The defective chromosome originates from a translocation, which merges the *ABL1* gene of chromosome 9 with *BCR* gene of chromosome 22. Chimeric protein encoded by the *BCR-ABL* gene is an oncogene; its hyperexpression activates pathways of intracellular signal transduction and causes excessive cell proliferation. *BCR-ABL* is viewed as a marker of myeloid leukemiaA and a promising target for therapy [65].

A hybrid of an androgen-regulated gene *TMPRSS2* and a member of an oncogene transcription factor family *ERG* often causes prostate cancer. For example, Demichelis et al. identified hyperexpression of the *TMPRSS2:ERG* gene in 50% of prostate cancer patients and linked it with more aggressive phenotype that leads to a fatal outcome [66].

Also, chimerism is interesting from the point of view of drug development; chimeric biomolecules combine functional properties of effector protein increasing recognition, binding, and toxicity, with the stability and specificity of a transporter protein [67].

Impaired protein processing. Maturation of protein implies transformation of the polypeptide into an active biomolecule with correct spatial structure. Aberrations in polypeptides can affect final tertiary structure of the protein molecule. Incorrect folding can lead to decreased stability and/or changes in functional properties of final protein product. For example, some types of cataract are associated with the R116C mutation in the α-crystallin A protein. This mutation changes the protein conformation and decreases its activity as a chaperon fourfold compared to the wild-type protein [68].

Processing errors that are not corrected in the endoplasmic reticulum are associated with various disorders, ranging from Alzheimer's disease to diabetes [69–71]. For example, SAP in the enzymes that catalyze protein cleavage in lysosomes decrease lysosome activity, which lead to lysosomal storage diseases, which are severe genetic disorders characterized by accumulation of intermediate proteoforms, which are normally degraded, inside the cell [72].

Sources of protein heterogeneity are schematically presented in Fig. 1.

PROTEOMIC APPROACHES TO PROTEOFORM IDENTIFICATION

Studies of DNA and RNA allowed to detect multiple genes involved in tumor processes; however, only two genome sets, Oncotype DX [73] and MammaPrint [74], have been introduced into clinical practice due to the high cost. These sets are used to determine activity of a group of breast cancer genes and predict relapses. Considerable disadvantage of genomic and transcriptomic methods is rooted in the stochastic nature of the proteome: only 10% of mutations observed in cancer cell lines at genomic and transcriptomic levels can be detected in the proteome; this is due to the low sensitivity of current techniques, as well as the heterogeneity of tumors and diverse oncogenesis mechanisms [75].

The proteome is a result of realization of hereditary information processes; therefore, it provides data on protein properties (structure, localization, expression level, modifications, and protein–protein interactions) that are actually occurring, not the plausible ones. PTMs play an important role in the regulation of protein activity, but they cannot be predicted at either the transcriptome or genome levels [76]; therefore, the proteome-based approaches can fully depict the disease profile at a molecular level.

Enzyme-Linked Immunosorbent Assay

Enzyme-linked immunosorbent assay (ELISA) is one of the methods based on mapping protein aberrations, which is also used to identify the pathophysiological conditions [77]. ELISA is based on two principles, i.e., the first principle is that immobilized proteins, including antibodies, are able to maintain their functional activity, while the second principle is that interactions underlying formation of the antibody–enzyme complex are highly specific (Fig. 2).

Modern diagnostic ELISA systems enable the detection of diseases both determined by a single factor (e.g., the E6 oncoprotein of the human papilloma virus) and a multifactorial nature (oncological, autoimmune, and neurodegenerative conditions), which requires the analysis of several molecular markers [78]. The application of multiplex panels allows for a shorter analysis time and lower cost, also saving biomaterial. Progress in the multiplex ELISA widens the profile of monitored markers, which allows distinguishing of diseases with similar phenotypes [79].

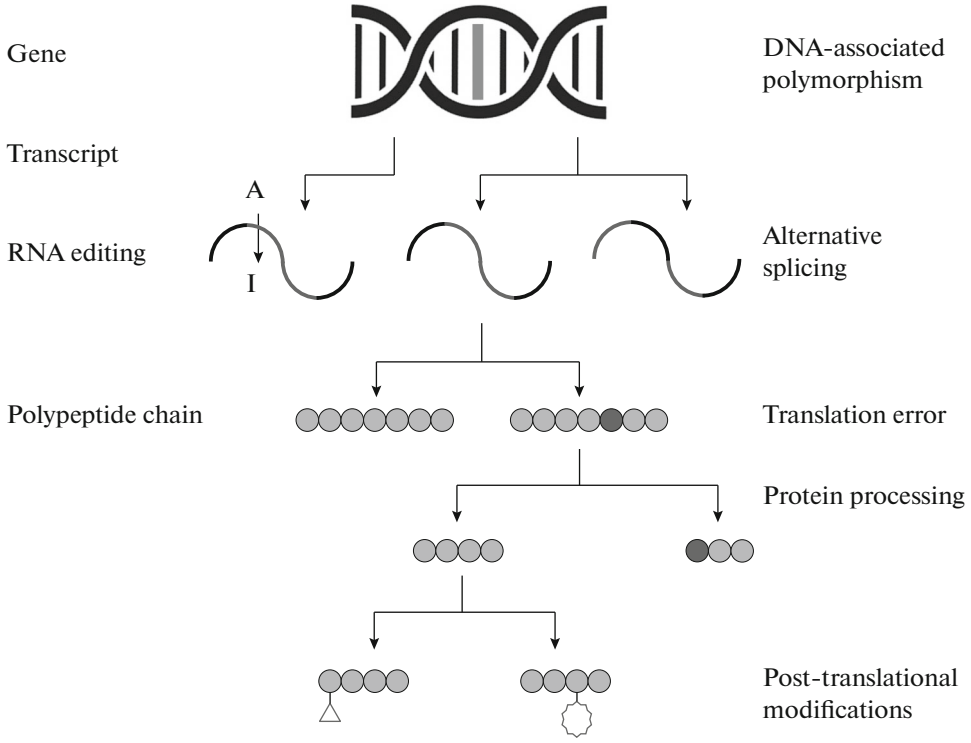

Fig. 1. Sources of proteome heterogeneity.

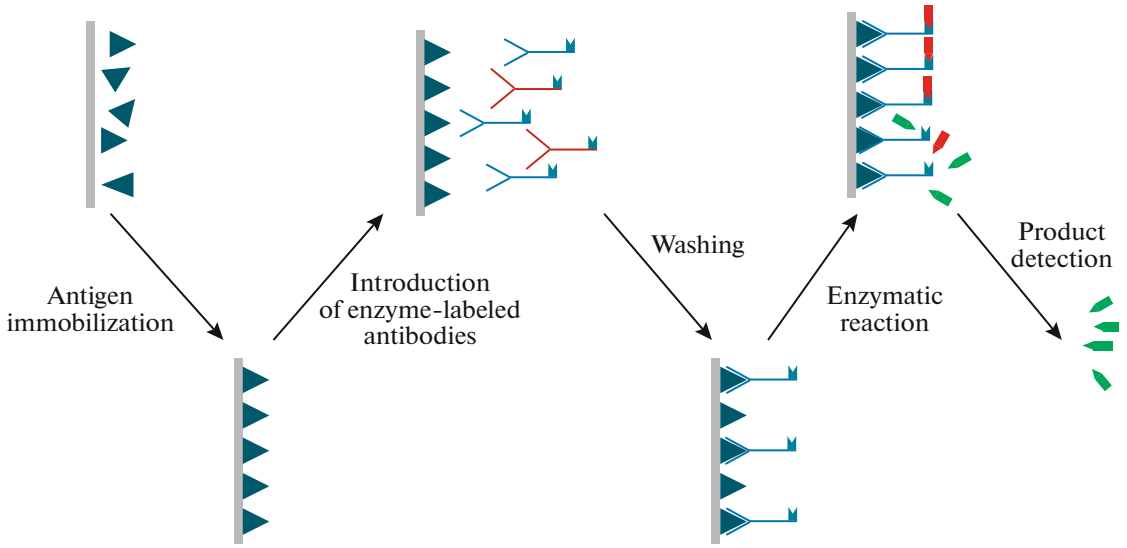

Fig. 2. General scheme of ELISA.

The expansion of the catalogue of antibodies specific to certain regions of the protein molecule enabled the development of new approaches in detection of aberrant proteoforms, including both ELISA and ELISA-PCR combination [80, 81]. For example, ELISA was used to detect the K196E substitution in the S-protein, a genetic risk factor of thromboembolism [82]. The application of specific antibodies enabled the detection of splice-mediated protein variants encoded by the *TNFR2* gene, which recognizes and binds the tumor necrosis factor [83]. Major success was achieved in solving the problem of PTMs of a number of signaling proteins, which could not be performed by transcriptomic methods [84]. ELISA results are used in the development of personalized strategies to treat diseases and control patients'

response to therapy. For example, determination of phosphorylation status in protein products of genes *EGFR* and *MEK* in colorectal cancer cell line provides data on the progression of the oncological disease in the patient's organism and allows treatment to be personalized [85].

Despite that ELISA is considered the golden standard for protein detection, this method has a number of faults associated with hybridization specificity, which typically manifests when proteins of a single family or aberrant forms are studied. A proteoform of the hepatitis B surface antigen (HBsAg) with the T118M substitution in its hydrophilic surface-exposed region has been detected; it allows the virus to evade host immune response. Now, this proteoform of HBsAg is included in ELISA kits designed for hepatitis B virus diagnostics. Earlier, up to 80% of the test systems used to analyze donor blood did not allow the detection of the HBsAg proteoform and showed false-negative results for hepatitis B patients [86].

Recently, inefficiency of a popular prostate cancer biomarker, the prostate-specific antigen (PSA), which yields large fractions of false-negative and false-positive results has been revealed. In studies of prostate gland bioptates, oncological diagnosis has been confirmed in as few as 25% of men with elevated PSA levels and, on the contrary, malignant transformations could not be excluded, even if the PSA level was within normal range [87, 88]. The reason for the poor performance is that the test used for PSA detection was suitable for detecting a canonical form of the protein, while there are over 20 proteoforms thereof [88]. The example confirms the idea that even well-proven laboratory-scale diagnostic tests often underestimate heterogeneity of the proteome; therefore, when applied in clinics, they produce results that are not always true.

The application of ELISA is also complicated by the fact that it is often difficult to select antibodies that would selectively interact precisely with the aberrant and not the canonical proteoform [89]; the search for potential high-affinity agents takes much time and resources. Application of the method is also limited by the constant of antigen–antibody complex formation; proteins with concentrations below 10^{-12} M cannot be detected [90].

Two-Dimensional Gel Electrophoresis

Two-dimensional gel electrophoresis (2DE) is used to separate complex protein mixtures by two independent physicochemical parameters, i.e., the isoelectric point (p*I*) of proteins and their molecular masses. Isoelectric focusing is achieved by passing current through the electrolyte solution in which the voltage gradient is formed. Under the effect of the electric field, proteins migrate according to their charge to the region of electrolyte, where pH matches the protein p*I*. The second axis of protein separation orthogonal to the first one is the molecular mass. Separation by molecular masses is performed by electrophoresis in polyacrylamide gel and is based on the differences in protein migration rate under the effect of an electrical field [91] (Fig. 3).

2DE technique holds one of the leading positions among methods of protein fractionation. Combined with approaches to enrich, specifically stain, and further identify proteins (e.g., by mass spectrometry), the method allows fractioning protein mixtures and differentiating proteoforms at structural level, according to their physicochemical properties [92–96].

The detection of proteoforms with 2DE can be performed in non-covalent and covalent modes [97]. The difference gel electrophoresis (DIGE) approach based on covalent binding of fluorescent probes differing by excitation and emission wavelengths, enables the reproducible quantitative analysis of the protein composition of several samples in a single-gel format [98–100].

Possibilities of 2DE in tandem with shotgun mass spectrometry for identifying proteoforms were demonstrated in both studies that are focused on the identification of certain aberrations (e.g., glycoforms [94], SAPs in protein products of several genes [101], or splice-mediated variants [102]) and large-scale screening projects [95, 96, 103].

Disadvantages of the 2DE method, which in most proteomic experiments is followed by mass spectrometry analysis, include low reproducibility and narrow range of detected protein masses. The method does not allow for the separation of low-copy proteins or proteins with extreme physicochemical characteristics, i.e., acidic or basic, hydrophobic, or too bulky or compact. Despite the above-mentioned limitations, 2DE remains a powerful tool for protein inventory and is successfully combined with other proteomic technologies.

Mass Spectrometry

Today, high-throughput strategies of chromatography–mass spectrometry analysis adapted for certain research task hold a central place among methods of proteoform studies. The classification of mass spectrometry-based proteomic approaches according to the aim of the study includes two strategies, i.e., a shotgun approach suitable for screening studies and a targeted approach for analyzing a certain proteoform.

Shotgun mass spectrometry. The prevailing fraction of proteomic data has been acquired by methods of shotgun proteomics [104]. The problem facing the shotgun proteomics is identification of the maximum number of proteins to fulfil a wide range of scientific tasks, from mapping of major proteins in cell lysates [105] to searching for sensitive and selective biomarkers of diseases [106].

Mass spectrometry is one of the most efficient methods to study protein mixtures, however even the

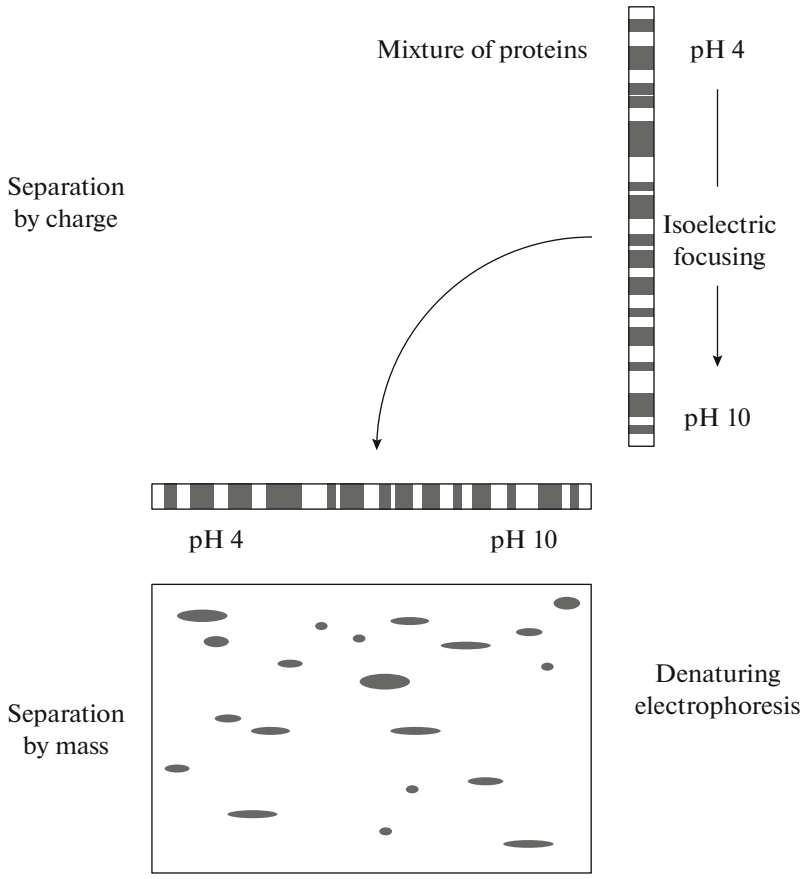

Fig. 3. Principal scheme of two-dimensional gel electrophoresis.

most up-to-date mass spectrometers are limited by the dynamic range of peptides that can be registered during a single scan. In a typical proteomic study mass spectrometry is preceded by chromatography fractionation, which allows to separate peptides in a system of two contacting nonmixing phases, one of which is typically mobile and moves with respect to another phase, i.e., the stationary phase. Chromatography is used to separate mixtures of macromolecules that enter mass spectrometer over time; each macromolecule has individual physicochemical properties and, thus, interacts with the sorbent surface with a varying strength. Chromatography allows to concentrate the analyte within a narrow zone (peak), which leads to an increase in sensitivity of further mass spectral analysis.

To fractionate complex protein mixtures, liquid chromatography is widely used; the variants include reverse-phase, ion-exchange, affinity, and exclusion chromatography, as well as others [107]. Adding the liquid chromatography stage preceding mass spectral analysis decreases the effect of interference on the result of mass spectrometry proteomic data obtained for complex biological mixtures.

After the chromatography stage in proteomic chromatography–mass spectrometry, an ionized and fragmented peptide or protein is characterized by mass spectra that reflect the stages of ion filtration by the mass-to-charge ratio. The key step in shotgun analyses is the identification of fragmented peptides in the mass spectra. In contrast to genomic and transcriptomic approaches, in which fragments of nucleotide sequences are read letter-wise, in proteomic mass spectrometry peptides are traditionally identified by matching experimental mass spectra with theoretical spectra of all possible peptides contained in a database.

Two complementary most popular strategies to mass spectrometry analysis are bottom-up and top-down approaches, which share common chromatography and mass spectrometry stages of the analysis, but differ by the sample preparation stage. In bottom-up approach, proteins are subject to proteolysis, after which the mixture of peptides is separated on a chromatography column and then analyzed in a mass spectrometer. In the top-down approach, full proteins are analyzed in a mass spectrometer without preliminary cleavage.

Bottom-up. In the last two decades, bottom-up mass spectrometry approach, when proteins are cleaved into 6–30 amino acid-long peptides, has been one of the primary methods to analyze amino acid

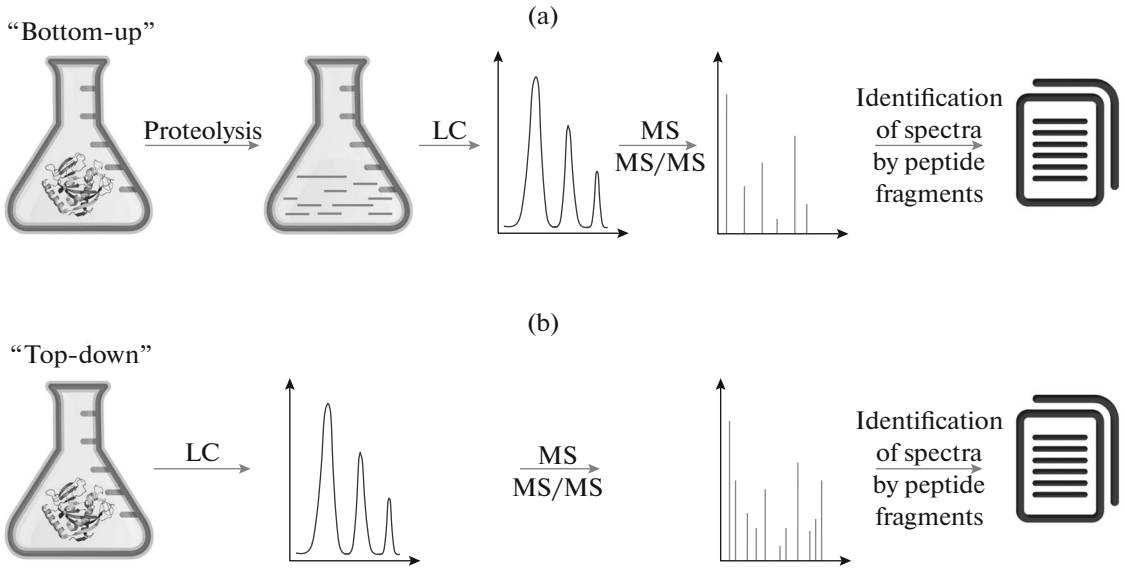

Fig. 4. Principles of shotgun mass spectrometry with preliminary chromatographic fractionation according to bottom-up (a) and top-down (b) approaches. In contrast to top-down strategy, in the bottom-up approach, liquid chromatography (LC) fractionation and the mass spectrometry (MS or MS/MS) experiment follow the stage of proteolytic protein cleavage.

sequence [108] (Fig. 4a). Simplicity of peptide separation and possibility to predict fragmentation patterns enabled the technology to be used for a high throughput analysis of complex protein mixtures and whole proteomes.

There is a number of experiments reported on aberrant proteoforms studied by a bottom-up approach aimed at both the investigation of a specific mutation (for example, substitution determining the human hemoglobin type) and an inventory of all variants of target biomaterial [109, 110]. Large-scale bottom-up studies of cell lines or human tissues (including those employing RNA sequencing-derived databases [95]) enable dozens of splice-variants, hundreds of SAPs, and thousands of PTMs to be detected, some of which are found to be associated with diseases [75, 111, 112]. For example, in several cell lines, it was possible to detect the Y186R substitution in the BCAT2 protein, which is associated with glioma progression [113].

The bottom-up approach is very popular, although it is not free from considerable disadvantages. When PTMs are analyzed, the technology does not allow to determine whether researcher deals with a protein modified twice at two sites or there are two populations of the protein with a single modification each. Use of bottom-up approach in the analysis of splice variants is also limited, since the fragment of amino acid sequence differing from the canonical proteoform may be too short or contain no proteolysis sites [114]. Up to 95% non-mapped mass spectra are referred to uninterpreted zone of amino acid sequence of the proteoform [115]; the fact can partially be explained by reasons common for data-dependent mass spectrometry approaches: insufficient quality of analyzed mass spectra [116] and imperfect search algorithms [117], as well as the incompleteness of reference amino acid sequence libraries, which do not include all possible modifications [118]. The identification of modified peptides is a difficult task. If a researcher does not indicate a specific modification in the search settings, the peptide with the modification will naturally not be identified, while listing of all possible modifications in the search might lead to a combinatorial explosion [119]. Spectra generated based on the results of modified peptide analysis are interpreted erroneously in 20–50% cases [116]. These erroneous identifications are typically characterized by a higher rate of intensity and reliability score than false-positive hits.

Top-down mass spectrometry allows to determine the amino acid sequences of protein ions and study PTMs without preliminary proteolysis (Fig. 4b). For example, the top-down analysis of HeLa cells enabled the identification of over 3000 proteoforms encoded by 1043 genes [120]. Generally, mass spectrometry of complete proteins has a rather narrow area of application and is aimed at a thorough analysis of individual short peptides or simple mixtures with effectively low throughput, although in some cases it can be used for analyses of complex biological samples [121, 122].

The classical example of identifying amino acid substitutions in hemoglobin was not ignored by modern top-down mass spectrometry. Coelho Graca et al. [123] used the top-down approach to register single amino acid substitutions in hemoglobin. The experiment was performed under the conditions of direct injection of high concentrations of hemoglobin extracted from red blood cells; four SAPs were detected virtually.

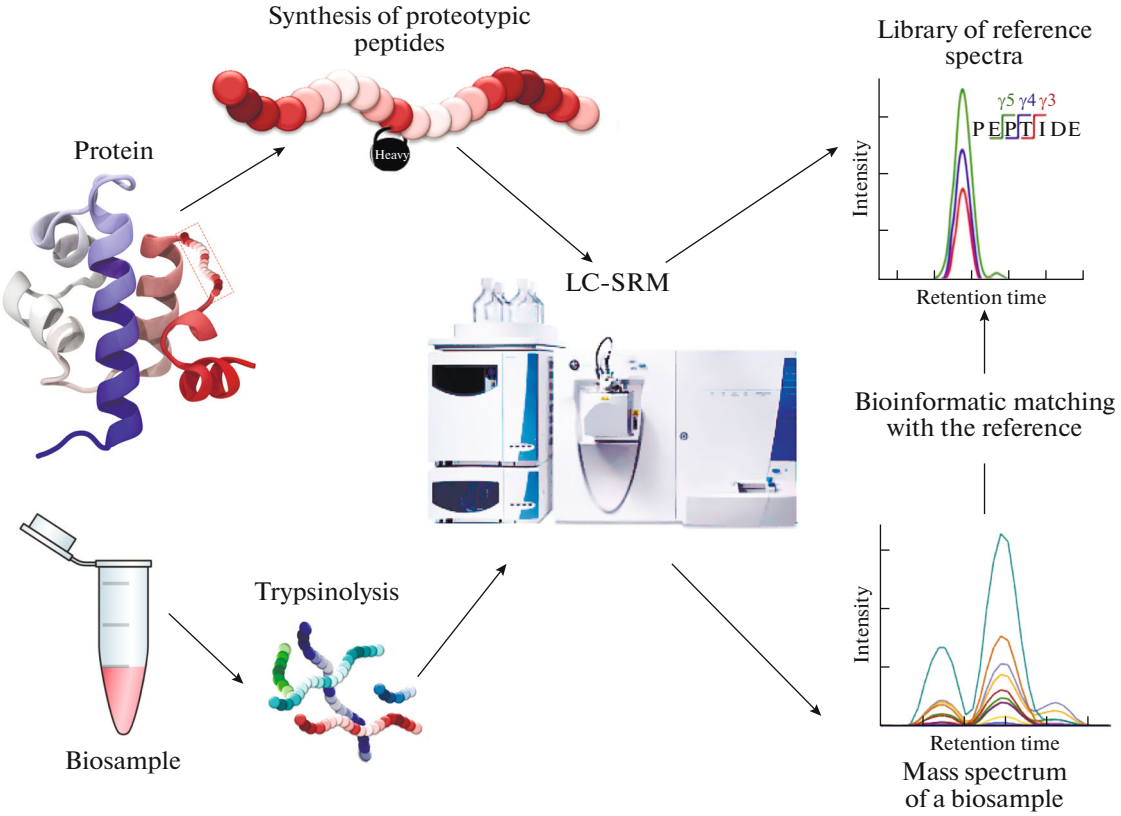

Fig. 5. Principles of targeted mass spectrometry.

Top-down approach allows to perform qualitative and quantitative analyses of various splice variants translated from a single gene, which has been demonstrated based on the example of two splice variants of tropomyosin associated with muscle diseases [124].

To correctly identify protein products of two or several related genes, which often have similar physicochemical properties, high-precision equipment and special software algorithms are required. For example, the trimethylation of lysine changes protein mass by 42.0470 Da, and acetylation changes by 42.0106; the difference between the two modifications is 0.0364 Da, which is not always reliably registered by top-down mass spectrometry. Moreover, in the absence of preliminary hydrolysis, only relatively short proteins (up to 50 kDa) can be studied. Along with limited sensitivity and sophisticated equipment, this limits applications of the top-down approach to studies of modifications, but not as strongly as with the bottom-up strategy. The identification of phosphorylation sites in glutathione-S-transferase P1 that are involved in metabolism of cancerogenic compounds and anticancer agents can be used in an example of a PTM study [125].

Despite significant progress in analyzing whole protein molecules, the sensitivity of the top-down approach is inferior to that of the peptide-based bottom-up method; however, all limitations being considered, the top-down approach remains indispensible in studies of short aberrant proteoforms.

Directed mass spectrometry is aimed at the quantitative determination of certain proteoforms with a known amino acid sequence. The main tool in the approach is the selected reaction monitoring (SRM) method. SRM relies on the mass-spectrometric registration of unique peptides that unambiguously characterize a protein. These peptides occur in the only protein among all proteins encoded by the genome and can be used as reproducibly detected protein equivalents in enzymatic hydrolysis (Fig. 5).

A double m/z (mass-to-charge ratio) filter of the proteotypic precursor peptide and ion fragments ensures the high selectivity of the method. Monitoring pre-selected transitions leads to an increase in sensitivity by two orders of magnitude compared to the shotgun approach of panoramic sequencing [127]. Directed chromatography–mass spectrometry gradually takes on the features of a universal tool of reliable and precise protein analysis for a wide range of medical and biological tasks starting from candidate biomarkers of pathological processes [128] to large-scale projects on the inventory of all proteins in a proteome and an investigation of their heterogeneity [9, 10, 129].

Several works have recognized the applicability of the SRM method in clinics for a reliable quantitative

analysis of medium and high-copied blood plasma proteins [130, 131].

Directed mass spectrometry can be efficiently used to detect SAPs associated with various diseases, including diabetes, obesity, and cancer [132, 133]. The study of a representative cohort of type II [133] diabetes patients revealed an imbalance between the number of aberrant and canonical proteoforms of the *C7* and *CFH* genes detected in homozygous and heterozygous samples. Diagnostically important relations found between levels of allelic gene expression enable reliable differentiation groups of obesity and diabetes patients. In another work, SRM was used to reveal the oncology-associated mutation of the RAS protein involved in various tumor processes [132]. According to the results of the experiments with tumor cell lines, spleen, and tumor tissues, the G12V substitution was detected in a proteotypic peptide and found to be responsible for the malignant transformation.

The efficiency of directed mass spectrometry was demonstrated in the course of identifying of various types of hemoglobin. Uncleaved proteins were subjected to a combination of SRM and electron transfer dissociation (ETD); the specificity of the approach was demonstrated by the example of hemoglobin A blood, into which a sample of hemoglobin C blood was introduced [123].

Wu et al. [134] used SRM to study modifications of osteopontin. They used immunoaffinity enrichment coupled with directed mass spectrometry to analyze splice-mediated variants of the protein. A quantitative analysis of levels of three forms of osteopontin, OPNa, OPNb, and OPNc, was performed; a total of 20 samples was used to demonstrate that isoform A is specific to lung cancer, which contains approximately five times more isoform A as normal cells. Isoform A is detected by SRM in both oncology patients and healthy volunteers; therefore, the results of quantitative analysis of this proteoform are of no interest for clinical studies, but are successfully used to demonstrate the potential of SRM in recognition of spliceoforms.

In a number of works [135–137], SRM was used to reveal and determine PTMs. For example, PNGase F glycosylation status was determined in human blood plasma. The sensitivity of the analysis was 10^{-10} M, which competes with that of ELISA methods [137].

Directed mass spectrometry was applied to the detection of five proteotypic peptides of a chimeric NPM-ALK protein in nine cell lines of lymphoproliferative disorders. The elaborated SRM method employing double isotope labeling was validated on 11 lymphopositive and 12 lymphonegative bioptates; sensitivity was 0.4 fmol/μL. The status of the chimeric protein-specific peptides was determined correctly in all studied clinical samples [138].

Despite the high sensitivity, selectivity, and flexibility of directed mass spectrometry, there are factors limiting its application in aberrant form investigation. Similar to the process of selection and synthesis of antibodies for ELISA, SRM development is limited by the stage of protein-specific peptide synthesis; the latter is used as internal reference. Similar to antibody in ELISA, peptide can fail, i.e., be poorly ionized and detected irreproducibly.

One of the key problems of both SRM and proteomic mass spectrometry as a whole is the selection of optimal size of the reference database. Proteotypic peptide for given proteoform is identified by matching experimental and theoretical spectra contained in different databases. These bases are incomplete in terms of information on possible modifications. The application of these unrepresentative reference libraries naturally does not allow registration of mutations in the directed mass spectrometry mode that were not annotated earlier, since they are absent from databases. As noted previously [139], expansion of library with not validated modifications cannot be considered a silver bullet solution because it increases the fraction of false-positive identifications and will devalue the analysis quality [140].

Interference rising from overlapping of spectral peaks of peptides from a complex biological matrix leads to errors in SRM spectra identification. It was demonstrated that the expansion of the reference library with the inclusion of possible PTMs exponentially increases the number of interfering peptides [141].

Difficulties in detection of aberrant proteoforms are also associated with their low occurrence compared to canonical forms. Moreover, the choice of proteotypic peptides is a complicated task due to high level of homology of amino acid sequences and limited number of proteotypic peptides that can differentiate alternative splice variants. Computer modeling of proteotypic peptides for alternative splicing and their comparison with UniProt data showed that only some 6% possible proteotypic peptides can be used as characteristic peptides of splice variants [87].

CONCLUSION

Modified proteoforms that have altered properties compared to a canonical variant can be used as biomarkers. However, major efforts that have been applied in the search for protein biomarkers by proteomic approaches in the past decade concentrated not on the investigation of aberrant variants, but on the detection of characteristic canonical proteoforms of clinical importance. Unfortunately, so far there have been found no canonical proteoforms with diagnostic characteristics allowing to use them in clinical practice [142].

Taking into account recent technical progress, the inefficiency of proteomics as a provider of new clinically important biomarkers continues to be justified by the imperfection of the equipment and its sensitivity,

reproducibility, and dynamic concentration range limitations. Nevertheless, the mechanical buildup of the number of detected proteins will hardly ever turn to surpassing the barrier on the way to new clinically important molecules. Despite the active development of technology, equipment, and bioinformatics basics [143], each of the existing approaches used to study aberrations at the proteomic level is not free from faults.

High throughput technologies of nucleotide or amino acid sequencing flood repositories with data on impaired constitutive splicing, SAPs, or PTMs, which are potentially associated with various diseases; the importance of each of the mutations detected in the development of the disease is doubtful. The absence of criteria for distinguishing harmful mutations is further aggravated by intricate molecular pathways and the complex multifactorial nature of many diseases, in particular oncological diseases. Terabytes of data on aberrant sequences are explained by the nature of oncological disease. In the course of spontaneous mutagenesis, the uniqueness of each tumor is created under the effect of individual combination of factors; therefore, the extrapolation of singular artifacts on a population is not free from adventurism.

Modern methods allow generation of enormous data sets on mutations in a phenomenological manner, that is, the research is aimed at data accumulation without their subjective interpretation. A huge number of "bad" poorly annotated mutations detected using "good" technologies represents jungles on the way to truly important markers that can be used in clinical practice. To identify a pathology-associated mutation, a representative set of clinical samples and bioinformatic evaluation of its contribution to origin and progression of the disease are needed.

We suppose that synergy of a large amount of transcriptome and proteome data, as well as the monitoring of not a single proteoform, but rather a panel of protein modifications, will allow the discovery of efficient disease biomarkers and elaborate relevant therapeutic strategies; this is supported by bioinformatics approaches. A shift in the emphasis towards mass spectrometry-based research methods will considerably simplify the process of a biomarker search and validation and lower the ultimate assay cost. In the future, the development of mass spectrometry-based assays will allow to personify the evaluation of potential risks of disease development, diagnose diseases, and monitor the efficiency of treatment based on rational assignment of agents and the selection of individual treatment schemes.

This work was financially supported by Russian Science Foundation (grant no. 15-15-30041).

REFERNCES

1. Kelleher N.L. 2012. A cell-based approach to the human proteome project. *J. Am. Soc. Mass Spectrom.* **23**, 1617–1624.
2. Ezkurdia I., del Pozo A., Frankish A., et al. 2012. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol. Biol. Evol.* **29**, 2265–2283.
3. Smith C.W., Valcaŕcel J. 2000. Alternative pre-mRNA splicing: The logic of combinatorial control. *Trends Biochem. Sci.* **25**, 381–388.
4. Ezkurdia I., Rodriguez J.M., Carillo-de Santa Pau E., Valencia A., Tress M.L. 2015. Most highly expressed protein-coding genes have a single dominant isoform. *J. Proteome Res.* **14**, 1880–1887.
5. Mucoz J., Heck A.J.R. 2014. From the human genome to the human proteome. *Angew. Chem. Int. Ed. Engl.* **53**, 10864–10866.
6. Xu Q., Modrek B., Lee C. 2002. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.* **30**, 3754–3766.
7. Hanahan D., Weinberg R.A. 2011. Hallmarks of cancer: The next generation. *Cell.* **144**, 646–674.
8. Martincorena I., Campbell P.J. 2015. Somatic mutation in cancer and normal cells. *Science*. **349**, 1483–1489.
9. Kim M.S., Pinto S.M., Getnet D., et al. 2014. A draft map of the human proteome. *Nature.* **509**, 575–581.
10. Wilhelm M., Schlegl J., Hahne H., et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature.* **509**, 582–587.
11. Archakov A., Zgoda V., Kopylov A., et al. 2012. Chromosome-centric approach to overcoming bottlenecks in the Human Proteome Project. *Expert Rev. Proteomics*. **9**, 667–676.
12. Lisitsa A.V., Poverennaya E.V., Ponomarenko E.A., Archakov A.I. 2015. The width of the human plasma proteome compared with a cancer cell line and bacteria. *J. Biomol. Res. Ther.* **4**, 10–13.
13. Ponomarenko E.A., Poverennaya E.V., Ilgisonis E.V., et al. 2016. The size of the human proteome: the width and depth. *Int. J. Anal. Chem.* **2016**, ID 7436849. doi 10.1155/2016/7436849
14. Cargill M., Altshuler D., Ireland J., et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238.
15. 1000 Genomes Project Consortium, Abecasis G.R., Altshuler, D., Auton A., Brooks L.D., et al. 2010. A map of human genome variation from population-scale sequencing. *Nature.* **467**, 1061–1073.
16. Roach J.C., Glusman G., Smit A.F.A., et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science.* **328**, 636–639.
17. Zhang Z., Miteva M.A., Wang L., Alexov E. 2012. Analyzing effects of naturally occurring missense mutations. *Comput. Math. Methods Med.* **2012**, 805827.
18. Gautieri A., Vesentini S., Redaelli A., Buehler M.J. 2012. Osteogenesis imperfecta mutations lead to local tropocollagen unfolding and disruption of H-bond network. *RSC Adv.* **2**, 3890.
19. Fujiwara H., Tatsumi K., Tanaka S., et al. 2000. A novel V59E missense mutation in the sodium iodide

symporter gene in a family with iodide transport defect iodide. *Thyroid*. **10**, 471–474.

20. Shirley B.A., Stanssens P., Hahn U., Pace C.N. 1992. Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. *Biochemistry*. **31** (3), 725–732.
21. Wang Z., Moult J. 2001. SNPs, protein structure, and disease. *Hum. Mutat.* **17**, 263–270.
22. Vogelstein B., Papadopoulos N., Velculescu V.E., et al. 2013. Cancer genome landscapes. *Science*. **339**, 1546–1558.
23. Abhishek N., Vihinen M. 2015. Harmful somatic amino acid substitutions affect key pathways in cancers. *BMC Med. Genomics*. **8**, 1–12.
24. Ng P.C., Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61–80.
25. Stenson P.D., Ball E., Howells K., et al. 2008. Human Gene Mutation Database: Towards a comprehensive central mutation database. *J. Med. Genet.* **45**, 122–124.
26. Maglott D., Amberger J.S., Hamosh A. 2002. Online Mendelian Inheritance in Man (OMIM): A directory of human genes and genetic disorders. *Nucleic Acids Res.* **30**, 1–7.
27. Kutlar A. 2007. Sickle cell disease: A multigenic perspective of a single gene disorder. *Hemoglobin*. **31**, 209–224.
28. Hart J.R., Zhang Y., Liao L., et al. 2014. The butterfly effect in cancer: A single base mutation can remodel the cell. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 1131–1136.
29. Wang E.T., Sandberg R., Luo S., et al. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature*. **456**, 470–476.
30. Tress M.L., Abascal F., Valencia A. 2017. Alternative splicing may not be the key to proteome complexity. *Trends Biochem Sci*. **42**, 98–110.
31. Buljan M., Chalancon G., Eustermann S., et al. 2012. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol. Cell*. **46**, 871–883.
32. Ellis J.D., Barrios-Rodiles M., Colak R., et al. 2012. Tissue-specific alternative splicing remodels protein–protein interaction networks. *Mol. Cell*. **46**, 884–892.
33. Melamud E., Moult J. 2009. Stochastic noise in splicing machinery. *Nucleic Acids Res.* **37**, 4873–4886.
34. Faustino N.A., Cooper T., Andre N. 2003. Pre-mRNA splicing and human disease. *Genes Dev.* **17**, 419–437.
35. Krawczak M., Reiss J., Cooper D. 1992. The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: Causes and consequences. *Hum. Genet.* **90**, 41–54.
36. Lacovich V., Espindola S.L., Alloatti M., et al. 2017. Tau isoforms imbalance impairs the axonal transport of the amyloid precursor protein in human neurons. *J. Neurosci.* **37**, 58–69.
37. Bonomi S., Gallo S., Catillo M., et al. 2013. Oncogenic alternative splicing switches: Role in cancer progression and prospects for therapy. *Int. J. Cell Biol.* **2013**, 962038.
38. Wei J., Zaika E., Zaika A. 2012. P53 family: Role of protein isoforms in human cancer. *J. Nucleic Acids.* 2012, **687359**. doi 10.1155/2012/687359
39. Hofstetter G., Berger A., Fiegl H., et al. 2010. Alternative splicing of p53 and p73: The novel p53 splice variant p53delta is an independent prognostic marker in ovarian cancer. *Oncogene*. **29**, 1997–2004.
40. Marabese M., Marchini S., Marrazzo E., et al. 2011. Expression levels of p53 and p73 isoforms in stage I and stage III ovarian cancer. *Eur. J. Cancer.* **4**, 131–141.
41. Flaherty K., Puzanov I., Kim K., et al. 2010. Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med*. **363**, 809–819.
42. Nazarian R., Shi H., Wang Q., et al. 2011. Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature*. **468**, 973–977.
43. Poulikakos P.I., Persaud Y., Janakiraman M., et al. 2011. RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). *Nature*. **480**, 387–390.
44. Walsh C.T., Garneau-Tsodikova S., Gatto G.J. 2005. Protein posttranslational modifications: The chemistry of proteome diversifications. *Angew. Chem. Int. Ed.* **44**, 7342–7372.
45. Komander D. 2009. The emerging complexity of protein ubiquitination. *Biochem. Soc. Trans.* **37**, 937–953.
46. Vidal C.J. 2011. *Post-Translational Modifications in Health and Disease*. New York: Springer.
47. Anderson N.L. 2010. The clinical plasma proteome: A survey of clinical assays for proteins in plasma and serum. *Clin. Chem*. **56**, 177–185.
48. Weissman A.M. 2001. Themes and variations on ubiquitylation. *Nat. Rev. Mol. Cell. Biol*. **2**, 169–178.
49. López-Otín C., Blasco M.A., Partridge L., et al. 2013. The hallmarks of aging. *Cell*. **153**, 1194–1217.
50. Powers E.T., Morimoto R.I., Dillin A., et al. 2009. Biological and chemical approaches to diseases of proteostasis deficiency. *Annu. Rev. Biochem*. **78**, 959–991.
51. Thomas S.N., Cripps D., Yang A.J. 2009. Proteomic analysis of protein phosphorylation and ubiquitination in Alzheimer's disease. *Meth. Mol. Biol*. **566**, 109–121.
52. Pagel O., Loroch S., Sickmann A., Zahedi R.P. 2015. Current strategies and findings in clinically relevant post-translational modification-specific proteomics. *Expert Rev. Proteomics*.**12**, 235–253.
53. Kashuba V.I., Li J., Wang F., et al. 2004. *RBSP3* (*HYA22*) is a tumor suppressor gene implicated in major epithelial malignancies. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 4906–4911.
54. Haas-Kogan D.A., Prados M.D., Tihan T., et al. 2005. Epidermal growth factor receptor, protein kinase B/Akt, and glioma response to erlotinib. *J. Natl. Cancer Inst.* **97**, 880–887.
55. Benne R., Van den Burg J., Brakenhoff J.P., et al. 1986. Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*. **46**, 819–826.
56. Gott J.M., Emeson R.B. 2000. Functions and mechanisms of RNA editing. *Annu. Rev. Genet.* **34**, 499–531.

57. Athanasiadis A., Rich A., Maas S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol.* **2**, e391.

58. Anant S., Davidson N.O. 2001. Molecular mechanisms of apolipoprotein B mRNA editing. *Curr. Opin. Lipidol.* **12**, 159–165.

59. Tariq A., Jantsch M.F. 2012. Transcript diversification in the nervous system: a to I RNA editing in CNS function and disease development. *Front. Neurosci.* **6**, 99.

60. Nishikura K. 2010. Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.* **79**, 321–349.

61. Dominissini D., Moshitch-Moshkovitz S., Amariglio N., Rechavi G. 2011. Adenosine-to-inosine RNA editing meets cancer. *Carcinogenesis.* **32**, 1569–1577.

62. Beghini A., Ripamonti C.B., Peterlongo P., et al. 2000. RNA hyperediting and alternative splicing of hematopoietic cell phosphatase (PTPN6) gene in acute myeloid leukemia. *Hum. Mol. Genet.* **9**, 2297–2304.

63. Ranieri M., Brajkovic S., Riboldi G., et al. 2013. Mitochondrial fusion proteins and human diseases. *Neurol. Res. Int.* **2013**, 293893.

64. Mitelman F., Johansson B., Mertens F. 2007. The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer.* **7**, 233–245.

65. Modi H., McDonald T., Chu S., et al. 2007. Role of BCR/ABL gene-expression levels in determining the phenotype and imatinib sensitivity of transformed human hematopoietic cells. *Blood.* **109**, 5411–5421.

66. Demichelis F., Fall K., Perner S., et al. 2007. *TMPRSS2:ERG* gene fusion associated with lethal prostate cancer in a watchful waiting cohort. *Oncogene.* **26**, 4596–4599.

67. Baldo B.A. 2015. Chimeric fusion proteins used for therapy: Indications, mechanisms, and safety. *Drug Saf.* **38**, 455–479.

68. Cobb B.A., Petrash J.M. 2000. Structural and functional changes in the alpha A-crystallin R116C mutant in hereditary cataracts. *Biochemistry.* **39**, 15791–15798.

69. Fiorelli T., Kirouac L., Padmanabhan J. 2013. Altered processing of amyloid precursor protein in cells undergoing apoptosis. *PLoS One.* **8**, e57979.

70. O'Brien R.J., Wong P.C. 2011. Amyloid precursor protein processing and Alzheimer's disease. *Annu. Rev. Neurosci.* **34**, 185–204.

71. Iwawaki T., Oikawa D. 2013. The role of the unfolded protein response in diabetes mellitus. *Semin. Immunopathol.* **35**, 333–350.

72. Schmidt B., Selmer T., Ingendoh A., von Figura K. 1995. A novel amino acid modification in sulfatases that is defective in multiple sulfatase deficiency. *Cell.* **82**, 271–278.

73. Paik S., Shak S., Tang G., et al. 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826.

74. Bedard P.L., Mook S., Piccart-Gebhart M.J., et al. 2009. MammaPrint 70-gene profile quantifies the likelihood of recurrence for early breast cancer. *Expert Opin.* **3**, 193–205.

75. Ruggles K.V., Tang Z., Wang X., et al. 2015. An analysis of the sensitivity of proteogenomic mapping of somatic mutations and novel splicing events in cancer. *Mol. Cell. Proteomics.* **15**, 1060–1071. doi 10.1074/mcp.M115.056226

76. Karve T.M., Cheema A.K. 2011. Small changes, huge impact: The role of protein posttranslational modifications in cellular homeostasis and disease. *J. Amino Acids.* **2011**, 207691. doi 10.4061/2011/207691

77. Khan S., Bennit H.F., Turay D., et al. 2014. Early diagnostic value of survivin and its alternative splice variants in breast cancer. *BMC Cancer.* **14**, 176.

78. Frampton J.P., White J.B., Simon A.B., et al. 2014. Aqueous two-phase system patterning of detection antibody solutions for cross-reaction-free multiplex ELISA. *Sci. Rep.* **4**, 871–874.

79. Tighe P.J., Ryder R.R., Todd I., Fairclough L.C. 2015. ELISA in the multiplex era: Potentials and pitfalls. *Proteomics Clin. Appl.* **9**, 406–422.

80. Vernes J.-M., Meng Y.G. 2015. Detection and quantification of VEGF isoforms by ELISA. *Methods Mol. Biol.* **1332**, 25–37.

81. Bartel F., Schulz J., Böhnke A., et al. 2005. Significance of HDMX-S (or MDM4) mRNA splice variant overexpression and *HDMX* gene amplification on primary soft tissue sarcoma prognosis. *Int. J. Cancer.* **117**, 469–475.

82. Maruyama K., Akiyama M., Kokame K., et al. 2015. ELISA-based detection system for protein S K196E mutation, a genetic risk factor for venous thromboembolism. *PLoS One.* **10**, 1–10.

83. Romero X., Cacete J.D., Engel P. 2014. Determination of soluble tumor necrosis factor receptor 2 produced by alternative splicing. *Methods Mol. Biol.* **1155**, 187–199.

84. Sheehan K.M., Calvert V.S., Kay E.W., et al. 2005. Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. *Mol. Cell. Proteomics.* **4**, 346–355.

85. Petricoin E.F., Bichsel V.E., Calvert V.S., et al. 2005. Mapping molecular networks using proteomics: A vision for patient-tailored combination therapy. *J. Clin. Oncol.* **23**, 3614–3621.

86. Yang Y., Nan Y., Cai J., Xu J., et al. 2016. The Thr to Met substitution of amino acid 118 in hepatitis B virus surface antigen escapes from immune-assay-based screening of blood donors. *J. Gen. Virol.* **97**, 1210–1217.

87. Lisitsa A., Moshkovskii S., Chernobrovkin A., et al. 2014. Profiling proteoforms: Promising follow-up of proteomics for biomarker discovery. *Expert Rev. Proteomics.* **11**, 121–129.

88. Rezeli M., Végvári A., Fehniger T.E., et al. 2011. Moving towards high density clinical signature studies with a human proteome catalogue developing multiplexing mass spectrometry assay panels. *J. Clin. Bioinforma.* **1**, 7. doi 10.1186/2043-9113-1-7

89. Juncker D., Bergeron S., Laforte V., Li H. 2014. Cross-reactivity in antibody microarrays and multi-

plexed sandwich assays: shedding light on the dark side of multiplexing. *Curr. Opin. Chem. Biol.* **18**, 29–37.

90. Archakov A., Ivanov Y., Lisitsa A., Zgoda V. 2009. Biospecific irreversible fishing coupled with atomic force microscopy for detection of extremely low-abundant proteins. *Proteomics*. **9**, 1326–1343.
91. Rabilloud T., Lelong C. 2011. Two-dimensional gel electrophoresis in proteomics: A tutorial. *J. Proteomics.* **74**, 1829–1841.
92. Rogowska-Wrzesinska A., Le Bihan M.-C., Thaysen-Andersen M., Roepstorff P. 2013. 2D gels still have a niche in proteomics. *J. Proteomics.* **88**, 4–13.
93. Jungblut P.R., Holzhütter H.G., Apweiler R., Schlüter H. 2008. The speciation of the proteome. *Chem. Cent. J.* **2**, 16.
94. Barrabés S., Sarrats A., Fort E., et al. 2010. Effect of sialic acid content on glycoprotein pI analyzed by two-dimensional electrophoresis. *Electrophoresis.* **31**, 2903–2912.
95. Naryzhny S.N., Zgoda V.G., Maynskova M.A., et al. 2016. Combination of virtual and experimental 2DE together with ESI LC-MS/MS gives a clearer view about proteomes of human cells and plasma. *Electrophoresis*. **37**, 302–309.
96. Naryzhny S.N., Lisitsa A.V., Zgoda V.G., et al. 2014. 2DE-based approach for estimation of number of protein species in a cell. *Electrophoresis*. **35**, 895–900.
97. Naryzhny S. 2016. Towards the full realization of 2DE power. *Proteomes*. **4**, E33.
98. Unlü M., Morgan M.E., Minden J.S. 1997. Difference gel electrophoresis: A single gel method for detecting changes in protein extracts. *Electrophoresis.* **18**, 2071–2077.
99. Collier T.S., Muddiman D.C. 2012. Analytical strategies for the global quantification of intact proteins. *Amino Acids*. **43**, 1109–1117.
100. Stephens A.N., Hannan N.J., Rainczuk A., et al. 2010. Post-translational modifications and protein-specific isoforms in endometriosis revealed by 2D DIGE. *J. Proteome Res.* **9**, 2438–2449.
101. Chung H.Y. 2011. Effects of SNPs using differentially expressed serum proteins at growth stages on average daily gain in pig. *Mol. Biol. Rep*. **38**, 3777–3785.
102. Huang H.L., Stasyk T., Morandell S., et al. 2006. Biomarker discovery in breast cancer serum using 2-D differential gel electrophoresis/MALDI-TOF/TOF and data validation by routine clinical assays. *Electrophoresis*. **27**, 1641–1650.
103. Wright E.P., Partridge M.A., Padula M.P., et al. 2014. Top-down proteomics: Enhancing 2D gel electrophoresis from tissue processing to high-sensitivity protein detection. *Proteomics.* **14**, 872–889.
104. Zhang Y., Fonslow B.R., Shan B., et al. 2013. Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.* **113**, 2343–2394.
105. Gavin A.-C., Aloy P., Grandi P., et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature.* **440**, 631–636.
106. Veitinger M., Oehler R., Umlauf E., et al. 2014. A platelet protein biochip rapidly detects an Alzheimer's disease-specific phenotype. *Acta Neuropathol.* **128**, 665–677.
107. Shi Y., Xiang R., Horváth C., Wilkins J.A. 2004. The role of liquid chromatography in proteomics. *J. Chromatogr. A*. **1053**, 27–36.
108. Gatlin C.L., Eng J.K., Cross S.T., et al. 2000. Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Anal. Chem. Am. Chem. Soc.* **72**, 757–763.
109. Brosch M., Saunders G.I., Frankish A., et al. 2011. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res.* **21**, 756–767.
110. Cesnik A.J., Shortreed M.R., Sheynkman G.M., et al. 2016. Human proteomic variation revealed by combining RNA-Seq proteogenomics and global post-translational modification (G-PTM. search strategy. *J. Proteome Res.* **15**, 800–808.
111. Bunger M.K., Cargile B.J., Sevinsky J.R., (2007. Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. *J. Proteome Res.* **6**, 2331–2340.
112. Zhang B., Wang J., Wang X., et al. 2014. Proteogenomic characterization of human colon and rectal cancer. *Nature.* **513**, 382–387.
113. Lichti C.F., Mostovenko E., Wadsworth P., et al. 2015. Systematic identification of single amino acid polymorphisms in glioma stem cell-derived chromosome 19 proteins. *J. Proteome Res*. **14**, 778–786.
114. Veenstra T.D. 2011. Where are all the biomarkers? *Expert Rev. Proteomics.* **8**, 681–683.
115. Nedelkov D., Kiernan U.A., Niederkofler E.E., et al. 2005. Investigating diversity in human plasma proteins. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10852–10857.
116. Resing K.A., Meyer-Arendt K., Mendoza A.M., et al. 2004. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem*. **76**, 3556–3568.
117. Chen Y., Zhang J., Xing G., Zhao Y. 2009. Mascot-derived false positive peptide identifications revealed by manual analysis of tandem mass spectra. *J. Proteome Res.* **8**, 3141–3147.
118. Stevens S.M., Jr., Prokai-Tatrai K., Prokai L. 2008. Factors that contribute to the misidentification of tyrosine nitration by shotgun proteomics. *Mol. Cell Proteomics.* **7**, 2442–2451.
119. Ahrné E., Müller M., Lisacek F. 2010. Unrestricted identification of modified proteins using MS/MS. *Proteomics*. **10**, 671–686.
120. Tran J.C., Zamdborg L., Ahlf D.R., et al. 2011. Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature.* **480**, 254–258.
121. Meng F., Du Y., Miller L.M., et al. 2004. Molecular-level description of proteins from *Saccharomyces cerevisiae* using quadrupole ft hybrid mass spectrometry for top down proteomics. *Anal. Chem. Am. Chem. Soc.* **76**, 2852–2858.
122. Roth M.J., Forbes A.J., Boyne M.T. 2nd, et al. 2005. Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in

human proteins by mass spectrometry. *Mol. Cell. Proteomics*. **4**, 1002–1008.

123. Coelho Graça D., Lescuyer P., et al. 2012. Electron transfer dissociation mass spectrometry of hemoglobin on clinical samples. *J. Am. Soc. Mass Spectrom*. **23**, 1750–1756.
124. Jin Y., Peng Y., Lin Z., et al. 2016. Comprehensive analysis of tropomyosin isoforms in skeletal muscles by top-down proteomics. *J. Muscle Res. Cell. Motil*. **37**, 41–52.
125. Ahn Y.H., Ji E.S., Kwon K.H., et al. 2007. Protein phosphorylation analysis by site-specific arginine-mimic labeling in gel electrophoresis and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Anal. Biochem*. **370**, 77–86.
126. Ntai I., LeDuc R.D., Fellers R.T., et al. 2016. Integrated bottom-up and top-down proteomics of patient-derived breast tumor xenografts. *Mol. Cell. Proteomics*. **15**, 45–56.
127. Whiteaker J.R., Lin C., Kennedy J., et al. 2011. A targeted proteomics-based pipeline for verification of biomarkers in plasma. *Nat. Biotechnol*. **29**, 625–634.
128. Paik Y.K., Omenn G.S., Uhlen M., et al. 2012. Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res*. **11**, 2005–2013.
129. Kusebauch U., Campbell D.S., Deutsch E.W., et al. 2016. Human SRMAtlas: A resource of targeted assays to quantify the complete human proteome. *Cell*. **166**, 766–778.
130. Anderson L., Hunter C.L. 2006. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics*. **5**, 573–588.
131. Kuzyk M.A., Smith D., Yang J., et al. 2009. Multiple reaction monitoring-based, multiplexed, absolute quantitation of 45 proteins in human plasma. *Mol. Cell. Proteomics*. **8**, 1860–1877.
132. Wang Q., Chaerkady R., Wu J., et al. 2011. Mutant proteins as cancer-specific biomarkers. *Proc. Natl. Acad. Sci. U. S. A*. **108**, 2444–2449.
133. Su Z.D., Sun L., Yu D.X., et al. 2011. Quantitative detection of single amino acid polymorphisms by targeted proteomics. *J. Mol. Cell Biol*. **3**, 309–315.
134. Wu J., Pungaliya P., Kraynov E., Bates B. 2012. Identification and quantification of osteopontin splice variants in the plasma of lung cancer patients using immunoaffinity capture and targeted mass spectrometry. *Biomarkers*. **17**, 125–133.
135. Mayor U., Peng J. 2012. Deciphering tissue-specific ubiquitylation by mass spectrometry. *Methods Mol. Biol*. **83**, 65–80.
136. Kettenbach A.N., Rush J., Gerber S.A. 2011. Absolute quantification of protein and post-translational modification abundance with stable isotope-labeled synthetic peptides. *Nat. Protoc*. **6**, 175–186.
137. Ossola R., Schiess R., Picotti P., et al. 2011. Biomarker validation in blood specimens by selected reaction monitoring mass spectrometry of N-glycosites. *Methods Mol. Biol*. **728**, 179–194.
138. Conlon K.P., Basrur V., Rolland D., et al. 2013. Fusion peptides from oncogenic chimeric proteins as putative specific biomarkers of cancer. *Mol. Cell. Proteomics*. **12**, 2714–2723.
139. Bogdanow B., Zauber H., Selbach M. 2016. Systematic errors in peptide and protein identification and quantification by modified peptides. *Mol. Cell. Proteomics*. **15**, 2791–2801.
140. Karpova M.A., Karpov D.S., Ivanov M.V., et al. 2014. Exome-driven characterization of the cancer cell lines at the proteome level: The NCI-60 case study. *J. Proteome Res*. **13**, 5551–5560.
141. Sherman J., Molloy M.P., Burlingame A.L. 2012. Why complexity and entropy matter: Information, post-translational modifications, and assay fidelity. *Proteomics*. **12**, 1147–1150.
142. Kondo T. 2014. Casting doubt on the traditional approach of cancer biomarker discovery through proteomics. *Expert Rev. Proteomics*. **11**, 9–12.
143. Sheynkman G.M., Shortreed M.R., Cesnik A.J., Smith L.M. 2016. Proteogenomics: Integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation. *Annu. Rev. Anal. Chem*. **9**, 521–545.

Translated by N. Onishchenko