IN SILICO AND *IN CELLULO* APPROACHES FOR FUNCTIONAL ANNOTATION OF HUMAN PROTEIN SPLICE VARIANTS

O.I. Kiseleva, V.A. Arzumanian, I.Yu. Kurbatov, E.V. Poverennaya*

Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; *e-mail: k.poverennaya@gmail.com

The elegance of pre-mRNA splicing mechanisms continues to interest scientists even after over a half century, since the discovery of the fact that coding regions in genes are interrupted by non-coding sequences. The vast majority of human genes have several mRNA variants, coding structurally and functionally different protein isoforms in a tissue-specific manner and with a linkage to specific developmental stages of the organism. Alteration of splicing patterns shifts the balance of functionally distinct proteins in living systems, distorts normal molecular pathways, and may trigger the onset and progression of various pathologies. Over the past two decades, numerous studies have been conducted in various life sciences disciplines to deepen our understanding of splicing mechanisms and the extent of their impact on the functioning of living systems. This review aims to summarize experimental and computational approaches used to elucidate the functions of splice variants of a single gene based on our experience accumulated in the Laboratory of Interactomics of Proteoforms at the Institute of Biomedical Chemistry (IBMC) and best global practices.

Key words: functional annotation; alternative splicing; splice form; proteoforms; proteome heterogeneity; multiomics studies

DOI: 10.18097/PBMC20247005315

INTRODUCTION

The completion of the Human Genome Project has finally dotted the i's and crossed the t's with regard to the central dogma of molecular biology, demonstrating the scale of the deviation from the principle “one gene — one protein”. For more than a quarter of a century, it has been generally accepted that alternative splicing (AS) in higher eukaryotes increases the diversity of variants of exon sequence combinations and thereby expands the spectrum of protein products (Fig. 1) encoded by a relatively small set of genes [1]. More than 95% of multi-exon human genes produce more than one (constitutive) splice variant [2]. With the development of high-throughput methods for analyzing nucleotide and amino acid sequences, a wide range of researchers have the opportunity to analyze molecular profiles of objects of interest not only at the level of the master protein (i.e., the generalized image of the protein products of a gene [3]), but also at the level of specific proteoforms.

1. ASSOCIATION OF ALTERNATIVE SPLICING WITH DISEASE

“Fine-tuning” of the entire biological system due to AS is capable of changing its molecular composition. For example, titin, which has the longest amino acid sequence (reaching 38,138 residues), changes its predominant splice form as a person ages [4]. Such modifications change the length of the protein and relative stiffness, and therefore affect ventricular tension at rest and are associated with acquired forms of heart failure. Transcriptional and post-translational changes that increase the length and extensibility of titin, making the sarcomere longer and softer, are associated with systolic dysfunction and left ventricular dilation. Titin modifications that shorten both the protein itself and the sarcomere are associated with diastolic dysfunction [5].

In recent years, a trend has emerged in studies of molecular heterogeneity to search for a relationship between splicing patterns and the occurrence and development of diseases. In general, such studies are panoramic in nature and are aimed at forming the most complete transcriptomic and translational profiles of the studied objects [6–9]. Good evidence now exists that impairments in splicing mechanisms, depending on their scale, can trigger production of functionally inactive proteins [10]. Such proteins, exhibiting altered functions, distort the well-established processes of differentiation, growth, intercellular communication and apoptosis, in other words, disrupt the normal functioning of the entire organism and can lead to its death [11]. The connection between splicing aberrations is especially clearly demonstrated by the example of “hallmarks of cancer”: each distinctive feature characteristic of oncological diseases corresponds to a case of disrupted splicing (Fig. 2). In 2022, the classic ten hallmarks of cancer (proliferative signaling, replicative immortality, angiogenesis induction, growth suppressor evasion, etc.) were supplemented by the epigenetic reprogramming [12], polymorphic microbiome [13], cell aging [14, 15], and unlocking of phenotypic plasticity [16].

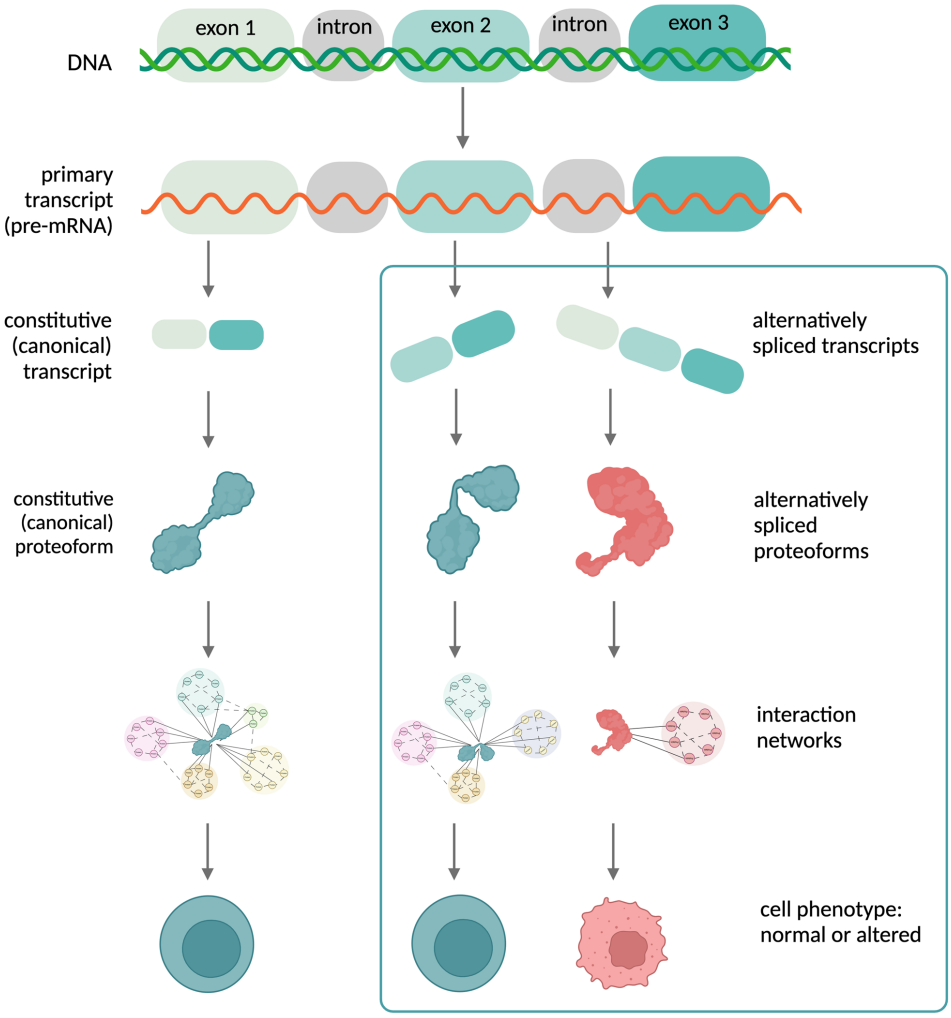

Figure 1. Stages of genetic information transfer that form the final phenotype of an object. As a result of alternative splicing (AS), one gene can produce several splice variants, which repersent different combinations of exons (less often with introns included) and transcribed from different 3′- and 5′- splice sites.

It is difficult to find an example of an oncological disease, for which cases of the disruption of normal splicing patterns have not been identified. Knowledge has been accumulated about splicing aberrations both in solid tumors (in the brain [17], liver [18], skin [19], kidneys [20], lungs [21], breast [22], cervix [23], ovaries [24], prostate [25]), and in oncohematology. Most cases of pathological AS are explained by genetic mutations in constitutive splicing sites or impairments in the expression level of spliceosome regulatory factors [26].

Numerous results of similar studies have revealed the switching effect, which describes a change or "switch" in the predominantly expressed splice variant of a gene during the transition from normal tissue to pathologically altered tissue [27, 28]. This phenomenon is determined as changes in the proportions of splice variants and/or their differential expression of all expressed transcripts. Sufficient evidence has been accumulated that changes in the abundance of splice variants of one gene affect the development and functioning of a living cell in normal conditions and pathology [29–31]. In general, such conclusions are made based on an analysis of the enrichment of gene sets or metabolites, rather than a detailed study of cause-and-effect relationships. In addition, switching analysis and interpretation of results of transcriptome profiling, consider all possible transcript variants, including those, which are not translated into the amino acid sequence [32]. As we have shown earlier, focusing exclusively on protein-producing transcripts expectedly reduces the array of cases of switching the predominant splice form, but does not reduce it to tens of cases [28], indirectly confirming the hypothesis about different functional properties of proteoforms.

The transition from the quantity of data to their qualitative (functional) understanding may serve as protection against the tempting "*post hoc, ergo propter hoc*" fallacy in searching for biological meaning in sequencing data on alternative splicing switching. In order to fully exploit knowledge of splicing patterns as a source of diagnostic, prognostic, predictive, and therapeutic tools, it is necessary not only to notice the numerical patterns, but also to determine the functions of splice variants encoded by a single gene.

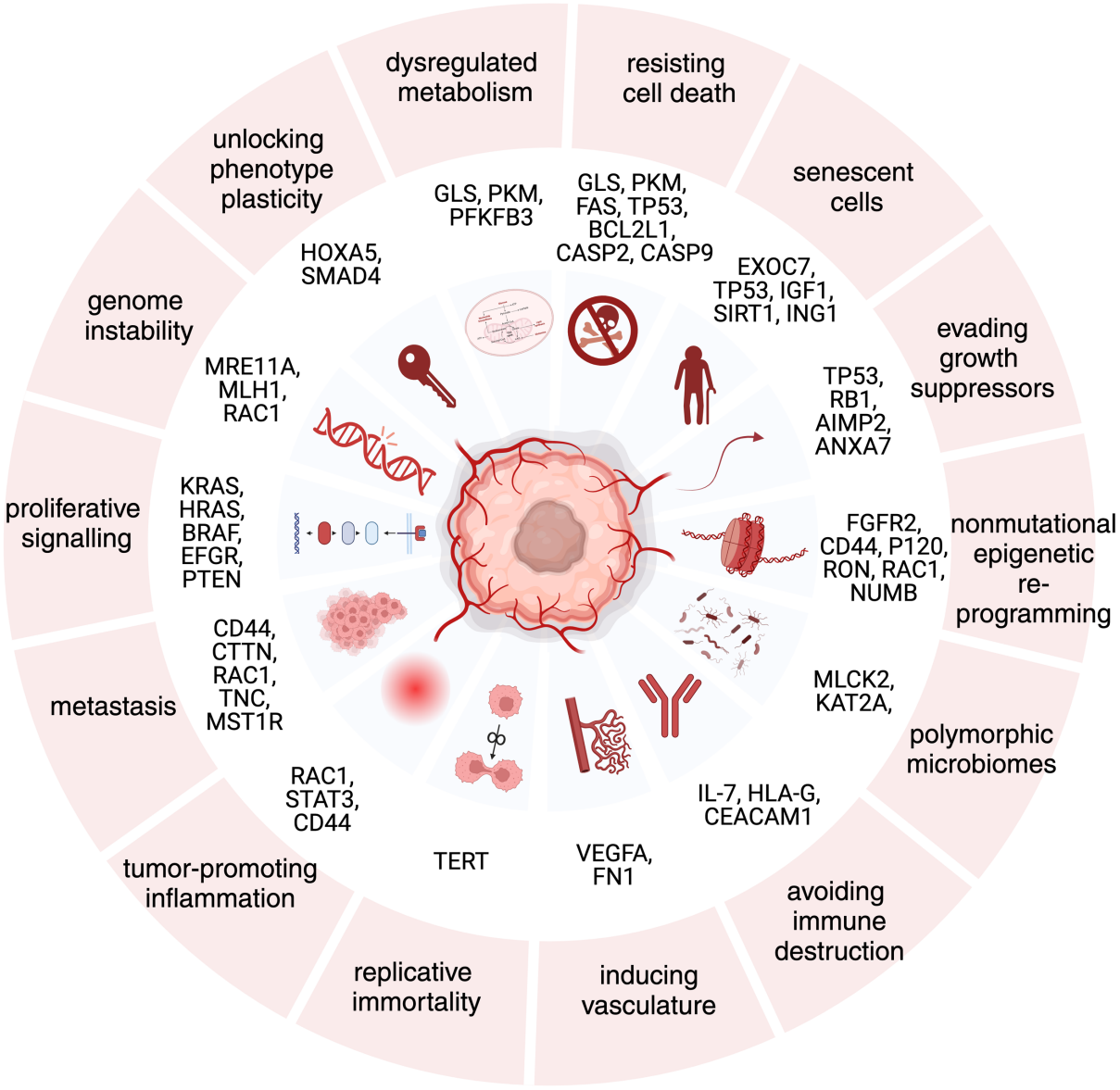

Figure 2. The place of alternative splicing in the realization of the current version of "hallmarks of cancer": a concept designed to reduce the complexity of oncophenotypes to a conventional set of principles common to the origin and development of various cancers. The illustration is built taking into account the evolution of the concept over the last 20 years [26–28] and provides landmark examples of genes, whose splicing is reliably associated with the occurrence and development of cancer.

2. FUNCTIONAL HETEROGENEITY OF SPLICES VARIANTS

Differently spliced protein products exhibit different enzymatic activities, are localized in different cellular compartments [33], and often behave as separate proteins rather than as interchangeable variants of each other. They can exhibit dominant-negative effects in relation to other forms encoded by the same gene, be expressed to a greater or lesser extent than the constitutive variant, or even have opposite functions.

The most demonstrative (but not the only) example of a gene with functionally different splice variants is BCL2L1 from the family of apoptosis regulators. Splice variants of this gene differ due to the presence of alternative 5′ splice sites in the second exon; the short form (BCL-XS) triggers cell death processes, while the long form (BCL-XL) has an anti-apoptotic function and is often activated in cancer [34]. The identified pattern is promising from a practical point of view: antisense therapy aimed at changing the ratio between two BCL proteoforms may increase the sensitivity of cells to chemotherapeutic drug-induced apoptosis [35]. A similar situation exists with a representative of the tumor necrosis factor receptor superfamily TNFR2, which also encodes two splice variants with antagonistic functions. The canonical variant of the TNFR2 receptor mediates TNF-α-induced apoptosis, while the shorter DS-TNFR2 variant, lacking the amino acid sequence encoded by the seventh and eighth exons, blocks apoptosis [35]. Another example is the IG20 gene, which

is overexpressed in cancer cells and encodes at least six splice forms (IG20, IG20-PA or IG20pa, DENN or MADD, DENN-SV, KIAA0358, and IG20-SV4), differing in their apoptotic properties [36, 37].

The above examples of functional differences between individual splice variants look impressive, but such a depth of study and elucidation of the role of splicing in the formation of a healthy or altered phenotype is rather exceptional and has been achieved only for several dozen protein-coding genes, often at the level of transcripts and translators. It took several years and several iterative approaches to studying the sources of heterogeneity in omics layers to achieve consensus on issues of alternative splicing. The path of the Laboratory of Structural and Computational Biology of the Spanish National Cancer Research Center is very demonstrative in this context. Initially, in 2015, based on the results of eight large-scale proteomic experiments and analysis of databases deposited in the database, it was suggested that AS did not play a significant role in the formation of protein diversity [38]. The modest set of alternatively spliced proteins detected led researchers to the idea that most protein-coding genes probably produced only one — canonical — protein product [39], and most alternative variants did not withstand selective pressure and could be functionally insignificant at all. Several years later, the same researchers refuted their previous assertions by analyzing alternative splicing at the protein level [40], which was tissue-specific for a third of genes. More recently, the same group, based on the results of large proteomic experiments, developed the bioinformatics tool TRIFID for predicting the functional significance of splice forms and returned to the assertion that 85% of alternative transcript variants were likely to be insignificant [41]. Such back-and-forth movements in the study of heterogeneity at different omics levels encourage caution in studying the adaptivity of AS and trying to establish a correspondence between the diversity of mRNA variants and the proteins they encode.

3. PROBLEMS IN FUNCTIONALITY STUDIES AT THE PROTEOME LEVEL

The findings obtained at the transcriptome level are difficult to transfer to the proteome, especially in the context of attempts to establish quantitative patterns. The Pearson coefficient for correlation between the transcript and protein abundances usually does not exceed 0.5, as shown in studies analyzing the available relationships to study the possibility of constructing a protein abundance model based on transcriptome and translatome data [42–45]. Additional efforts to model the effect of protein synthesis regulation after pre-mRNA splicing were able to explain 30% of the differences in protein and mRNA ratios [43]. Another achievement in the search for consistency between transcript and protein abundance was the rather natural observation that exceeding certain transcript expression levels was a good predictor of protein expression [46]. Nevertheless, it is obvious that a complete understanding of the living system functioning requires information about the proteome component.

The peptide-centric nature of proteomic data do not often discriminate individual splice forms [43–45, 47] because of problems of isolation and reliable detection of proteoform-specific peptides [48]. Protease cocktails [49], *de novo* data processing [50, 51], and orthogonal sequencing technologies [52, 53] have been used to improve the quality of protein sequence coverage; however, even the sum of these efforts does not provide a full assessment of the number of proteoforms.

Today, there are still many blank spots on the map of systems biology, but the general consensus exists that proteins are the driving force of living systems. The study of individual protein variants has already achieved significant success [54–58]. We have proposed to enhance the results of mass spectrometric profiling of HepG2 cell line proteins distributed across the cells of a two-dimensional gel according to their physicochemical properties [59] by using a customized search library. Such library, built on the basis of transcriptome sequencing data for the studied HepG2 cells, generates a most accurate search space of expected proteoforms. On the one hand, this is achieved by taking into account splice variants and sequences with point substitutions specific to a particular sample. On the other hand, the volume of the search space can be limited by ignoring sequences whose production, according to transcriptome data, should not be expected in the studied sample. Integrated analysis in HepG2 cells increased the number of identified proteoforms by 76% compared to standard panoramic profiling without preliminary fractionation on a two-dimensional gel. This effect is achieved as a result of the synergy of two factors: firstly, a decrease in the complexity of the biological mixture, and secondly, additional knowledge about the physicochemical properties of proteoforms [52, 53].

The use of original mass spectrometric approaches [60] and antibody enrichment technologies [61] has brought proteomics closer to answering the question: how does the final sequence and structure form the functionality of protein variants and how do individual protein variants affect the viability of the entire living system?

4. TRANSITION FROM DIFFERENTIAL EXPRESSION DATA TO FUNCTIONAL ANNOTATION

The popularization of RNA sequencing methods and the ability to analyze the abundance of genes and individual splice variants have made it possible

to accumulate data demonstrating the difference in their expression under different conditions. Hundreds of papers have been published comparing the abundance of gene products in normal and tumor tissues and assessing changes in molecular profiles after gene knockout or knockdown [62, 63]. The results of such experiments can be reused, in particular, to identify new protein functions [33]. Detection of differentially expressed genes to determine protein functions is a standard approach applicable to transcriptomics. Transcriptomics has matured as a field of science, and protocols for transcriptomic data collection and processing are optimized, reliable, and efficient [64]. This allows to develop computational methods for systematical studies of protein functions at the isoform level [65–68].

Elucidation of protein function is the main "activation barrier" in the formation of a systematic understanding of the structure of living systems. High-throughput sequencing is becoming increasingly accessible, data are multiplying, but a qualitative transition from information on the expression of a single gene to understanding the role of a specific protein molecule has not yet occurred. Despite significant advances in proteome cataloging, fundamental questions regarding the roles of individual proteins in the complex proteome mechanism have not been resolved yet. The reason for this is a significant difference in the structures and sizes of the information space of the transcriptome and proteome [69].

5. PRACTICE OF PROTEIN FUNCTION DETERMINATION

In the absence of a generally accepted standard for protein function determination, most of the functional annotations are predictive due to the widespread use of bioinformatics methods in addition to large omics data and the existing, essentially fragmentary, information on cellular processes regulation. The existing gap between the methods of experimental and computational biochemistry in terms of labor intensity, cost, and rapidity explains the predominance of bioinformatic predictions over empirical evidence in determining protein functions. In our previous retrospective studies, using the example of the neXtProt database, known for the completeness and reliability of published information on human proteins, we analyzed the evolution of the terminology used to describe protein function [70]. We have noticed that in most cases the accumulated annotations are achieved by computational methods, but even the best bioinformatics tools often yield unsatisfactory results, when it comes to annotating non-canonical variants. The guilty by association postulate is often used: based on the results of affinity purification – mass spectrometry (AP-MS) and yeast-two-hybrid (Y2H) analysis technology, protein function is attempted to be mapped to biochemical processes by studying contacts or "handshakes" of target proteins. Another problem we have noticed studying the trends in functional annotation of proteins concerns repositories with manual verification of deposited data. Dataframes (i.e., tabular systems of the "observations – variables" data architecture) of such repositories (e.g., neXtProt) are not optimized for efficient storage of information in a proteoform-centric mode. Currently, neXtProt provides information for approximately 10,000 splice variants, 10% of which are differently localized in the cell and have distinct functions within a single gene [71].

Within the framework of existing experimental approaches for functional annotation of proteins, two directions can be distinguished: 1) knockout or alteration of the expression level of the gene of interest to identify altered molecular pathways on the basis of the analysis of one or more omics levels and 2) interactome analysis.

5.1. Loss-of-Function or Gain-of-Function

Suppression of gene expression apparently alters the biological processes, in which the protein it encodes is involved. These changes are not easy to detect and/or unambiguously interpret: the available information on molecular pathways is fragmentary, and the pathways themselves are non-linear and often duplicate a number of steps in cellular processes. Additional complexity in such studies is introduced by targeted changes in expression: introducing point mutations that disrupt the reading frame, or using the interference phenomenon. Currently, the most popular method of knockout, as well as knockdown or knocking (in situations where knockout is impossible or, conversely, expression is too low) is the use of genetic editing based on the CRISPR-Cas9 system [72], which gives a more predictable and stable result than interfering microRNAs. As we have previously shown, the capabilities of CRISPR-Cas9-based methods for studying the properties of proteins and their diversity in proteomic studies are impressive, but such technologies have not yet been widely used to study splice forms [72].

Results of several studies, applying genetic editing methods to splice forms, demonstrate the importance of their presence for maintaining the function or phenotype of the studied object. For example, knockout of splice forms encoded by the Reep6 gene has shown that the canonical Reep6.1 variant is critically important for retinal rod functioning [73]. At the same time, the second splice form is important for maintaining fertility in male mice, and both variants are expressed in the testes. Comparative proteomic and phenotypic analysis of ES-2 and OVCAR-8 cell lines with knocked-out splice forms of the TGFβ receptor revealed the different roles of these proteoforms

in the development of ovarian cancer [74]. Evaluation of splice forms specific to gastric cancer by using promoter knockout revealed that in the ZFHX3 tumor suppressor, splice forms had opposite functions, similar to BCL2 variants [75]. In the case of Duchenne muscular dystrophy, a genetic disease caused by dystrophin translational defects due to frameshift mutations, methods for excluding exons with such errors to express a shorter version of the protein are developed for treatment of this disease. Suppression of expression of the canonical variant has been shown to be effective in partially preserving the function of alternatively spliced dystrophin [76, 77].

In our study of the function of the mitochondrial protein importer TOMM34, we assessed changes before/after its knockout at the transcriptomic, proteomic, and metabolomic levels. One of the criteria for choosing TOMM34 (in addition to its poorly investigated role) was the presence of only one translated mRNA, according to UniProt, which made it possible to describe the functional role of a specific amino acid sequence [63]. For complete suppression of TOMM34 expression during TOMM34 knocking out, we had to introduce five mutations into the first exon. Focusing on a specific splice form, in addition to technical difficulties, increases the risks of frameshifting and additional off-target effects, which can in turn affect the cascade of molecular events [78].

5.2. Functional Annotation via the Interactome

Most splice variants within a single gene have less than 50% common partners for intermolecular interactions [33]. It is noteworthy that partners, interacting with a particular splice variant, are usually expressed in a highly tissue-specific manner and belong to separate functional modules [34]. The association of protein partners with certain functions helps to assume that the studied protein is also associated with it (i.e., using the guilty by association concept). In the case of splice forms, an additional complication is that functional annotation is carried out using the combined information for proteoforms encoded by a single gene.

At present, interactome profiles have been formed for 80% of human genes [79, 80], and a trend has emerged towards identifying protein-protein interactions (PPIs) for splice forms. Experimentally, binary interactions can be studied by two-hybrid methods using characteristic sequences as a target protein or by AP-MS, which provides information about the molecular complex. One of the most impressive studies on the definition of PPIs for splice forms was performed in 2016, in which interactome profiles were determined for 366 out of more than a thousand splice forms studied using Y2H [36]. The authors of the study showed that even on a small sample, the interactome network increased by 3.2 times as compared to the gene-centric approach.

In AP-MS experiments, protein partners are identified by mass spectrometry, i.e., by peptides that are uniquely mapped to the amino acid sequence of a specific proteoform [81]. The IntACT database is one of the first resources containing data on PPIs for splice forms [82]; it deposits the results of interactome experiments, including those performed using AP-MS methods. Despite the fact that AP-MS-based approaches make it possible not only to use the splice form as a target protein but also to describe specific proteoforms in the resulting complexes, no emphasis is placed on splice forms [83]. An illustrative example is one of the largest BioPlex projects: despite the mention of splice forms, the data are provided in a gene-centric format [84].

The accumulated AP-MS results make it possible to identify new interactions, including splice-specific PPIs. For example, we have reanalyzed the mass spectrometric data array of the BioPlex 2.0 project. Based on the statistically significant frequency of co-occurrence, we were able to identify 287,474 interactions, predicted for the first time a function for 391 proteins and for 31 genes demonstrated a difference in the interactome profiles of the splice forms encoded by them [85].

Although PMIs are the basis of cellular processes, protein-metabolite interactions (PMIs) also play a significant role; the methods for their identification we described in [86]. The value of PMIs for unraveling the splice form interactome can be illustrated by the example of 505 splice forms: their unique role in biochemical processes is noticeable, when analyzing interactions with small molecules, but is not distinguished by standard analysis protocols [87].

Bioinformatic algorithms represent a significant part of the interactome methods. They compare sequences and extrapolate interactome profiles and functions from studied objects to unstudied ones or integrate different types of data to improve interactome annotation. Five years ago we described existing approaches to proteoform annotation [88], and their basic principles did not change since that time. However, at present, more and more works are devoted to the development of methods for integrating PPI data with expression/translation information to recognize specific interactions directly for splice forms. For example, using the DIGGER method, based on the combined analysis of interactomes, data on the interaction of domains and amino acid residues, as well as expression, it is possible to extract splice form-specific subnetworks [89]. In a more advanced method, LINDA, data on transcription factors, as well as the DIGGER results, are additionally used to decipher the interactomes of splice forms [90]. NEASE (Network Enrichment for AS Events) is another example of an even more specialized resource for predicting splice variant functions [91]; during identification of subnetworks for splice forms this resource takes into consideration not only interactome

data and quantitative changes, but also structural features of proteins according to DOMINE [92], 3did [93], Eukaryotic Linear Motif [94], and PDB [95].

5.3. Bioinformatics Methods and Tools

Bioinformatics methods for functional annotation of proteoforms are developing not only within the framework of interactome analysis. Over 20 years of active omics research, it was possible to build the architecture of various knowledge bases. These include both highly specialized knowledge bases focused exclusively on splicing [96–99], and comprehensive knowledge bases such as UniProt [100], neXtProt [101], RefSeq (NCBI) [102], Ensembl-GENCODE (EMBL) [103], and MANE [104]. These resources summarize the results of hundreds and thousands of experiments capable of distinguishing splice variants of a single gene at different omics levels. Each of the listed resources with annotations of human amino acid and nucleotide sequences provides information on more than 100 thousand protein-coding transcripts [105, 106]. At the same time, the pools of reported data differ greatly: for example, in 2018 it was shown that the ENCODE and RefSeq databases agreed on only 1/6 of all the presented transcripts [105].

Accumulated knowledge and experimental data have facilitated the development of bioinformatic methods for assessing protein function. Sequence alignment and expression analysis are commonly used to study gene function. The main idea for using these approaches is that sequences that are conserved across species are likely to be functional [107]. Similar logic is used for splice variants: the conservation of a particular sequence (e.g., a protein encoded by an alternative exon) in species that diverged evolutionarily tens of millions of years ago indicates its functional significance [108]. Similarly, the more evidence there is of the expression of a particular transcript in different species, the more likely it is to be functional. Functional annotation is complicated by the fact [109] that the expression of most splice variants in humans is very low (often at the level of biological noise [110]), and tissue-specific [105].

Almost any functional analysis is focused on comparison of data on the known role of the protein of interest in cellular processes. Among the various resources describing and cataloging the functional annotation of genes and their products, the most popular is the Gene Ontology (GO) terminology system [111]. The system includes 45,000 terms and combines them into three subontologies: molecular functions (i.e., terms describing protein functions, such as kinase activity); biological processes (to describe the sequence of events occurring in a cell or organism that involve genes and proteins they encode, such as cell division or immune response), and cellular components (terms describing protein localization: nucleus, membrane, etc.). GO offers functional annotations only at the gene level, without detailing information for individual splice variants [112]. To date, several algorithms have been developed: iMILP [67], mi-SVM [113], WLRM [114], IsoResolve [115], DeepIsoFun [116], and DIFFUSE [117], designed to refine GO terms in relation to splice forms, based on multivariate learning. The current limitation of such methods is the low prediction accuracy due to the difficulty of taking into account the hierarchical structure and extensive semantics of GO terms during analysis. Moreover, most algorithms assume that only one splice form is responsible for the implementation of the gene function, although in reality several variants can interact simultaneously to perform this function [70]. In a more advanced algorithm, IsofunGO [116], special attention is paid to detection of different annotations for individual isoforms.

Three-dimensional protein structure prediction technologies, feeding an RNA or protein sequence as input, are also used to determine the function of splice variant functions. The popular AlphaFold2 program allows automatic prediction of 3D protein structures with high accuracy [108, 118, 119], including splice forms. It is assumed that protein molecules, whose sequence folds into an ordered structure (about 68% of the human proteome [118, 120]), are most likely functional, and, conversely, poorly folded molecules are most likely inoperative. A similar approach was tested on a data pool, summarizing the results of more than 10 thousand transcriptomic experiments [108], using an optimized version of AlphaFold known as the ColabFold program [121]. An illustration of the determination of the splice form function is the analysis of the ASMT gene (encoding N-acetylserotonin-O-methyltransferase), which is involved in melatonin biosynthesis. Based on the combined transcriptome profiling of the pineal gland, biopsied at night (due to its melatonin synthesis) [122], it was determined that the variant that showed a more stable protein assembly was expressed more than 10 times higher than other splice forms.

The scientific community has long relied on two methods for protein function discovery and prediction: RNA-seq and DNA and protein sequence alignment to detect evolutionary conservation. Currently, hope is pinned on computational methods based on GO, protein structure prediction, and interactome interactions. It is worth recognizing that these are just the beginning. The development of computational methods for splice variant function discovery will remain an area of active research for many years.

CONCLUSIONS

Researchers tend to study the same proteins recurrently. Only 30 proteins of the human brain account for 2/3 of all scientific literature

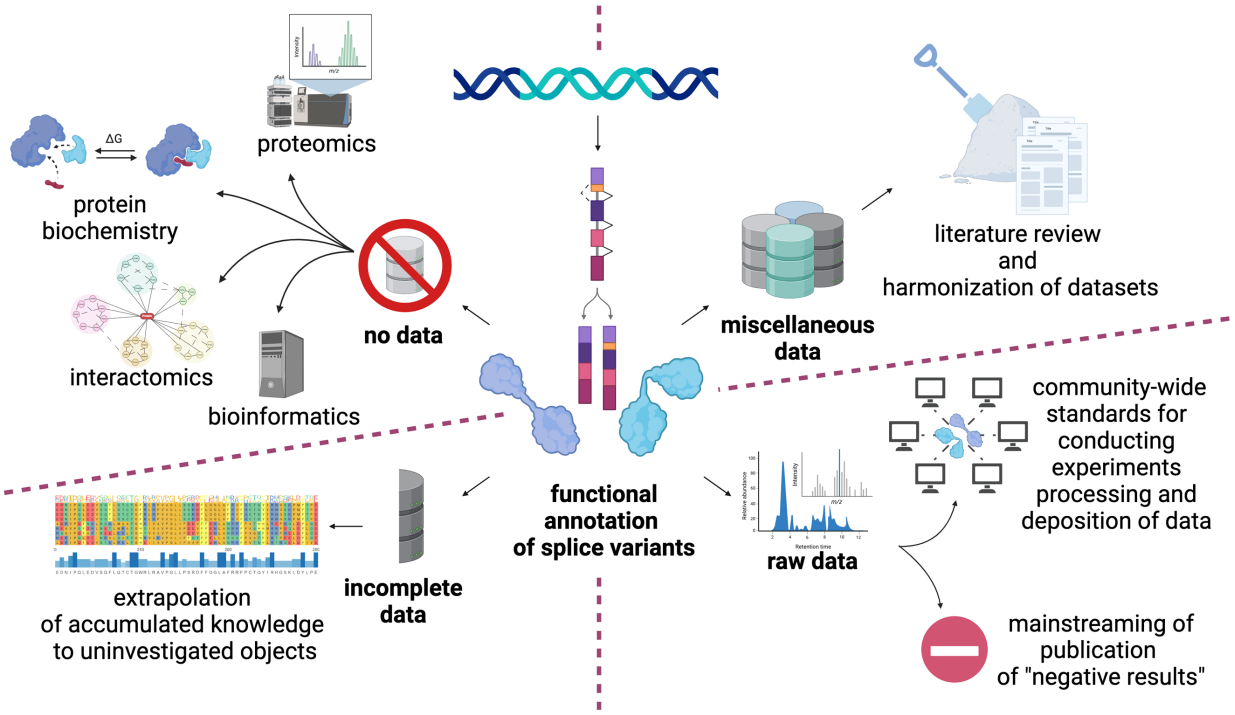

Figure 3. Proposed roadmap for functional annotation of the heterogeneous proteome.

devoted to the analysis of the brain proteome. The Matthew effect, according to which "the rich get richer and the poor get poorer," is also reflected in the issue of functional annotation of proteins. This state of affairs is facilitated rather by the availability of funding and existing technological capabilities than by the fundamental or practical value of studying a certain set of already well-studied proteins. The situation is naturally aggravated when moving from master proteins, which we define as a set of protein products of a single gene without specifying proteoforms [69], to specific protein variants.

Having studied the degree of elaboration of the issue of functional annotation of individual splice forms, we propose a roadmap (Fig. 3) and hope that the organized efforts of all community members will allow significant progress in understanding the structure of living systems.

FUNDING

This study was supported by the Russian Science Foundation (project no. 21-74-10061).

COMPLIANCE WITH ETHICAL STANDARDS

This article does not contain any research involving humans or the use of animals as objects.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. *Graveley B.R.* (2001) Alternative splicing: Increasing diversity in the proteomic world. Trends Genet., **17**(2), 100–107. DOI: 10.1016/S0168-9525(00)02176-4
2. *Lee Y., Rio D.C.* (2015) Mechanisms and regulation of alternative pre-mRNA splicing. Annu. Rev. Biochem., **84**, 291–323. DOI: 10.1146/annurev-biochem-060614-034316
3. *Archakov A., Aseev A., Bykov V., Grigoriev A., Govorun V., Ivanov V., Khlunov A., Lisitsa A., Mazurenko S., Makarov A.A., Ponomarenko E., Sagdeev R., Skryabin K.* (2011) Gene-centric view on the human proteome project: The example of the Russian roadmap for chromosome 18. Proteomics, **11**(10), 1853–1856. DOI: 10.1002/pmic.201000540
4. *Tharp C.A., Haywood M.E., Sbaizero O., Taylor M.R.G., Mestroni L.* (2019) The giant protein titin's role in cardiomyopathy: Genetic, transcriptional, and post-translational modifications of TTN and their contribution to cardiac disease. Front. Physiol., **10**, 1436. DOI: 10.3389/fphys.2019.01436
5. *Tharp C., Mestroni L., Taylor M.* (2020) Modifications of titin contribute to the progression of cardiomyopathy and represent a therapeutic target for treatment of heart failure. J. Clin. Med., **9**(9), 2770. DOI: 10.3390/jcm9092770
6. *Vitting-Seerup K., Sandelin A.* (2019) IsoformSwitchAnalyzeR: Analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. Bioinformatics, **35**(21), 4469–4471. DOI: 10.1093/bioinformatics/btz247
7. *Vitting-Seerup K., Sandelin A.* (2017) The landscape of isoform switches in human cancers. Mol. Cancer Res., **15**(9), 1206–1220. DOI: 10.1158/1541-7786.MCR-16-0459
8. *Nowicka M., Robinson M.D.* (2016) DRIMSeq: A Dirichlet-multinomial framework for multivariate count outcomes in genomics. F1000Research, **5**, 1356. DOI: 10.12688/f1000research.8900.2

9. *Anders S., Reyes A., Huber W.* (2012) Detecting differential usage of exons from RNA-seq data. Genome Res., **22**(10), 2008–2017. DOI: 10.1101/gr.133744.111
10. *Liu Q., Fang L., Wu C.* (2022) Alternative splicing and isoforms: From mechanisms to diseases. Genes, **13**(3), 401. DOI: 10.3390/genes13030401
11. *Sterne-Weiler T., Sanford J.R.* (2014) Exon identity crisis: Disease-causing mutations that disrupt the splicing code. Genome Biol., **15**, 201. DOI: 10.1186/gb4150
12. *Pradella D., Naro C., Sette C., Ghigna C.* (2017) EMT and stemness: Flexible processes tuned by alternative splicing in development and cancer progression. Mol. Cancer, **16**, 8. DOI: 10.1186/s12943-016-0579-2
13. *Zou C., Zan X., Jia Z., Zheng L., Gu Y., Liu F., Han Y., Xu C., Wu A., Zhi Q.* (2023) Crosstalk between alternative splicing and inflammatory bowel disease: Basic mechanisms, biotechnological progresses and future perspectives. Clin. Transl. Med., **13**(11), e1479. DOI: 10.1002/ctm2.1479
14. *Georgilis A., Klotz S., Hanley C.J., Herranz N., Weirich B., Morancho B., Leote A.C., d'Artista L., Gallage S., Seehawer M., Carroll T., Dharmalingam G., Wee K.B., Mellone M., Pombo J., Heide D., Guccione E., Arribas J., Barbosa-Morais N.L., Heikenwalder M., Thomas G.J., Zender L., Gil J.* (2018) PTBP1-mediated alternative splicing regulates the inflammatory secretome and the pro-tumorigenic effects of senescent cells. Cancer Cell, **34**(1), 85–102.e9. DOI: 10.1016/j.ccell.2018.06.007
15. *Deschênes M., Chabot B.* (2017) The emerging role of alternative splicing in senescence and aging. Aging Cell, **16**(5), 918–933. DOI: 10.1111/acel.12646
16. *Yuan S., Norgard R.J., Stanger B.Z.* (2019) Cellular plasticity in cancer. Cancer Discov., **9**(7), 837–851. DOI: 10.1158/2159-8290.CD-19-0015
17. *Babic I., Anderson E.S., Tanaka K., Guo D., Masui K., Li B., Zhu S., Gu Y., Villa G.R., Akhavan D., Nathanson D., Gini B., Mareninov S., Li R., Camacho C.E., Kurdistani S.K., Eskin A., Nelson S.F., Yong W.H., Cavenee W.K., Cloughesy T.F., Christofk H.R., Black D.L., Mischel P.S.* (2013) EGFR mutation-induced alternative splicing of Max contributes to growth of glycolytic tumors in brain cancer. Cell Metab., **17**(6), 1000–1008. DOI: 10.1016/j.cmet.2013.04.013
18. *Duriez M., Mandouri Y., Lekbaby B., Wang H., Schnuriger A., Redelsperger F., Guerrera C.I., Lefevre M., Fauveau V., Ahodantin J., Quetier I., Chhuon C., Gourari S., Boissonnas A., Gill U., Kennedy P., Debzi N., Sitterlin D., Maini M.K., Kremsdorf D., Soussan P.* (2017) Alternative splicing of hepatitis B virus: A novel virus/host interaction altering liver immunity. J. Hepatol., **67**(4), 687–699. DOI: 10.1016/j.jhep.2017.05.025
19. *Jensen M.A., Wilkinson J.E., Krainer A.R.* (2014) Splicing factor SRSF6 promotes hyperplasia of sensitized skin. Nat. Struct. Mol. Biol., **21**(2), 189–197. DOI: 10.1038/nsmb.2756
20. *Sokół E., Kędzierska H., Czubaty A., Rybicka B., Rodzik K., Tański Z., Bogusławska J., Piekiełko-Witkowska A.* (2018) MicroRNA-mediated regulation of splicing factors SRSF1, SRSF2 and hnRNP A1 in context of their alternatively spliced 3′UTRs. Exp. Cell Res., **363**(2), 208–217. DOI: 10.1016/j.yexcr.2018.01.009
21. *Sheng J., Zhao Q., Zhao J., Zhang W., Sun Y., Qin P., Lv Y., Bai L., Yang Q., Chen L., Qi Y., Zhang G., Zhang L., Gu C., Deng X., Liu H., Meng S., Gu H., Liu Q., Coulson J.M., Li X., Sun B., Wang Y.* (2018) SRSF1 modulates PTPMT1 alternative splicing to regulate lung cancer cell radioresistance. EBioMedicine, **38**, 113–126. DOI: 10.1016/j.ebiom.2018.11.007
22. *Xie R., Chen X., Chen Z., Huang M., Dong W., Gu P., Zhang J., Zhou Q., Dong W., Han J., Wang X., Li H., Huang J., Lin T.* (2019) Polypyrimidine tract binding protein 1 promotes lymphatic metastasis and proliferation of bladder cancer via alternative splicing of MEIS2 and PKM. Cancer Lett., **449**, 31–44. DOI: 10.1016/j.canlet.2019.01.041
23. *Liu F., Dai M., Xu Q., Zhu X., Zhou Y., Jiang S., Wang Y., Ai Z., Ma L., Zhang Y., Hu L., Yang Q., Li J., Zhao S., Zhang Z., Teng Y.* (2018) SRSF10-mediated IL1RAP alternative splicing regulates cervical cancer oncogenesis via mIL1RAP-NF-κB-CD47 axis. Oncogene, **37**(18), 2394–2409. DOI: 10.1038/s41388-017-0119-6
24. *Iborra S., Hirschfeld M., Jaeger M., Zur Hausen A., Braicu I., Sehouli J., Gitsch G., Stickeler E.* (2013) Alterations in expression pattern of splicing factors in epithelial ovarian cancer and its clinical impact. Int. J. Gynecol. Cancer, **23**(6), 990–996. DOI: 10.1097/IGC.0b013e31829783e3
25. *Fan L., Zhang F., Xu S., Cui X., Hussain A., Fazli L., Gleave M., Dong X., Qi J.* (2018) Histone demethylase JMJD1A promotes alternative splicing of AR variant 7 (AR-V7) in prostate cancer cells. Proc. Natl. Acad. Sci. USA, **115**(20), E4584–E4593. DOI: 10.1073/pnas.1802415115
26. *Zhang Y., Qian J., Gu C., Yang Y.* (2021) Alternative splicing and cancer: A systematic review. Signal Transduct. Target. Ther., **6**, 78. DOI: 10.1038/s41392-021-00486-7
27. *Sebestyén E., Zawisza M., Eyras E.* (2015) Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. Nucleic Acids Res., **43**(3), 1345–1356. DOI: 10.1093/nar/gku1392
28. *Dolgalev G., Poverennaya E.* (2023) Quantitative analysis of isoform switching in cancer. Int. J. Mol. Sci., **24**(12), 10065. DOI: 10.3390/ijms241210065
29. *Khan F., Anelo O.M., Sadiq Q., Effah W., Price G., Johnson D.L., Ponnusamy S., Grimes B., Morrison M.L., Fowke J.H., Hayes D.N., Narayanan R.* (2023) Racial differences in androgen receptor (AR) and AR splice variants (AR-SVs) expression in treatment-naïve androgen-dependent prostate cancer. Biomedicines, **11**(3), 648. DOI: 10.3390/biomedicines11030648
30. *Bonnal S.C., López-Oreja I., Valcárcel J.* (2020) Roles and mechanisms of alternative splicing in cancer — implications for care. Nat. Rev. Clin. Oncol., **17**(8), 457–474. DOI: 10.1038/s41571-020-0350-x
31. *West S., Kumar S., Batra S.K., Ali H., Ghersi D.* (2019) Uncovering and characterizing splice variants associated with survival in lung cancer patients. PLoS Comput. Biol., **15**(10), e1007469. DOI: 10.1371/journal.pcbi.1007469
32. *Tress M.L., Abascal F., Valencia A.* (2017) Alternative splicing may not be the key to proteome complexity. Trends Biochem. Sci., **42**(2), 98–110. DOI: 10.1016/j.tibs.2016.08.008
33. *Kelemen O., Convertini P., Zhang Z., Wen Y., Shen M., Falaleeva M., Stamm S.* (2013) Function of alternative splicing. Gene, **514**, 1–30. DOI: 10.1016/j.gene.2012.07.083
34. *Yang X., Coulombe-Huntington J., Kang S., Sheynkman G.M., Hao T., Richardson A., Sun S., Yang F., Shen Y.A., Murray R.R., Spirohn K., Begg B.E., Duran-Frigola M., MacWilliams A., Pevzner S.J., Zhong Q., Wanamaker S.A., Tam S., Ghamsari L., Sahni N., Yi S., Rodriguez M.D., Balcha D., Tan G., Costanzo M., Andrews B., Boone C., Zhou X.J., Salehi-Ashtiani K., Charloteaux B., Chen A.A.,*

Calderwood M.A., Aloy P., Roth F.P., Hill D.E., Iakoucheva L.M., Xia Y., Vidal M. (2016) Widespread expansion of protein interaction capabilities by alternative splicing. Cell, **164**(4), 805–817. DOI: 10.1016/j.cell.2016.01.029

35. *Lainez B., Fernandez-Real J.M., Romero X., Esplugues E., Cacete J.D., Ricart W., Engel P.* (2004) Identification and characterization of a novel spliced variant that encodes human soluble tumor necrosis factor receptor 2. Int. Immunol., **16**(1), 169–177. DOI: 10.1093/intimm/dxh014
36. *Kurada B.R.V.S.N., Li L.C., Mulherkar N., Subramanian M., Prasad K.V., Prabhakar B.S.* (2009) MADD, a splice variant of IG20, is indispensable for MAPK activation and protection against apoptosis upon tumor necrosis factor-alpha treatment. J. Biol. Chem., **284**(20), 13533–13541. DOI: 10.1074/jbc.M808554200
37. *Efimova E.V., Al-Zoubi A.M., Martinez O., Kaithamana S., Lu S., Arima T., Prabhakar B.S.* (2004) IG20, in contrast to DENN-SV, (MADD splice variants) suppresses tumor cell survival, and enhances their susceptibility to apoptosis and cancer drugs. Oncogene, **23**(5), 1076–1087. DOI: 10.1038/sj.onc.1207210
38. *Ezkurdia I., Rodriguez J.M., Carrillo-de Santa Pau E., Vázquez J., Valencia A., Tress M.L.* (2015) Most highly expressed protein-coding genes have a single dominant isoform. J. Proteome Res., **14**(4), 1880–1887. DOI: 10.1021/pr501286b
39. *Tress M.L., Abascal F., Valencia A.* (2017) Most alternative isoforms are not functionally important. Trends Biochem. Sci., **42**(6), 408–410. DOI: 10.1016/j.tibs.2017.04.002
40. *Rodriguez J.M., Pozo F., di Domenico T., Vazquez J., Tress M.L.* (2020) An analysis of tissue-specific alternative splicing at the protein level. PLoS Comput. Biol., **16**(10), e1008287. DOI: 10.1371/journal.pcbi.1008287
41. *Pozo F., Martinez-Gomez L., Walsh T.A., Rodriguez J.M., di Domenico T., Abascal F., Vazquez J., Tress M.L.* (2021) Assessing the functional relevance of splice isoforms. NAR Genom. Bioinform., **3**(2), lqab044. DOI: 10.1093/nargab/lqab044
42. *Ponomarenko E.A., Krasnov G.S., Kiseleva O.I., Kryukova P.A., Arzumanian V.A., Dolgalev G.V., Ilgisonis E.V., Lisitsa A.V., Poverennaya E.V.* (2023) Workability of mRNA sequencing for predicting protein abundance. Genes, **14**(11), 2065. DOI: 10.3390/genes14112065
43. *Eraslan B., Wang D., Gusic M., Prokisch H., Hallström B.M., Uhlén M., Asplund A., Pontén F., Wieland T., Hopf T., Hahne H., Kuster B., Gagneur J.* (2019) Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. Mol. Syst. Biol., **15**(2), e8513. DOI: 10.15252/msb.20188513
44. *Liu Y., Gonzàlez-Porta M., Santos S., Brazma A., Marioni J.C., Aebersold R., Venkitaraman A.R., Wickramasinghe V.O.* (2017) Impact of alternative splicing on the human proteome. Cell Rep., **20**(5), 1229–1241. DOI: 10.1016/j.celrep.2017.07.025
45. *Tay A.P., Pang C.N.I., Twine N.A., Hart-Smith G., Harkness L., Kassem M., Wilkins M.R.* (2015) Proteomic validation of transcript isoforms, including those assembled from RNA-seq data. J. Proteome Res., **14**(9), 3541–3554. DOI: 10.1021/pr5011394
46. *Vogel C., Marcotte E.M.* (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet.., **13**(4), 227–232. DOI: 10.1038/nrg3185
47. *Kosti I., Jain N., Aran D., Butte A.J., Sirota M.* (2016) Cross-tissue analysis of gene and protein expression in normal and cancer tissues. Sci. Rep., **6**, 24799. DOI: 10.1038/srep24799
48. *Tran T.T., Bollineni R.C., Strozynski M., Koehler C.J., Thiede B.* (2017) Identification of alternative splice variants using unique tryptic peptide sequences for database searches. J. Proteome Res., **16**(7), 2571–2578. DOI: 10.1021/acs.jproteome.7b00126
49. *Wang X., Codreanu S.G., Wen B., Li K., Chambers M.C., Liebler D.C., Zhang B.* (2018) Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. Mol. Cell. Proteomics, **17**(3), 422–430. DOI: 10.1074/mcp.RA117.000155
50. *Karunratanakul K., Tang H.-Y., Speicher D.W., Chuangsuwanich E., Sriswasdi S.* (2019) Uncovering thousands of new peptides with sequence-mask-search hybrid *de novo* peptide sequencing framework. Mol. Cell. Proteomics, **18**(12), 2478–2491. DOI: 10.1074/mcp.TIR119.001656
51. *Bogdanow B., Zauber H., Selbach M.* (2016) Systematic errors in peptide and protein identification and quantification by modified peptides. Mol. Cell. Proteomics, **15**(8), 2791–2801. DOI: 10.1074/mcp.M115.055103
52. *Kiseleva O., Ponomarenko E., Poverennaya E.* (2020) Empowering shotgun mass spectrometry with 2DE: A HepG2 study. Int. J. Mol. Sci., **21**(11), 3813. DOI: 10.3390/ijms21113813
53. *Poverennaya E.V., Kiseleva O.I., Ponomarenko E.A., Naryzhny S.N., Zgoda V.G., Lisitsa A.V.* (2017) Multiomics study of HepG2 cell line proteome. Biomeditsinskaia Khimiya, **63**(5), 373–378. DOI: 10.18097/PBMC20176305373
54. *Kiseleva O.I., Lisitsa A.V., Poverennaya E.V.* (2018) Proteoforms: Methods of analysis and clinical prospects. Mol. Biol. (Mosk), **52**(3), 394–410. DOI: 10.7868/S0026898418030047
55. *Smith L.M., Agar J.N., Chamot-Rooke J., Danis P.O., Ge Y., Loo J.A., Paša-Tolić L., Tsybin Y.O., Kelleher N.L., Consortium for Top-Down Proteomics* (2021) The human proteoform project: Defining the human proteome. Sci Adv., **7**(46), eabk0734. DOI: 10.1126/sciadv.abk0734
56. *Smith L.M., Kelleher N.L.* (2018) Proteoforms as the next proteomics currency. Science, **359**(6380), 1106–1107. DOI: 10.1126/science.aat1884
57. *Carbonara K., Andonovski M., Coorssen J.R.* (2021) Proteomes are of proteoforms: Embracing the complexity. Proteomes, **9**(3), 38. DOI: 10.3390/proteomes9030038
58. *Forgrave L.M., Wang M., Yang D., de Marco M.L.* (2022) Proteoforms and their expanding role in laboratory medicine. Pract. Lab. Med., **28**, e00260. DOI: 10.1016/j.plabm.2021.e00260
59. *Naryzhny S.* (2016) Towards the full realization of 2DE power. Proteomes, **4**(4), 33. DOI: 10.3390/proteomes4040033
60. *Fornelli L., Toby T.K., Schachner L.F., Doubleday P.F., Srzentić K., deHart C.J., Kelleher N.L.* (2018) Top-down proteomics: Where we are, where we are going? J. Proteomics, **175**, 3–4. DOI: 10.1016/j.jprot.2017.02.002
61. *Chang A., Leutert M., Rodriguez-Mias R.A., Villén J.* (2023) Automated enrichment of phosphotyrosine peptides for high-throughput proteomics. J. Proteome Res., **22**(6), 1868–1880. DOI: 10.1021/acs.jproteome.2c00850
62. *Romashin D., Rusanov A., Arzumanian V., Varshaver A., Poverennaya E., Vakhrushev I., Netrusov A., Luzgina N.* (2024) Exploring the functions of mutant p53 through

TP53 knockout in HaCaT keratinocytes. Curr. Issues Mol. Biol., **46**(2), 1451–1466. DOI: 10.3390/cimb46020094

63. *Poverennaya E.V., Pyatnitskiy M.A., Dolgalev G.V., Arzumanian V.A., Kiseleva O.I., Kurbatov I.Y., Kurbatov L.K., Vakhrushev I.V., Romashin D.D., Kim Y.S., Ponomarenko E.A.* (2023) Exploiting multi-omics profiling and systems biology to investigate functions of TOMM34. Biology, **12**(2), 198. DOI: 10.3390/biology12020198
64. *Rosati D., Palmieri M., Brunelli G., Morrione A., Iannelli F., Frullanti E., Giordano A.* (2024) Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. Comput. Struct. Biotechnol. J., **23**, 1154–1168. DOI: 10.1016/j.csbj.2024.02.018
65. *Li W., Liu C.-C., Kang S., Li J.-R., Tseng Y.-T., Zhou X.J.* (2016) Pushing the annotation of cellular activities to a higher resolution: Predicting functions at the isoform level. Methods, **93**, 110–118. DOI: 10.1016/j.ymeth.2015.07.016
66. *Tseng Y.-T., Li W., Chen C.-H., Zhang S., Chen J.J., Zhou X.J., Liu C.-C.* (2015) IIIDB: A database for isoform-isoform interactions and isoform network modules. BMC Genomics, **16**(Suppl 2), S10. DOI: 10.1186/1471-2164-16-S2-S10
67. *Li W., Kang S., Liu C.-C., Zhang S., Shi Y., Liu Y., Zhou X.J.* (2014) High-resolution functional annotation of human transcriptome: Predicting isoform functions by a novel multiple instance-based label propagation method. Nucleic Acids Res., **42**(6), e39. DOI: 10.1093/nar/gkt1362
68. *Cruz L.M., Trefflich S., Weiss V.A., Castro M.A.A.* (2017) Protein function prediction. Methods Mol. Biol., **1654**, 55–75. DOI: 10.1007/978-1-4939-7231-9_5
69. *Ponomarenko E.A., Poverennaya E.V., Ilgisonis E.V., Pyatnitskiy M.A., Kopylov A.T., Zgoda V.G., Lisitsa A.V., Archakov A.I.* (2016) The size of the human proteome: The width and depth. Int. J. Anal. Chem., **2016**, 7436849. DOI: 10.1155/2016/7436849
70. *Ilgisonis E.V., Pogodin P.V., Kiseleva O.I., Tarbeeva S.N., Ponomarenko E.A.* (2022) Evolution of protein functional annotation: Text mining study. J. Pers. Med., **12**(3), 479. DOI: 10.3390/jpm12030479
71. *Zahn-Zabal M., Lane L.* (2020) What will neXtProt help us achieve in 2020 and beyond? Expert Rev. Proteomics, **17**(2), 95–98. DOI: 10.1080/14789450.2020.1733418
72. *Dolgalev G., Poverennaya E.* (2021) Applications of CRISPR-Cas technologies to proteomics. Genes, **12**(11), 1790. DOI: 10.3390/genes12111790
73. *Liang Q., Wu N., Zaneveld S., Liu H., Fu S., Wang K., Bertrand R., Wang J., Li Y., Chen R.* (2021) Transcript isoforms of Reep6 have distinct functions in the retina. Hum. Mol. Genet., **30**(21), 1907–1918. DOI: 10.1093/hmg/ddab157
74. *Jacobs Catane L., Moshel O., Smith Y., Davidson B., Reich R.* (2021) Splice-variant knock-out of TGFβ receptors perturbates the proteome of ovarian carcinoma cells. Int. J. Mol. Sci., **22**(23), 12647. DOI: 10.3390/ijms222312647
75. *Davies R., Liu L., Taotao S., Tuano N., Chaturvedi R., Huang K.K., Itman C., Mandoli A., Qamra A., Hu C., Powell D., Daly R.J., Tan P., Rosenbluh J.* (2021) CRISPRi enables isoform-specific loss-of-function screens and identification of gastric cancer-specific isoform dependencies. Genome Biol., **22**, 47. DOI: 10.1186/s13059-021-02266-6
76. *Amoasii L., Hildyard J.C.W., Li H., Sanchez-Ortiz E., Mireault A., Caballero D., Harron R., Stathopoulou T.-R., Massey C., Shelton J.M., Bassel-Duby R., Piercy R.J., Olson E.N.* (2018) Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy. Science, **362**(6410), 86–91. DOI: 10.1126/science.aau1549
77. *Long C., Amoasii L., Mireault A.A., McAnally J.R., Li H., Sanchez-Ortiz E., Bhattacharyya S., Shelton J.M., Bassel-Duby R., Olson E.N.* (2016) Postnatal genome editing partially restores dystrophin expression in a mouse model of muscular dystrophy. Science, **351**(6271), 400–403. DOI: 10.1126/science.aad5725
78. *Dours-Zimmermann M.T., Zimmermann D.R.* (2012) A novel strategy for a splice-variant selective gene ablation: The example of the versican V0/V2 knockout. Methods Mol. Biol., **836**, 63–85. DOI: 10.1007/978-1-61779-498-8_5
79. *Dimitrakopoulos G.N., Klapa M.I., Moschonas N.K.* (2022) How far are we from the completion of the human protein interactome reconstruction? Biomolecules, **12**(1), 140. DOI: 10.3390/biom12010140
80. *Huttlin E.L., Bruckner R.J., Navarrete-Perea J., Cannon J.R., Baltier K., Gebreab F., Gygi M.P., Thornock A., Zarraga G., Tam S., Szpyt J., Gassaway B.M., Panov A., Parzen H., Fu S., Golbazi A., Maenpaa E., Stricker K., Guha Thakurta S., Zhang T., Rad R., Pan J., Nusinow D.P., Paulo J.A., Schweppe D.K., Vaites L.P., Harper J.W., Gygi S.P.* (2021) Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell, **184**(11), 3022–3040.e28. DOI: 10.1016/j.cell.2021.04.011
81. *Nesvizhskii A.I.* (2012) Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. Proteomics, **12**(10), 1639–1655. DOI: 10.1002/pmic.201100537
82. *Hermjakob H., Montecchi-Palazzi L., Lewington C., Mudali S., Kerrien S., Orchard S., Vingron M., Roechert B., Roepstorff P., Valencia A., Margalit H., Armstrong J., Bairoch A., Cesareni G., Sherman D., Apweiler R.* (2004) IntAct: An open source molecular interaction database. Nucleic Acids Res., **32**(Database issue), D452–D455. DOI: 10.1093/nar/gkh052
83. *Frommelt F., Fossati A., Uliana F., Wendt F., Xue P., Heusel M., Wollscheid B., Aebersold R., Ciuffa R., Gstaiger M.* (2024) DIP-MS: Ultra-deep interaction proteomics for the deconvolution of protein complexes. Nat. Methods, **21**(4), 635–647. DOI: 10.1038/s41592-024-02211-y
84. *Huttlin E.L., Ting L., Bruckner R.J., Gebreab F., Gygi M.P., Szpyt J., Tam S., Zarraga G., Colby G., Baltier K., Dong R., Guarani V., Vaites L.P., Ordureau A., Rad R., Erickson B.K., Wühr M., Chick J., Zhai B., Kolippakkam D., Mintseris J., Obar R.A., Harris T., Artavanis-Tsakonas S., Sowa M.E., de Camilli P., Paulo J.A., Harper J.W., Gygi S.P.* (2015) The BioPlex network: A systematic exploration of the human interactome. Cell, **162**(2), 425–440. DOI: 10.1016/j.cell.2015.06.043
85. *Poverennaya E., Kiseleva O., Romanova A., Pyatnitskiy M.* (2020) Predicting functions of uncharacterized human proteins: From canonical to proteoforms. Genes, **11**(6), 677. DOI: 10.3390/genes11060677
86. *Kurbatov I., Dolgalev G., Arzumanian V., Kiseleva O., Poverennaya E.* (2023) The knowns and unknowns in protein-metabolite interactions. Int. J. Mol. Sci., **24**(4), 4155. DOI: 10.3390/ijms24044155

87. *Hernández Sánchez L.F., Burger B., Castro Campos R.A., Johansson S., Njølstad P.R., Barsnes H., Vaudel M.* (2023) Extending protein interaction networks using proteoforms and small molecules. Bioinformatics, **39**(10), btad598. DOI: 10.1093/bioinformatics/btad598
88. *Poverennaya E.V., Kiseleva O.I., Ivanov A.S., Ponomarenko E.A.* (2020) Methods of computational interactomics for investigating interactions of human proteoforms. Biochemistry (Moscow), **85**(1), 68–79. DOI: 10.1134/S000629792001006X
89. *Louadi Z., Yuan K., Gress A., Tsoy O., Kalinina O.V., Baumbach J., Kacprowski T., List M.* (2021) DIGGER: Exploring the functional role of alternative splicing in protein interactions. Nucleic Acids Res., **49**(D1), D309–D318. DOI: 10.1093/nar/gkaa768
90. *Gjerga E., Naarmann-de Vries I.S., Dieterich C.* (2023) Characterizing alternative splicing effects on protein interaction networks with LINDA. Bioinformatics, **39**(Suppl 1), i458–i464. DOI: 10.1093/bioinformatics/btad224
91. *Louadi Z., Elkjaer M.L., Klug M., Lio C.T., Fenn A., Illes Z., Bongiovanni D., Baumbach J., Kacprowski T., List M., Tsoy O.* (2021) Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases. Genome Biol., **22**, 327. DOI: 10.1186/s13059-021-02538-1
92. *Yellaboina S., Tasneem A., Zaykin D.V., Raghavachari B., Jothi R.* (2011) DOMINE: A comprehensive collection of known and predicted domain-domain interactions. Nucleic Acids Res., **39**(Database issue), D730–D735. DOI: 10.1093/nar/gkq1229
93. *Mosca R., Céol A., Stein A., Olivella R., Aloy P.* (2014) 3did: A catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res., **42**(Database issue), D374–D379. DOI: 10.1093/nar/gkt887
94. *Kumar M., Gouw M., Michael S., Sámano-Sánchez H., Pancsa R., Glavina J., Diakogianni A., Valverde J.A., Bukirova D., Čalyševa J., Palopoli N., Davey N.E., Chemes L.B., Gibson T.J.* (2020) ELM — the eukaryotic linear motif resource in 2020. Nucleic Acids Res., **48**(D1), D296–D306. DOI: 10.1093/nar/gkz1030
95. *Berman H.M.* (2008) The protein data bank: A historical perspective. Acta Crystallogr. A, **64**(Pt 1), 88–95. DOI: 10.1107/S0108767307035623
96. *Zhang Y., Yao X., Zhou H., Wu X., Tian J., Zeng J., Yan L., Duan C., Liu H., Li H., Chen K., Hu Z., Ye Z., Xu H.* (2022) OncoSplicing: An updated database for clinically relevant alternative splicing in 33 human cancers. Nucleic Acids Res., **50**(D1), D1340–D1347. DOI: 10.1093/nar/gkab851
97. *Li Q., Lai H., Li Y., Chen B., Chen S., Li Y., Huang Z., Meng Z., Wang P., Hu Z., Huang S.* (2021) RJunBase: A database of RNA splice junctions in human normal and cancerous tissues. Nucleic Acids Res., **49**(D1), D201–D211. DOI: 10.1093/nar/gkaa1056
98. *Ling J.P., Wilks C., Charles R., Leavey P.J., Ghosh D., Jiang L., Santiago C.P., Pang B., Venkataraman A., Clark B.S., Nellore A., Langmead B., Blackshaw S.* (2020) ASCOT identifies key regulators of neuronal subtype-specific splicing. Nat. Commun., **11**, 137. DOI: 10.1038/s41467-019-14020-5
99. *Tian J., Wang Z., Mei S., Yang N., Yang Y., Ke J., Zhu Y., Gong Y., Zou D., Peng X., Wang X., Wan H., Zhong R., Chang J., Gong J., Han L., Miao X.* (2019) CancerSplicingQTL: A database for genome-wide identification of splicing QTLs in human cancer. Nucleic Acids Res., **47**(D1), D909–D916. DOI: 10.1093/nar/gky954
100. *UniProt Consortium* (2022) UniProt: The universal protein knowledgebase in 2023. Nucleic Acids Res., **51**(D1), D523–D531. DOI: 10.1093/nar/gkac1052
101. *Lane L., Argoud-Puy G., Britan A., Cusin I., Duek P.D., Evalet O., Gateau A., Gaudet P., Gleizes A., Masselot A., Zwahlen C., Bairoch A.* (2012) NeXtProt: A knowledge platform for human proteins. Nucleic Acids Res., **40**(Database issue), D76–D83. DOI: 10.1093/nar/gkr1179
102. *O'Leary N.A., Wright M.W., Brister J.R., Ciufo S., Haddad D., McVeigh R., Rajput B., Robbertse B., Smith-White B., Ako-Adjei D., Astashyn A., Badretdin A., Bao Y., Blinkova O., Brover V., Chetvernin V., Choi J., Cox E., Ermolaeva O., Farrell C.M., Goldfarb T., Gupta T., Haft D., Hatcher E., Hlavina W., Joardar V.S., Kodali V.K., Li W., Maglott D., Masterson P., McGarvey K.M., Murphy M.R., O'Neill K., Pujar S., Rangwala S.H., Rausch D., Riddick L.D., Schoch C., Shkeda A., Storz S.S., Sun H., Thibaud-Nissen F., Tolstoy I., Tully R.E., Vatsan A.R., Wallin C., Webb D., Wu W., Landrum M.J., Kimchi A., Tatusova T., di Cuccio M., Kitts P., Murphy T.D., Pruitt K.D.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res., **44**(D1), D733–D745. DOI: 10.1093/nar/gkv1189
103. *Frankish A., Diekhans M., Jungreis I., Lagarde J., Loveland J.E., Mudge J.M., Sisu C., Wright J.C., Armstrong J., Barnes I., Berry A., Bignell A., Boix C., Carbonell Sala S., Cunningham F., di Domenico T., Donaldson S., Fiddes I.T., García Girón C., Gonzalez J.M., Grego T., Hardy M., Hourlier T., Howe K.L., Hunt T., Izuogu O.G., Johnson R., Martin F.J., Martínez L., Mohanan S., Muir P., Navarro F.C.P., Parker A., Pei B., Pozo F., Riera F.C., Ruffier M., Schmitt B.M., Stapleton E,. Suner M.M., Sycheva I., Uszczynska-Ratajczak B., Wolf M.Y., Xu J., Yang Y.T., Yates A., Zerbino D., Zhang Y., Choudhary J.S., Gerstein M., Guigó R., Hubbard T.J.P., Kellis M., Paten B., Tress M.L., Flicek P.* (2021) GENCODE 2021. Nucleic Acids Res., **49**(D1), D916–D923. DOI: 10.1093/nar/gkaa1087
104. *Morales J., Pujar S., Loveland J.E., Astashyn A., Bennett R., Berry A., Cox E., Davidson C., Ermolaeva O., Farrell C.M., Fatima R., Gil L., Goldfarb T., Gonzalez J.M., Haddad D., Hardy M., Hunt T., Jackson J., Joardar V.S., Kay M., Kodali V.K., McGarvey K.M., McMahon A., Mudge J.M., Murphy D.N., Murphy M.R., Rajput B., Rangwala S.H., Riddick L.D., Thibaud-Nissen F., Threadgold G, Vatsan A.R., Wallin C., Webb D., Flicek P., Birney E., Pruitt K.D., Frankish A., Cunningham F., Murphy T.D.* (2022) A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature, **604**(7905), 310–315. DOI: 10.1038/s41586-022-04558-8
105. *Pertea M., Shumate A., Pertea G., Varabyou A., Breitwieser F.P., Chang Y.-C., Madugundu A.K., Pandey A., Salzberg S.L.* (2018) CHESS: A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol., **19**, 208. DOI: 10.1186/s13059-018-1590-2
106. *Howe K.L., Achuthan P., Allen J., Allen J., Alvarez-Jarreta J., Amode M.R., Armean I.M., Azov A.G., Bennett R., Bhai J., Billis K., Boddu S., Charkhchi M., Cummins C., da Rin Fioretto L., Davidson C., Dodiya K., El Houdaigui B., Fatima R., Gall A., Garcia Giron C., Grego T., Guijarro-Clarke C., Haggerty L., Hemrom A., Hourlier T., Izuogu O.G., Juettemann T., Kaikala V., Kay M., Lavidas I.,*

Le T., Lemos D., Gonzalez Martinez J., Marugán J.C., Maurel T., McMahon A.C., Mohanan S., Moore B., Muffato M., Oheh D.N., Paraschas D., Parker A., Parton A., Prosovetskaia I., Sakthivel M.P., Salam A.I.A., Schmitt B.M., Schuilenburg H., Sheppard D., Steed E., Szpak M., Szuba M., Taylor K., Thormann A., Threadgold G., Walts B., Winterbottom A., Chakiachvili M., Chaubal A., de Silva N., Flint B., Frankish A., Hunt S.E., Ilsley G.R., Langridge N., Loveland J.E., Martin F.J., Mudge J.M., Morales J., Perry E., Ruffier M., Tate J., Thybert D., Trevanion S.J., Cunningham F., Yates A.D., Zerbino D.R., Flicek P. (2021) Ensembl 2021. Nucleic Acids Res., **49**(D1), D884–D891. DOI: 10.1093/nar/gkaa942

107. *Lindblad-Toh K., Garber M., Zuk O., Lin M.F., Parker B.J., Washietl S., Kheradpour P., Ernst J., Jordan G., Mauceli E., Ward L.D., Lowe C.B., Holloway A.K., Clamp M., Gnerre S., Alföldi J., Beal K., Chang J., Clawson H., Cuff J., di Palma F., Fitzgerald S., Flicek P., Guttman M., Hubisz M.J., Jaffe D.B., Jungreis I., Kent W.J., Kostka D., Lara M., Martins A.L., Massingham T., Moltke I., Raney B.J., Rasmussen M.D., Robinson J., Stark A., Vilella A.J., Wen J., Xie X., Zody M.C., Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baldwin J., Bloom T., Chin C.W., Heiman D., Nicol R., Nusbaum C., Young S., Wilkinson J., Worley K.C., Kovar C.L., Muzny D.M., Gibbs R.A., Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Cree A., Dihn H.H., Fowler G., Jhangiani S., Joshi V., Lee S., Lewis L.R., Nazareth L.V., Okwuonu G., Santibanez J., Warren W.C., Mardis E.R., Weinstock G.M., Wilson R.K., Genome Institute at Washington University, Delehaunty K., Dooling D., Fronik C., Fulton L., Fulton B., Graves T., Minx P., Sodergren E., Birney E., Margulies E.H., Herrero J., Green E.D., Haussler D., Siepel A., Goldman N., Pollard K.S., Pedersen J.S., Lander E.S., Kellis M.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature, **478**(7370), 476–482. DOI: 10.1038/nature10530
108. *Sommer M.J., Cha S., Varabyou A., Rincon N., Park S., Minkin I., Pertea M., Steinegger M., Salzberg S.L.* (2022) Structure-guided isoform identification for the human transcriptome. eLife, **11**, e82556. DOI: 10.7554/eLife.82556
109. *Palazzo A.F., Lee E.S.* (2015) Non-coding RNA: What is functional and what is junk? Front. Genetics, **6**, 2. DOI: 10.3389/fgene.2015.00002
110. *Ponting C.P., Haerty W.* (2022) Genome-wide analysis of human long noncoding RNAs: A provocative review. Annu. Rev. Genomics Hum. Genet., **23**, 153–172. DOI: 10.1146/annurev-genom-112921-123710
111. *Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G.* (2000) Gene ontology: Tool for the unification of biology. The gene ontology consortium. Nat. Genet., **25**(1), 25–29. DOI: 10.1038/75556
112. *Qiu S., Yu G., Lu X., Domeniconi C., Guo M.* (2022) Isoform function prediction by Gene Ontology embedding. Bioinformatics, **38**(19), 4581–4588. DOI: 10.1093/bioinformatics/btac576
113. *Eksi R., Li H.-D., Menon R., Wen Y., Omenn G.S., Kretzler M., Guan Y.* (2013) Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. PLoS Comput. Biol., **9**(11), e1003314. DOI: 10.1371/journal.pcbi.1003314
114. *Luo T., Zhang W., Qiu S., Yang Y., Yi D., Wang G., Ye J., Wang J.* (2017) Functional annotation of human protein coding isoforms via non-convex multi-instance learning. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA. pp. 345–354. DOI: 10.1145/3097983.3097984
115. *Li H.-D., Yang C., Zhang Z., Yang M., Wu F.-X., Omenn G.S., Wang J.* (2021) IsoResolve: Predicting splice isoform functions by integrating gene and isoform-level features with domain adaptation. Bioinformatics, **37**(4), 522–530. DOI: 10.1093/bioinformatics/btaa829
116. *Shaw D., Chen H., Jiang T.* (2019) DeepIsoFun: A deep domain adaptation approach to predict isoform functions. Bioinformatics, **35**(15), 2535–2544. DOI: 10.1093/bioinformatics/bty1017
117. *Chen H., Shaw D., Zeng J., Bu D., Jiang T.* (2019) DIFFUSE: Predicting isoform functions from sequences and expression profiles via deep learning. Bioinformatics, **35**(14), i284–i294. DOI: 10.1093/bioinformatics/btz367
118. *Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., Žídek A., Potapenko A., Bridgland A., Meyer C., Kohl S.A.A., Ballard A.J., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Back T., Petersen S., Reiman D., Clancy E., Zielinski M., Steinegger M., Pacholska M., Berghammer T., Bodenstein S., Silver D., Vinyals O., Senior A.W., Kavukcuoglu K., Kohli P., Hassabis D.* (2021) Highly accurate protein structure prediction with AlphaFold. Nature, **596**(7873), 583–589. DOI: 10.1038/s41586-021-03819-2
119. *Tunyasuvunakool K., Adler J., Wu Z., Green T., Zielinski M., Žídek A., Bridgland A., Cowie A., Meyer C., Laydon A., Velankar S., Kleywegt G.J., Bateman A., Evans R., Pritzel A., Figurnov M., Ronneberger O., Bates R., Kohl S.A.A., Potapenko A., Ballard A.J., Romera-Paredes B., Nikolov S., Jain R., Clancy E., Reiman D., Petersen S., Senior A.W., Kavukcuoglu K., Birney E., Kohli P., Jumper J., Hassabis D.* (2021) Highly accurate protein structure prediction for the human proteome. Nature, **596**(7873), 590–596. DOI: 10.1038/s41586-021-03828-1
120. *Deiana A., Forcelloni S., Porrello A., Giansanti A.* (2019) Intrinsically disordered proteins and structured proteins with intrinsically disordered regions have different functional roles in the cell. PloS ONE, **14**(8), e0217889. DOI: 10.1371/journal.pone.0217889
121. *Mirdita M., Schütze K., Moriwaki Y., Heo L., Ovchinnikov S., Steinegger M.* (2022) ColabFold: Making protein folding accessible to all. Nat. Methods, **19**(6), 679–682. DOI: 10.1038/s41592-022-01488-1
122. *Chang E., Fu C., Coon S.L., Alon S., Bozinoski M., Breymaier M., Bustos D.M., Clokie S.J., Gothilf Y., Esnault C., Michael Iuvone P., Mason C.E., Ochocinska M.J., Tovin A., Wang C., Xu P., Zhu J., Dale R., Klein D.C.* (2020) Resource: A multi-species multi-timepoint transcriptome database and webpage for the pineal gland and retina. J. Pineal Res., **69**(3), e12673. DOI: 10.1111/jpi.12673

Received: 27. 04. 2024.
Revised: 21. 06. 2024.
Accepted: 18. 07. 2024.

IN SILICO И *IN CELLULO* ПОДХОДЫ ДЛЯ ФУНКЦИОНАЛЬНОЙ АННОТАЦИИ СПЛАЙС-ФОРМ БЕЛКОВ ЧЕЛОВЕКА

О.И. Киселева, В.А. Арзуманян, И.Ю. Курбатов, Е.В. Поверенная*

Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича,
119121, Москва, ул. Погодинская, 10; *эл. почта: k.poverennaya@gmail.com

Изящество механизмов сплайсинга пре-мРНК не перестаёт интересовать ученых даже спустя более полувека с момента открытия того факта, что кодирующие участки в генах прерываются некодирующими. Превалирующее большинство генов человека имеют несколько вариантов мРНК, которые, в свою очередь, кодируют структурно и функционально разные варианты белков — в тканезависимой манере и с привязкой к конкретным этапам развития организма. Нарушение паттернов сплайсинга смещает баланс функционально различающихся белков в живой системе, искажает нормальные молекулярные пути и может спровоцировать возникновение и развитие патологий. За последние два десятилетия выполнено множество исследований в различных областях наук о жизни для более глубокого понимания механизмов сплайсинга и степени его влияния на функционирование живых систем. Целью данного обзора было суммирование экспериментальных и вычислительных подходов, используемых для выяснения функций сплайс-опосредованных белковых продуктов одного гена: на основе собственного опыта, накопленного в лаборатории интерактомики протеоформ Института биомедицинской химии, и лучших мировых практик.

Полный текст статьи на русском языке доступен на сайте журнала (http://pbmc.ibmc.msk.ru).

Ключевые слова: функциональная аннотация; альтернативный сплайсинг; сплайс-варианты; протеоформы; гетерогенность протеома; мультиомные исследования

Финансирование. Данное исследование поддержано грантом РНФ № 21-74-10061.

Поступила в редакцию: 27.04.2024; после доработки: 21.06.2024; принята к печати: 18.07.2024.