Validating amino acid variants in proteogenomics using sequence coverage by multiple reads

L.I. Levitsky [1], K.G. Kuznetsova [2], A.A. Kliuchnikova [2,3], I.Y. Ilina [2], A.O. Goncharov [2,3], A.A. Lobas [1], M.V. Ivanov [1], V.N. Lazarev [2], R.H. Ziganshin [4], M.V. Gorshkov [1], S.A. Moshkovskii [2,3*]

[1] V.L. Talrose Institute for Energy Problems of Chemical Physics, N.N. Semenov Federal Research Center for Chemical Physics, Russian Academy of Sciences, 38, bld. 1, Leninsky Prospect, Moscow, 119334, Russia

[2] Federal Research and Clinical Center of Physical-Chemical Medicine, 1a, Malaya Pirogovskaya, Moscow, 119435, Russia

[3] Pirogov Russian National Research Medical University, 1, Ostrovityanova, Moscow, 117997, Russia

[4] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, 16/10, Miklukho-Maklaya, Moscow, 117997, Russia

* To whom correspondence should be addressed.

E-mail: smosh@mail.ru (Moshkovskii S.A.)

Abstract

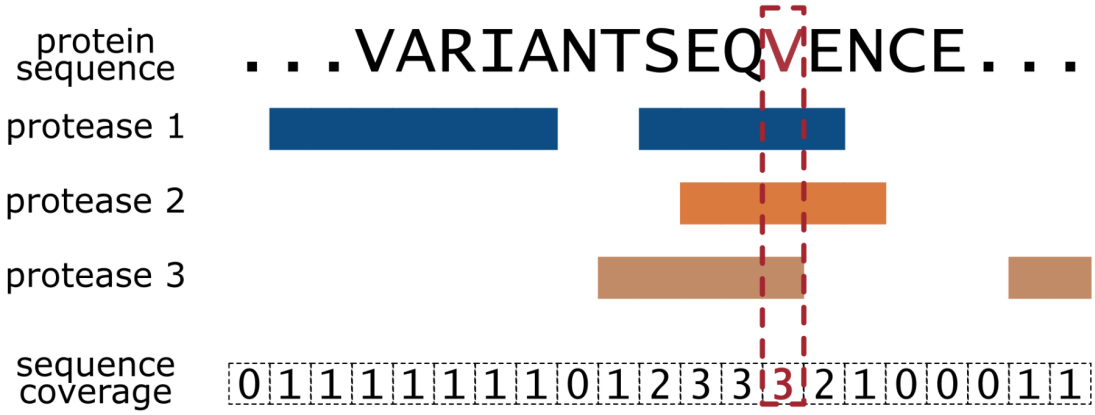
Mass spectrometry-based proteome analysis usually implies matching mass spectra of proteolytic peptides to amino acid sequences predicted from nucleic acid sequences. At the same time, due to the stochastic nature of the method when it comes to proteome-wide analysis, in which only a fraction of peptides are selected for sequencing, the completeness of protein sequence identification is undermined. Likewise, the reliability of peptide variant identification in proteogenomic studies is suffering. We propose a way to interpret shotgun proteomics results, specifically in data-dependent acquisition mode, as protein sequence coverage by multiple reads, just as it is done in the field of nucleic acid sequencing for the calling of single nucleotide variants. Multiple reads for each position in a sequence could be provided by overlapping distinct peptides, thus, confirming the presence of certain amino acid residues in the overlapping stretch with much lower false discovery rate than conventional 1%. The source of overlapping distinct peptides are, first, miscleaved tryptic peptides in combination with their properly cleaved counterparts, and, second, peptides generated by several proteases with different specificities after the same specimen is subject to parallel digestion and analyzed separately. We illustrate this approach

using publicly available multiprotease proteomic datasets and our own data generated for HEK-293 cell line digests obtained using trypsin, LysC and GluC proteases. From 5000 to 8000 protein groups are identified for each digest corresponding to up to 30% of the whole proteome coverage. Most of this coverage was provided by a single read, while up to 7% of the observed protein sequences were covered two-fold and more. The proteogenomic analysis of HEK-293 cell line revealed 36 peptide variants associated with SNP, seven of which were supported by multiple reads. The efficiency of the multiple reads approach depends strongly on the depth of proteome analysis, the digesting features such as the level of miscleavages, and will increase with the number of different proteases used in parallel proteome digestion.

Keywords

Proteogenomics, single nucleotide variant, single amino acid variant, missense mutation, shotgun proteomics, data dependent acquisition, SNP calling, false discovery rate, protease

Graphical abstract



1. Introduction

A majority of mass spectrometric studies of proteomes currently use the bottom-up, also called shotgun, technique [1]. The key part of the bottom-up workflow is protein digestion by a protease to produce a peptide mixture which is subject to the follow-up liquid chromatography/tandem mass spectrometry (LC-MS/MS) analysis. Peptides are much simpler macromolecules for identification from both LC and MS/MS viewpoints compared with proteins, yet, this approach complicates subsequent protein identification based on the peptide-spectrum

matches (PSM), posing the so-called protein inference problem [2,3]. The most effective way to attribute a measured tandem mass spectrum is by means of a genomic database, predicting a limited set of peptides from the theoretical proteome. An algorithm will select the best match to any eligible mass spectrum. To filter out unreliable predictions, the false discovery rate (FDR) is then calculated based typically on the so-called target-decoy approach (TDA) [4]. The PSMs passing the FDR threshold of, typically, 1%, constitute the list of identified peptides used further for protein inference.

Although this practice has been challenged by bioinformaticians [5], protein identifications are generally considered much more reliable if they are supported by multiple distinct peptide identifications. Otherwise, if one reports a new finding based on a single peptide identified by one or more PSMs, some additional validation is usually required. Unfortunately, this is a typical story in proteogenomics, in which important single peptide-based reports include variants predicted from genetic polymorphisms [6], alternative splicing events [7], RNA editing [8], or so-called missing proteins [9].

The above cases are typically resolved by targeted proteomic methods, such as multiple/selected reaction monitoring (MRM/SRM) or parallel reaction monitoring (PRM) [10]. For most applications, where amino acid variants to be detected originate from actionable or neoantigenic mutations, one can use targeted methods directly, omitting the shotgun proteome analysis step; however, this approach requires a large cohort of synthetic peptide standards and a lot of instrumentation time, further limiting its utility for wider clinical applications.

Several orthogonal methods have been introduced to increase reliability of peptide identifications by random sampling in shotgun proteomics. Addition of specific peptide features is used for scoring PSMs, such as a chromatographic retention time [11], intensity pattern in fragmentation spectra [12–14], etc. Rapidly accumulating data from large-scale studies made it possible to explore deep learning approaches for efficient prediction of these features to further enhance identification of PSMs. At the same time, there are still no standards in the field for validation of single peptide-based findings and further improvements to this end may be suggested.

For applications of shotgun proteomics such as proteogenomics, where single amino acid polymorphisms and, correspondingly, single peptide identification is often the only available option, it is desirable to improve reliability of these identifications within the scope of the method, ideally, avoiding additional procedures, such as targeted mass-spectrometry. Elaborating further on the similarities between shotgun proteomics and next generation sequencing (NGS) of nucleic acids, which is also a shotgun approach, note that single nucleotide polymorphisms are confirmed

by multiple overlapping reads in NGS [15]. The idea of this work is to use the same strategy for shotgun proteomics data. Overlapping reads in NGS are intrinsic for many methods of sequencing [16]. In a similar fashion, one can define the 'read' in shotgun proteomics. It is, naturally, a single PSM, which is a hypothesis that a given mass spectrum is produced by a given peptide compound listed in a genomic database, with some measure of reliability [17]. PSMs from different mass spectra and with different retention times and other measurable parameters are formally independent. Thus, if two or more different PSMs map to overlapping parts of a protein sequence, the corresponding identification of the overlap becomes increasingly more reliable. A common source of overlapping PSMs in shotgun proteomics is identification of a properly cleaved tryptic peptide and its miscleaved counterpart. If, for example, a sequence variant of interest would fall into the overlapping part, we must think that it is identified more reliably than from a single PSM. A hunt for overlapping reads in shotgun proteomics is the rationale behind the approach explored in this work.

The embodiment of this idea is focused primarily on the data dependent acquisition (DDA) mode as the concept of peptide/PSM FDR for data independent acquisition (DIA) has yet to be established and such notations as the group-specific FDR [18] are not available. At the same time, for many DIA applications, DDA runs are also made to attribute identities to peptides by the conventional target decoy-method and then DIA provides better quantification of proteins being not focused on single amino acid variants [6].

Controlling the efficiency of trypsinolysis to intentionally produce overlapping cleaved and miscleaved peptides does not seem feasible due to different kinetics of the cleavage process for different protein sequences [19]. An easier way to obtain overlapping reads is the use of multiple proteases known in the field [20–23]. In this work, we analyzed publicly available datasets and generated our own experimental data for the model HEK-293 cell line using multiple proteases to conceptualize the approach of utilizing the sequence coverage by multiple reads for validation of single amino acid variants.

2. Experimental procedures

2.1. Cell culture

Cell culture was managed in accordance with our previous work [24]. Briefly, HEK-293 cells were obtained from ATCC (accession number CRL-1573; Manassas, VA). The 40th passage of HEK-293 was used for proteome analysis.

Cryogenically preserved cells were thawed and expanded in culture medium (DMEM) supplemented with 10% (w/v) fetal bovine serum (FBS) and 100 units/mL gentamicin (all from

Gibco, Thermo Fisher Scientific, Bremen, Germany) in a humidified CO2 incubator under standard conditions (5% CO_2, 37 °C). The medium was exchanged every 2 days. To prepare cell samples for protein extraction, the cells were detached with 0.05% Trypsin-EDTA solution (PanEko, Moscow, Russia), washed 3 times with PBS, and counted. Aliquots of the resulting cell suspension were centrifuged, the supernatant was removed, and cells were frozen in liquid nitrogen. The cell pellets were kept frozen in liquid nitrogen vapor until use.

2.2. Cell lysis and protein digestion

Cell pellets of one million cells each were resuspended in lysis buffer containing 4% SDS in 100 mM TEABC, pH 8.5, incubated for 5 minutes and subjected to sonication by Qsonica Q55 ultrasonic homogenizer (Qsonica, USA) at 70% amplitude using 10 series of 10 one-second-duration impulses. After that, the samples were incubated for 10 minutes at 85 °C and centrifuge for 10 minutes at 16000 × g. Later on, protein disulfide bonds were reduced by addition of dithiothreitol (DTT) up to 5 mM and incubation for 30 minutes at 56 °C and the cysteine residues were alkylated with chloroacetamide (CAM) in the final concentration of 10 mM for 15 minutes at room temperature at dark.

At the next step, the proteins were precipitated with cold acetone. First, we precipitated the proteins from 10 µL of the lysate to measure protein concentration after the precipitation. Next, the volume containing 30 µg of total protein was taken from each sample for the future analysis.

For precipitation, 4 volumes of cold (-20 °C) acetone were added to the samples. The samples were vortexed and left for 120 minutes at -20 °C to let the proteins thoroughly precipitate. After the incubation period, the samples were centrifuged for 10 minutes at 13,000-15,000 × g at 4 °C followed by rinsing the pellet with cold acetone twice without mixing. Then the acetone was carefully removed and the tubes were left with the lids open for 30 minutes to let the pellets dry up.

For digestion, 30 µL of each protease, namely trypsin (Promega Gold), LysC and GluC (all from Promega, USA) at a concentration of 0.02 µg/µl in 50 mM TEABC were added straight to the pellets. The final w/w proportion of each protease to total protein was 1:50. Then the samples were incubated overnight at 37 °C. To stop the reaction, trifluoroacetic acid (TFA) up to 1% (v/v) was added to each tube.

2.3. Peptide desalting and clean-up

For peptide desalting and clean-up, in-house made stage tips containing SDB-RPS membrane (Empore-3M, CDS Analytical, USA) were used. The tips were prepared as described earlier [25] with the use of 3 pieces of membrane in each tip. The samples were loaded into the tips and the tips were centrifuged at 1200 rpm (about 70 × g) in the BioSan Multi-spin MSV-6000

centrifuge until the solution had passed through the membrane. At the next step, washing with 100 µL 0.2% TFA was performed at the same speed. The peptides were eluted by passing 60 µL of 70% acetonitrile (ACN) with 5% ammonia through the tips into the clean tubes at the speed as low as 1000 rpm (about 50 × g) in the same centrifuge. The peptide samples were dried up in the vacuum concentrator (Labconco, USA).

2.4. Liquid chromatography and mass spectrometry

For the LC-MS analysis, the samples were reconstituted in 0.1% TFA and loaded onto an Acclaim PepMap 100 C18 (100 mm × 2 cm) trap column (Thermo Fisher Scientific, USA) in the loading mobile phase (2% ACN, 98% H_2O, 0.1% TFA) at 10 mL/min flow and separated at 40 °C on a 75 mm × 50 cm Acclaim PepMap 100 C18 LC column (Thermo Fisher Scientific, USA) with particle size 2 mm. Reverse-phase chromatography was performed with an Ultimate 3000 Nano LC System (Thermo Fisher Scientific), which was coupled to the Orbitrap QExactive HF mass spectrometer via a nano electrospray source (Thermo Fisher Scientific). Water containing 0.1% (v/v) formic acid (FA) was used as a mobile phase A and ACN containing 0.1% FA (v/v), 20% (v/v) H_2O as a mobile phase B. Peptides were eluted from the trap column with a linear gradient: 3–35% solution B for 105 min; 35–55% B for 18 min, 55–99% B for 0.1 min, 99% B during 10 min, 99–2% B for 0.1 min at a flow rate of 300 nL/min. After each gradient, the column was re-equilibrated with the phase A for 10 min. MS data was collected in DDA mode (TopN = 15). MS1 parameters were as follows: 120K resolution, 350–1400 scan range, max injection time 50 msec, AGC target 3×10^6. Ions were isolated with 1.2 m/z window, preferred peptide match and isotope exclusion. Dynamic exclusion was set to 30 s. MS2 fragmentation was carried out at 15K resolution with HCD collision energy 28, max injection time – 80 msec, AGC target – 1×10^5. Other settings were as follows: charge exclusion - unassigned, 1, 6–8, >8.

2.5. MS/MS data processing and coverage calculation

All raw data were converted to mzML format using Proteowizard's MSConvert [26] and searched using IdentiPy v0.3.4 [27] with search parameters specified below. Search results were post-processed with Scavager v0.2.9 [28], used both to increase sensitivity at 1% FDR and to perform the merging of search results across replicates and fractions as needed. Additional reanalysis was performed by merging results across different proteases to produce summary statistics, such as the total number of peptides and proteins identified at 1% FDR, also using Scavager. Proteome coverage calculations were performed using in-house Python scripts.

Human melanoma cell line 82 data [29] were searched against the Swissprot human database. Precursor and fragment ion mass tolerances were optimized using the auto-tune feature of IdentiPy; carbamidomethylation of cysteine was set as a fixed modification, and

oxidation of methionine was allowed as a variable modification. Tryptic digestion rule was specified, with full specificity and up to two miscleavages allowed. Other data sets mentioned in Table 1 were processed using the same settings, except that cleavage on arginine and lysine was specified for samples where LysC was added to trypsin for digestion. Human brain data [30] were processed as above, except fragment ion mass tolerance was set at 0.01 Da and precursor ion tolerance was set at 10 ppm (no auto-tune). For LysC digestion, the cleavage rule was set to LysC and up to two miscleavages were allowed. Search parameters for Confetti data [20] were the same, except the fixed modification on cysteine was set to N-ethylmaleimide (this decision was motivated by previous analysis of modifications with AA_stat [31,32]. Also, for each protease, its corresponding rule was specified, and the number of allowed miscleavages was set as follows: two for ArgC, four for AspN, six for chymotrypsin, four for GluC, two for LysC, two for trypsin. For elastase, non-specific cleavage was used. Own data from the HEK-293 cell line were searched against the proteogenomic database described in the previous work [24] for the analysis of SAPs and against the RefSeq database for the analysis of alternative splicing. Search parameters were the same as above, with fixed cysteine carbamidomethylation and variable methionine oxidation, four miscleavages allowed for GluC and two for LysC and trypsin.

HEK-293 proteomics data have been deposited to the ProteomeXchange Consortium via the MassIVE (https://massive.ucsd.edu) partner repository with the dataset identifier PXD030226.

2.6. MS1 only search to identity variants in HEK-293 cells

HEK-293 data were additionally analyzed using DirectMS1 (ms1searchpy v. 2.1.7) approach [33]. Search parameters were default except missed cleavages were set to 0, 0 and 2 for trypsin, LysC and GluC, respectively. For the analysis, only variant proteins were considered; of those, we only kept the proteins which contain at least one theoretical variant peptide in the 7-30 length range for all three proteases. Thus, only 308 variant proteins were considered in DirectMS1 results.

3. Results and discussion

3.1. Bad and good of the protein miscleavage

The most widely used protease for proteome-wide digestion in shotgun proteomics, and obviously the best in terms of specificity, is trypsin. As all enzymes, this protease has its difficulties even in nominally specific sites, which are dictated by the surrounding protein sequence context [19,34]. In the field, there is a trend to prevent the protein miscleavage by trypsin, as miscleaved peptides are less suitable for mass-spectrometry pipelines due to being longer; also, their

presence inflates the search space, increasing FDR [35]. Lys-C, a protease with a similar specificity, is added to trypsin for better digestion of problematic cleavage sites [36].

When we aimed to identify and validate tryptic peptides which contained single amino acid variants originating from genomic variants [24,29] or RNA editing [8], some of those variants were found in peptides with missed cleavages. Manual inspection of the corresponding mass spectra, which was conventionally used to confirm true identifications, led us to the observation that peptide hits represented by miscleaved peptides alone were enriched with false positive results. In contrast, if the sequence variants were represented by both properly cleaved and miscleaved peptides, these results were always true [8]. Indeed, two different peptides containing the same residue of interest may be considered as independent events with multiplied probabilities. Thus, if a miscleaved peptide was identified along with one or more of its cleaved counterparts, the overlapping sequence was validated with higher probability than for any peptide alone. We then used the target-decoy analysis to illustrate this concept in selected available datasets for shotgun proteomes produced in DDA mode.

Group-specific FDR values were calculated separately, for groups of (i) properly cleaved tryptic peptides; (ii) peptides containing a single miscleavage site and lacking a confirmation with an overlapping properly cleaved peptide; and (iii) peptides containing a single miscleavage site and also confirmed by at least one overlapping properly cleaved peptide, in most cases forming pairs (Table 1). Corresponding decoy subsets were used for FDR calculations. Other groups, such as peptides containing more than one miscleavage, were also considered but not reported here as the peptides were found in low numbers. As expected, FDR values for the group of miscleaved/cleaved pairs were approximately an order of magnitude lower than for properly cleaved peptides without additional confirmation by overlapping peptides with miscleavage, with zero FDR in some datasets (Table 1). In other words, miscleaved peptides, in some cases, provide more reliable identification for the strands of amino acid sequence which have a double coverage. Interestingly, in the dataset of murine microglia [37], the proportion of miscleaved peptides is higher than in other exemplary datasets (14% of pairs vs. 5-8%, correspondingly), more likely, due to some sample preparation features. At the same time, the double sequence coverage, according to our teaching, in this dataset is also higher. As mentioned above, it is generally difficult to control trypsinolysis to guarantee efficiency. An easier and better established way to produce overlapping peptides proteome-wide is multiprotease cleavage [20].

Table 1. Group-specific false discovery rates calculated for properly cleaved tryptic peptides, miscleaved peptides with and without confirmation by their cleaved counterparts. Selected sets of shotgun proteomic data were publicly available and generated by high-resolution LC-MS/MS using Orbitrap detectors in DDA mode.

Source of biomaterial, digestion, PRIDE accession, reference	Properly cleaved peptides		Single miscleavages: Peptides with 1 miscleavage site, not accompanied by cleaved peptides		Pairs of miscleaved/cleaved: Peptides with 1 miscleavage site, accompanied by at least 1 cleaved counterparts (coverage 2X)	
	Amount	FDR, % (standard deviation*)	Amount	FDR, % (standard deviation)	Amount	FDR, % (standard deviation)
Human melanoma cell line 82, trypsin PXD007662 [29]	31293	1.02 (0.20)	3799	1.13 (0.63)	1658	0.12 (0.43)
Human cancer cell lines, trypsin/LysC, PXD007686 [38]	42733	0.92 (0.17)	5027	2.08 (0.75)	3617	0.06 (0.2)
Murine microglia culture, trypsin/LysC, PXD14466 [37]	41685	0.87 (0.16)	6474	2.45 (0.7)	5680	0.00 (0.09)
Fruit fly brain, trypsin, PXD009590 [39]	20582	0.94 (0.35)	2840	1.33 (1.16)	1412	0.00 (0.6)
Murine brain, fractionated, trypsin/LysC, PXD001250 [40]	77363	1.00 (0.13)	7817	1.53 (0.49)	4903	0.04 (0.14)

* The standard deviation for FDR is calculated as described earlier [41].

3.2. Sequence coverage in multiprotease proteomic datasets

The concept of sequence coverage by multiple reads as defined here relates mostly to more reliable identification of single amino acid or peptide events in proteome, such as amino acid sequence variants, splice junctions etc. In this case, the coverage can be defined as easy as the number of distinct peptides spanning the site of interest, with the majority of conventionally identified sites having a coverage of 1X. At the same time, to compare between datasets and methods for retrieval of the coverage it is convenient to estimate its fold proteome-wide. This task is more complicated because a whole proteome, similarly to the genome, contains many degenerated sequences which may corrupt the coverage estimation. Indeed, there is no linear “proteome” onto which all peptides can be mapped; instead, there are concurrent isoforms, splice products, etc. One way of approaching this issue is to map all peptides onto the genome directly; however, in this work we take a simpler approach and consider a subset of identified peptides that can be mapped to proteins without ambiguity.

To calculate the overall proteome coverage, we started with the output of Scavager post-search validation tool [28] (Fig.1). Scavager produces several tables with identified PSMs, peptides, proteins and protein groups. Each of the tables is filtered to the desired FDR level (1% in our case). Each of the protein groups has a “leader” protein, and it is known that no proteins in any group except the leader have any unique peptides. We considered all identified peptides, and considered which proteins they come from. Then, we only kept the peptides which have exactly one protein group leader in their list of proteins; these peptides could be unambiguously mapped to this leader protein. The leader proteins from this peptide subset formed the simplified, “linear” model of the proteome, for which the average coverage can be easily calculated. This approach to define the linear proteome was similar to the method of genomic SNP calling where the sequences shared between different genes are excluded from analysis to provide so-called "mappability" [42].

To calculate the proteome coverage, each of the leader proteins is represented with an array of zeros; each position in the array corresponds to a single amino acid residue in the protein sequence. For each of the distinct peptides from the subset defined above, the corresponding portion of the protein array is incremented by one. After that, all protein arrays are concatenated, and the distribution of values in the resulting array is considered as a representation of proteome coverage. Note that only unique peptides were considered in this procedure; repetitive PSMs for the same peptide were not taken into account, even when they come from samples treated with different cleavage agents. This is done to exclude possible systematic errors that could lead to repetitive incorrect peptide assignments.

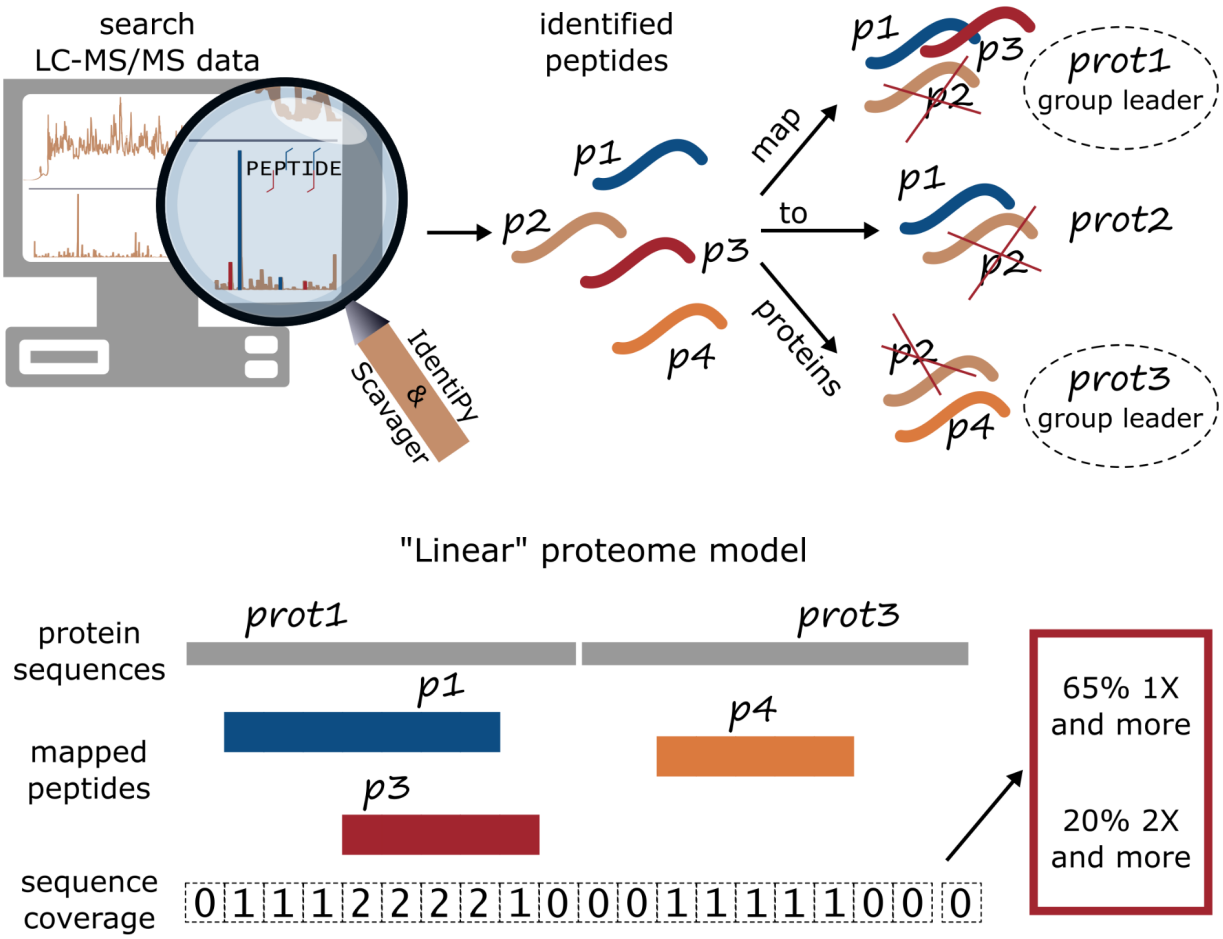


Figure 1. The concept of proteome-wide sequence coverage by multiple reads. To quantify the multiplicity of sequence coverage throughout the observed proteome, we build its "linear" model, composed of "group leader" proteins as reported by Scavager tool [28]. Peptides shared between protein groups are discarded, and the rest are mapped onto the "linear proteome". The "linear proteome" is represented computationally by an array of integer values (initialized with zeros), one for each amino acid position in every detected "group leader" protein. Then, each mapped peptide increments the values corresponding to its place in the protein sequence. Afterwards, the distribution of array values is summarized as proteome coverage statistics.

Two most frequent enzymes used for shotgun proteomics are trypsin and LysC, which have similar specificity, except that digestion after arginine residues is not feasible by LysC. However, in most cases these enzymes are used concurrently. For purposes of this work, it was interesting to find data where two enzymes acted separately, to estimate the gain in coverage

due to the difference in their specificities. A good example of such data was a human brain dataset by Wingo et al [30], where trypsin and LysC were used separately. This proteome, deeply covered using prefractionation of enzymatic lysates, was characterized by 8598 protein groups in total, according to our reprocessing of the raw data. First, we estimated if miscleaved peptides add to the sequence coverage in the trypsin subset of this brain proteome. At a total proteome coverage of 24%, about two percent of the linear proteome sequence was covered 2-fold by miscleaved peptides (see Section 3.1). Combining trypsin data with LysC runs slightly increased the general coverage to 28%. At the same time, coverage of 2-fold or more reached about 6%. Thus, about 22% of all identified proteome regions in this dataset were covered by multiple reads, when both trypsin- and LysC-digested samples were considered .

A unique collection of multiprotease shotgun proteomic data is Confetti, which represents a multiprotease proteome map of the HeLa cell line [20]. Expectedly, the use of seven proteases with different specificities enhanced the multiple coverage up to 35% of the visible proteome regions, while the overall proteome coverage was relatively low (20%) due to relatively poor sensitivity, with 4,8 thousand proteins identified (Table 2).

To demonstrate the sequence coverage approach with our own data, we used a multiprotease digestion scheme on the HEK-293 model cell line which was previously studied in our work focused on its amino acid variants [24]. Proteins extracted from the same sample used before were digested separately by three proteases, including trypsin, LysC and endoproteinase GluC, followed by shotgun analysis of the digests in DDA mode in three technical replicates. These analyses yielded 5435, 4950, and 3307 protein groups for trypsin, LysC, and GluC, respectively. Combined, 6109 protein groups were identified. These data provided a general coverage of the linear proteome about 19%, which is lower than for the brain proteome data described above [30]. Note that the brain proteome results were obtained using prefractionation of the peptide mixture. At the same time, a relative proteome coverage of 2X and more for our HEK-293 data was higher than in the brain data and reached about 25% of the covered regions vs. 22% in the brain dataset. This relative improvement, indicated as significant by Fisher's exact test, was provided by the addition of GluC to the protease set. At the same time, much deeper analysis in the brain dataset provided a two-fold gain in absolute characteristics of multiple coverage, i.e. about 0.34 mln amino acid residues covered 2x and more vs. 0.17 mln in the HEK-293 data. A high portion of identified proteome with multiple coverage in the Confetti dataset (35%) also looks not so impressive in absolute numbers spanning 0.19 mln residues (Fig.2A).

Not many datasets with distinct treatment of multiple proteases are publicly available. However, the three examples provided here confirmed an intuitively clear conclusion that a gain

in proteome sequence coverage in multiple reads may be reached by two experimental conditions. First, a number of proteases used for independent, simultaneous proteome analysis may be increased. Second, both the general coverage and the coverage by multiple reads may be increased with deeper proteomic analysis, attainable e.g. by prefractionation of the peptide mixture and analyzing the fractions separately. There are many examples of successful use of multiprotease analysis to increase the proteome coverage [20–23]. The novelty of our approach is the use of this analysis to confirm selected single amino acid events in contrast to other works which were focused on more complete inference of whole proteins.

As shown in Fig.2B, the total length of the proteome covered with at least *k* distinct peptides decreases exponentially as a function of *k*. This is expected if we consider that these peptide identifications are largely independent. The relationship becomes linear when switching to the logarithmic scale. Straight lines obtained from least-squares linear fit succinctly describe the multi-proteases analysis: the slope of the line depends on the number of proteases used (the more proteases, the flatter), while the intercept corresponds to the depth of the proteome analysis of each of the parallel analyses. For example, the brain datasets processed with trypsin only and with two proteases have nearly identical total numbers of residues covered exactly once (two left blue bars in Fig.2A, left-most blue and yellow points in Fig. 2B, see inset). Thus, the blue and yellow lines have the same value of intercept, however, the addition of LysC results in a flatter slope.

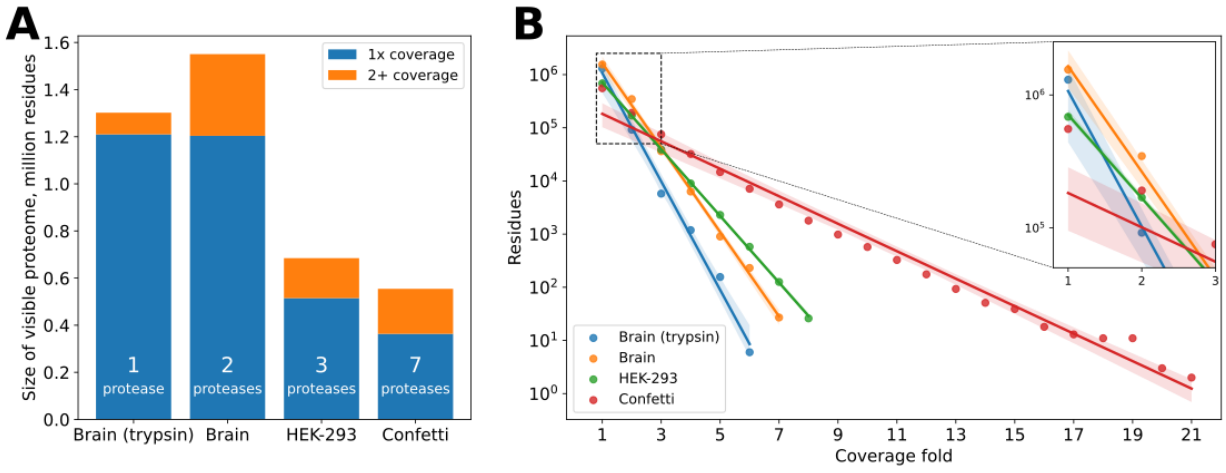


Figure 2. Size comparison of visible proteome sizes in the human brain, HEK-293 and Confetti datasets. The brain sample is sampled much deeper due to fractionation; addition of LysC analysis adds more coverage to the same sequence regions in already detected proteins. Increasing the number of proteases used in parallel results in a higher proportion of multiply-covered sequences (*A*). Distributions of coverage multiplicity (fold) values in detected proteins

follow exponential trends (straight lines in logarithmic scale). Line parameters depend on the number of proteases (slope) and sensitivity in a single analysis (intercept) (B).

Table 2. Characteristics of the sequence coverage calculated proteome-wide for shotgun proteomic datasets generated using multiple proteases

Source of biomaterial, PRIDE accession, reference	Human brain, PXD004143 [30]*	Human brain, PXD004143 [30]	HeLa human cell line (Confetti multiprotease proteome map), PXD000900 [20]	HEK-293 human cell line, PXD030226, this paper
Proteases	Trypsin	Trypsin, LysC	Trypsin, chymotrypsin, elastase, LysC, ArgC, AspN, GluC	Trypsin, LysC, GluC
Total proteome depth, protein groups	8314	8598	4819	6109
Peptide amount, total / considered	110,356 / 104,486	142,599 / 135,264	64,288 / 60,575	72,144 / 64,925
Visible "linear" proteome length, residues	5,382,815	5,468,413	2,756,099	3,572,977
Sequence coverage, 1X and more	24.2%	28.4%	20.1%	19.2%
Sequence coverage, 2X and more	**1.7%**	**6.3%**	**6.9%**	**4.7%**

* Shown for a digestion by trypsin only, to illustrate what gain in 2X coverage was provided by trypsin miscleavage.

3.3. Multiprotease shotgun analysis of HEK-293 proteome: sequence coverage of single amino acid variants

A modified genomic database for peptide identification with added amino acid variants was taken from our previous study [24]. The new dataset made it possible to estimate (i) the reproducibility of amino acid variant identification between different datasets; (ii) the gain in variant identification from the addition of treatment by GluC, a protease with complementary specificity

to trypsin and LysC; and, finally, (iii) the portion of coverage of these variants by multiple peptide reads.

Fortunately, the majority of 36 protein sequence variants found in HEK-293 cells (Table 3) were also identified before in at least one of three reported proteomes of this cell line [24,43,44] described in detail in our background paper [24]. More specifically, 28 of them were reproduced here (Table 3). This amount of variants was generally comparable with two background studies, where 32 and 38 variants were reported from our own [24] and Geiger et al. data [44]. Reprocessing the data for the much deeper proteome by Chick et al [43] yielded 84 amino acid variants identified. The use of GluC in our study led to the addition of 5 new variants which fell into parts of sequences poorly compatible with mass-spectrometric identification after digestion by trypsin/LysC. Two variants covered only by GluC-generated peptides in our data, namely, in PIH1D1 and TP53BP1 proteins, were also detected by tryptic peptides in the data by Geiger et al [44] and by Chick et al [43], respectively (Table 3).

The variant sequence coverage was estimated as exemplified in Fig.3. The portion of variants with coverage 2X or more was quite similar to that observed proteome-wide. Namely, of 36 variants identified with use of any enzyme, seven were covered by at least two distinct peptides (19.4%). Of these, the K216Q sequence variant of A-Kinase Anchoring Protein 12 (AKAP12) and the G514E variant of cancer-related Transforming Acidic Coiled-Coil Containing Protein 3 (TACC3) were covered by three distinct peptides generated by all three enzymes used, i.e. these sites had 3X coverage. Variants of proteins encoded by MMUT, PCMT1, PRRC2C, and SWAP70 all had 2X coverage provided by trypsin and LysC. The identification of the Q505E variant of SWAP70 may either result from genomic mutation or post-translational glutamine deamination [6]. However, we report it for illustration of our approach. Finally, the R1324L variant of RRBP1 was also covered 2-fold by LysC and GluC (Table 3).

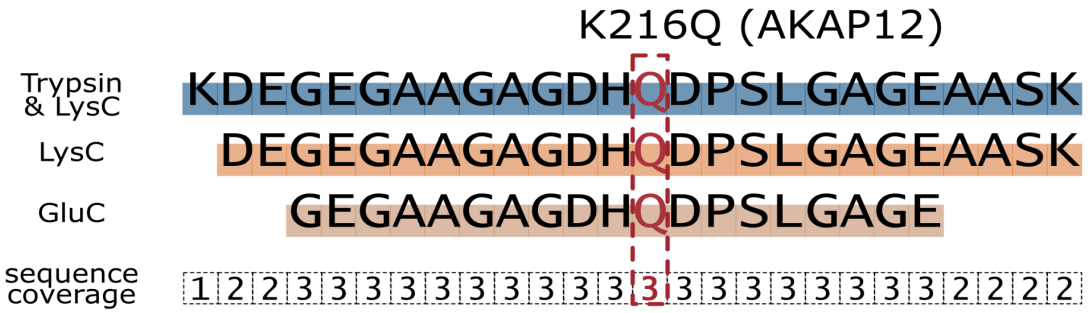


Figure 3. Sequence coverage of the K216Q single amino acid variant of A-Kinase Anchoring Protein 12 (AKAP12) by three distinct peptides generated by trypsin, LysC and GluC.

Table 3. Missense genomic variants identified in the shotgun proteome of the HEK-293 cell line using the multi-enzyme approach to provide enhanced sequence coverage by distinct peptide reads. Genomic database supplemented by amino acid variants was taken from the previous study [24]. The amino acid numbering is represented according to the isoform 1 in the NextProt knowledgebase, data release: 2021-02-15 [45].

Gene name	Enzyme	Peptide sequence	Amino acid change	Sequence coverage, fold	Match to 1 to 3 HEK-293 datasets described earlier [24]
Amino acid sequence variants with coverage 2X or more					
AKAP12*	Trypsin, LysC	KDEGEGAAGAGDH**[Q]**DPSLGAGEAASK	K216Q	3	1
	LysC	DEGEGAAGAGDH**[Q]**DPSLGAGEAASK			
	GluC	GEGAAGAGDH**[Q]**DPSLGAGE			
MMUT	Trypsin	YQLEKED**[T]**VEVLAIDNTSVR	A499T	2	1
	LysC	ED**[T]**VEVLAIDNTSVRNRQIEK			
PCMT1	Trypsin	ELVDDS**[I]**NNVR	V120I	2	3
	LysC	ELVDDS**[I]**NNVRK			
PRRC2C	Trypsin	TLS**[T]**PQEER	A906T	2	2
	LysC	TLS**[T]**PQEERISAVESQPSRK			

RRBP1	LysC	LTAEFEEAQTSAC**[L]**LQEELEK	R1324L	2	2
	GluC	FEEAQTSAC**[L]**LQEE			
SWAP70 **	Trypsin	EQALQEAMEQLE**[E]**LELER	Q505E	2	1
	LysC	EQALQEAMEQLE**[E]**LELERK			
TACC3	Trypsin	ALNSASTSLPTSCPGSEPVPTH QQ[**E**]QPALELK	G514E	3	2
	LysC	ERALNSASTSLPTSCPGSEPVPT HQQ**[E]**QPALELK			
	GluC	RALNSASTSLPTSCPGSEPVPT HQQ**[E]**QPALE			
Amino acid sequence variants with coverage 1X					
ACOX1	Trypsin	HQSE**[M]**KPGEPEPQILDFQTQQ YK	I312M	1	2
ALCAM	Trypsin	SSNTYTL**[M]**DVR	T301M	1	2
ALDH1B 1	LysC	EAGFPPGVVNIITGYGPTAGAAI AQH**[M]**DVDK	V253M	1	1
CCDC10 4	GluC	TSSLPQK[**G**]LKIPGLE	D243G	1	-
DCK	Trypsin	**[S]**VLFFER	P122S	1	1
DNTTIP2	LysC	DLDEDANGITD**[D]**GK	E309D	1	2
DST	GluC	ASDSKG**[A]**SDVLLQVE	T5138A	1	-
EIF4G1	GluC	VTAS**[V]**APPTIPSATPATAPSAT SPAQEEE	M432V	1	-

GALNT8	LysC	R**[D]**GAVIK	Y53D	1	-
GLRX3	LysC	VQRHASSGSFL**[S]**SANEHLK	P123S	1	3
LBR	Trypsin	F**[N]**LSQESSYIATQYSLRPR	S154N	1	2
MDN1	Trypsin	LVASELHTSL**[Y]**SSMVGADR	H3423Y	1	1
MPRIP	LysC	AEHMETNAVGPS**[Q]**SSDTRQG RSEK	P327Q	1	2
NES	GluC	SAAGAEPG**[L]**GQGVGGLGDPG HLTREE	P1101L	1	-
PIH1D1	GluC	**[I]**DLPKLDGALGLSLE	V224I	1	1***
PITRM1	Trypsin, LysC	DPSWII**[R]**	Q1037R	1	2
PLCG1	LysC	AQREDELTF**[T]**K	I813T	1	2
RRP1B	LysC	HHLQPENPGPGGAAPSLEQNR GREPEASG**[P]**K	L436P	1	1
SEC23A	LysC	VP**[V]**TQATRGPQVQQPPPSNRF LQPVQK	L211V	1	3
SKA1	Trypsin	EN**[I]**PSHLPQVTVTQSCVK	V91I	1	2
SULT1A 1	Trypsin	SLPEETVDF**[M]**VQHTSFK	V223M	1	2
SYNE2	LysC	QATSDVQESTQESA**[T]**VEK	A2395T	1	-
TBC1D5	Trypsin	GQGQSVQMSGA**[V]**K	I696V	1	1

TP53	Trypsin	**[R]****** VAPAPAAPTPAAPAPAPSWPLS SSVPSQK	P72R	1	3
TP53BP1	GluC	GGCSLASTPATTLHLLQLSGQR SLVQ**[E]**	D353E	1	1***
TRUB1	LysC	LLAEAGMPSP**[A]**WTK	E103A	1	2
TSPYL4	GluC	HASGDPD**[L]**DQCQGLREE	R30L	1	-
ZFYVE16	Trypsin	EQQND**[T]**SSELQNR	I192T	1	1
ZWINT	LysC	TGTQQELD**[G]**VFQK	R187G	1	-

* The site was not reported in [24] as K-to-Q substitution has a small mass difference, although it was present in the output. However, the current data are produced using high-resolution mass spectrometry for ion fragments, where the variant is confirmed by mass-spectra reliably.

** The variant identification may result from genomic mutation or post-translational glutamine deamination.

*** The variants were covered by tryptic peptides in other datasets, such as the PIH1D1 variant in the data by Geiger et al [44] and the TP53BP1 variant in the data by Chick et al [43].

**** Arginine residue is before the tryptic peptide which was identified. In some cases, this peptide could be a result of in-source decay, which, however, specifically relates to N- and C-terminal residues in a peptide, in contrast to the current peptide [46].

One of the ways to increase reliability of findings confirmed by specific PSMs, such as single amino acid variants discussed above, is inspection of mass spectra. Mass spectra of the variants with the sequence coverage of two or more (Table 3) were visualized by xiSPEC spectrum viewer [47] to determine if the specific amino acid residue at the alleged mutation site was confirmed by the corresponding b- and/or y-ions. For the K216Q variant in AKAP12, the corresponding glutamine residue was confirmed by fragmentation mass-spectra attributed to all three distinct peptides containing the variant (Fig. 4). Correspondingly, single amino acid variants in PCMT1, PRRC2C, and RRBP1 were confirmed by fragment mass spectra for all distinct peptides containing these variants (Figs. S1-S3, respectively). Further, the G514E variant in TACC3 was supported by fragment ions in two of three possible cases (Fig.S4). The variants in MMUT and SWAP70 were confirmed by fragmentation of one of two peptides (Figs. S5-S6).

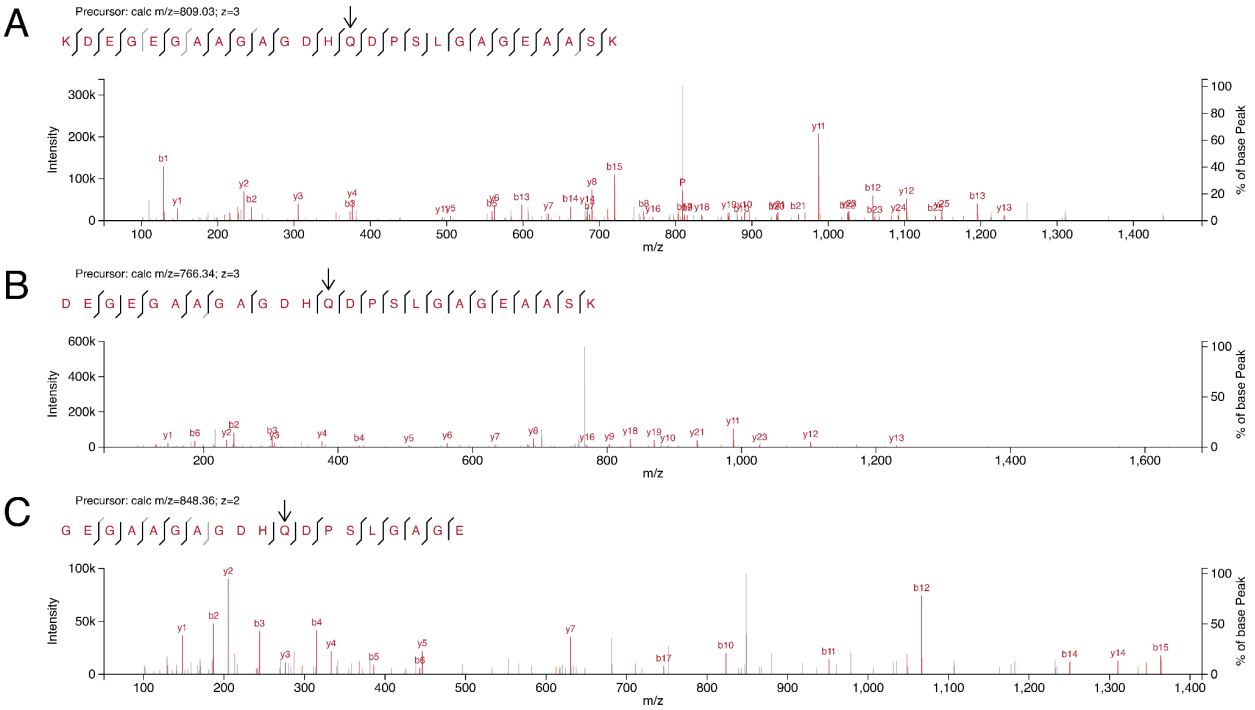


Figure 4. Fragmentation mass spectra of peptides of A-Kinase Anchoring Protein 12 (AKAP12) containing a genetically encoded K216Q amino acid sequence variant. Mass spectra recorded from the shotgun proteomes of HEK-293 cell line after protein digestion by trypsin (A), LysC (B), and GluC (C). Vertical bars around amino acid letters show that residues were confirmed by y-ions (upper branches) or b-ions (under branches). Mass spectra were visualized by xiSPEC spectrum viewer [47].

A question might arise whether the proteins with variants covered 2X or more were highly abundant in HEK-293 proteome. Using NSAF label-free quantitation method [48], percentiles of abundance could be defined for each of them in the trypsin-digested dataset. Indeed, AKAP12, a structural component of protein kinase A signalosome, was abundant in this cell line, placed at the 84th percentile. However, the other six proteins on the list were more or less evenly distributed along the list, for example, with 28th percentile for SWAP70. Thus, the coverage of specific sequences is controlled by physical properties of the corresponding peptides.

3.4. Sequence coverage of splice junctions for alternative splicing

Alternative splicing site coverage was calculated in the following way. Based on the RefSeq reference genome annotation (GRCh38) all consecutive coding sequence (CDS) pairs were considered in all annotated transcripts of all human genes. Theoretical peptide sequences were generated by translating these pairs and performing *in silico* digestion in accordance with

enzyme specificity of trypsin, LysC, or GluC, allowing for 2, 2 and 4 miscleavages, respectively. Peptides lying entirely within a single CDS, as well as peptides present in all protein products of the given gene, were discarded. Thus, only alternative splicing peptide products were kept for each possible site.

For 116,018 possible alternative splicing events in 12,968 genes, a total of 464,078 theoretical confirming peptides were predicted for trypsin, 253,815 for LysC and 7,118 for GluC. To calculate the coverage of each splicing site, a database search was performed against the regular RefSeq protein database. Then, reliably identified peptides confirming each site were counted across all experiments.

In our experiments, peptide evidence was found for 4209 alternative splicing events in 2101 genes of HEK-293 cells (Table S1). However, some peptides could not be mapped unambiguously to a single gene due to paralogy. A total of 273 unique junction-spanning peptides from paralogous proteins were identified; they were excluded from further analysis, leaving 3893 unique peptides. Additionally, some of these peptides attributed to a single gene still did not point unambiguously to a specific alternative splicing event, due to the fact that the splicing site was too close to the start or end of the peptide sequence. In those cases, almost all of the sequence coding the peptide is located within one exon, and the remaining few nucleotides are not enough to unambiguously identify the other exon. These cases were also excluded, leaving 3865 peptides as evidence for 3350 splice events. Of these events, only 478 (or 14%) were covered 2-fold or more. This percentage is much lower than for the total "linear" proteome, where it was about 25%. This observation can be presumably explained by decreased abundances of alternatively spliced proteoforms, as well as by their enrichment by lysine and arginine sites, limiting identification by trypsin and LysC [49].

3.5. MS1 approach to confirm amino acid sequence variants in HEK-293 proteome

An approach of express proteomic analysis which uses only precursor *m/z* values without MS/MS, called DirectMS1, was recently elaborated by authors of this work [33]. In this method, peptide-level FDR is usually uncontrolled and high, which makes variant identification tricky. However, in the case of three different proteases, the variant could be reliably confirmed if it was detected in three distinct peptides. In our analysis, two variants met this criteria, K216Q in the abundant AKAP12 protein, which was also covered 3X in conventional analysis (see Table 3), and, unexpectedly, I813T from PLCG1. In addition to the AQREDELTFTK peptide identified from the LysC subset (Table 3), DirectMS1 identified a tryptic fragment, EDELTFTK, and a GluC-produced fragment, DELTFTKSAIIQNVE. Probably, these peptides failed to pass the intensity

threshold during conventional MS/MS-based data acquisition. Indeed, relative MS1 intensities for the LysC, tryptic and GluC peptides were 8×10^6 (the one detected in MS/MS analysis), 1×10^6 and 4×10^5, respectively. Thus, the orthogonal identification method, based on exact *m/z* of precursor ions, added a variant hit to the results of MS/MS search.

4. Concluding remarks

The omics technologies today generate increasingly large data arrays providing biologically relevant information at an accelerating pace compared with the pre-genomic era in molecular biology. However, the omics, while providing the ways to measure thousands of molecular variables simultaneously, are suffering from reliability issues compared with the one-by-one approaches of the old days. New discoveries based on the high throughput omics are not self-sufficient and need extensive validation steps by orthogonal methods. This limits wider and more useful acceptance of omics in some applications, such as the clinical ones. NGS improved significantly the reliability issue in genomics, thus, paving its way to clinical discoveries and diagnosis [50].

However, mass spectrometry-based shotgun proteomics, despite the recent progress in its performance, is still lacking reliability in identification of every single amino acid of proteins proteome-wide. Here we considered a more conservative way to interpret the proteome analysis results obtained in the context of proteogenomic studies using data-dependent acquisition mode to distill the more reliable part of them, better suitable for clinical discoveries. Instead of considering all variant peptides identified at a given FDR level as equal, we suggest their further ranking using the sequence coverage by multiple reads approach in a similar way as it is done in nucleic acid sequencing for the calling of single nucleotide variants. Multiple reads for each letter in the proteome sequence can be obtained by overlapping distinct peptides, which confirm the presence of certain amino acid residues in the overlapping stretch with much lower FDR compared with 1% accepted for the whole group of identifications. These overlapping distinct peptides can be formed by, first, the pairs of miscleaved tryptic peptides and their fully cleaved counterparts, and, second, the peptides generated by several proteases with different specificities applied to the same specimen. The corresponding digests should be analyzed separately using the multiprotease proteome analysis workflow well known in proteomics [20].

Note that contrary to transcriptomics, which provides rich sequence coverage content, the coverage of each protein by peptide "reads" is relatively small in proteomics. For example, the coverage of human proteome in exemplary datasets, even with a single read, is typically in the range of 20% to 30%, with 5% to 7% of the proteome covered two-fold or more. Similarly, of 36

single amino acid variants identified here for the HEK293 cell line, 7 were covered at least two-fold. However, these 7 variants can be considered as bullet-proof identifications which can be reliably explored further in clinical studies.

The approach of sequence coverage with multiple reads may be useful for clinical applications, for example, identification of cancer missense mutations which may serve as neoantigens, including experimental schemes of neoantigen vaccine production [51]. More reliable validation of cancer mutations at the proteome level may facilitate prioritization of candidate neoantigens for personalized vaccines [52]. Further, actionable or diagnostic cancer mutations and splice junctions may be identified more reliably with the proposed method, which would make it possible to omit a further validation stage, such as targeted proteomics or antibody-based methods. Outside of medicine, sequence coverage may validate basic scientific findings derived from the identifications of short sequences, such as novel classes of micro proteins derived from RNAs which were previously thought as non-coding [53].

The work was funded by the Russian Science Foundation, grant #20-15-00072, to S.M. We thank the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Federal Research and Clinical Center of Physical-Chemical Medicine of the Federal Medical Biological Agency for providing computational resources for this project.

The authors have declared no conflict of interest.

5. References

[1] Wolters, D.A., Washburn, M.P., Yates, J.R., An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* 2001, 73, 5683–90.

[2] Nesvizhskii, A.I., Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* 2005, 4, 1419–40.

[3] Elias, J.E., Haas, W., Faherty, B.K., Gygi, S.P., Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* 2005, 2, 667–75.

[4] Elias, J.E., Gygi, S.P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, 4, 207–14.

[5] Gupta, N., Pevzner, P.A., False discovery rates of protein identifications: a strike against the two-peptide rule. *J. Proteome Res.* 2009, 8, 4173–81.

[6] Moshkovskii, S.A., Ivanov, M. V, Kuznetsova, K.G., Gorshkov, M. V, Identification of Single

Amino Acid Substitutions in Proteogenomics. *Biochemistry. (Mosc).* 2018, 83, 250–258.

[7] Liu, Y., Gonzàlez-Porta, M., Santos, S., Brazma, A., et al., Impact of Alternative Splicing on the Human Proteome. *Cell Rep.* 2017, 20, 1229–1241.

[8] Levitsky, L.I., Kliuchnikova, A.A., Kuznetsova, K.G., Karpov, D.S., et al., Adenosine-to-Inosine RNA Editing in Mouse and Human Brain Proteomes. *Proteomics* 2019, 19, 1900195.

[9] Paik, Y.-K., Lane, L., Kawamura, T., Chen, Y.-J., et al., Launching the C-HPP neXt-CP50 Pilot Project for Functional Characterization of Identified Proteins with No Known Function. *J. Proteome Res.* 2018, 17, 4042–4050.

[10] Caron, E., Kowalewski, D., Chiek Koh, C., Sturm, T., et al., Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry*. *Mol. Cell. Proteomics* 2015, 14, 3105–3117.

[11] Wen, B., Li, K., Zhang, Y., Zhang, B., Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat. Commun.* 2020, 11, 1759.

[12] Gabriels, R., Martens, L., Degroeve, S., Updated MS^2PIP web server delivers fast and accurate MS^2 peak intensity prediction for multiple fragmentation methods, instruments and labeling techniques. *Nucleic Acids Res.* 2019, 47, W295–W299.

[13] Gessulat, S., Schmidt, T., Zolg, D.P., Samaras, P., et al., Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* 2019, 16, 509–518.

[14] C Silva, A.S., Bouwmeester, R., Martens, L., Degroeve, S., Accurate peptide fragmentation predictions allow data driven approaches to replace and improve upon proteomics search engine scoring functions. *Bioinformatics* 2019, 35, 5243–5248.

[15] Henson, J., Tischler, G., Ning, Z., Next-generation sequencing and large genome assemblies. *Pharmacogenomics* 2012, 13, 901–15.

[16] Slatko, B.E., Gardner, A.F., Ausubel, F.M., Overview of Next-Generation Sequencing Technologies. *Curr. Protoc. Mol. Biol.* 2018, 122, e59.

[17] Chen, C., Hou, J., Tanner, J.J., Cheng, J., Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *Int. J. Mol. Sci.* 2020, 21.

[18] Nesvizhskii, A.I., Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* 2014, 11, 1114–25.

[19] Šlechtová, T., Gilar, M., Kalíková, K., Tesařová, E., Insight into Trypsin Miscleavage: Comparison of Kinetic Constants of Problematic Peptide Sequences. *Anal. Chem.* 2015, 87, 7636–43.

[20] Guo, X., Trudgian, D.C., Lemoff, A., Yadavalli, S., Mirzaei, H., Confetti: a multiprotease

map of the HeLa proteome for comprehensive proteomics. *Mol. Cell. Proteomics* 2014, 13, 1573–84.

[21] Meyer, J.G., Kim, S., Maltby, D.A., Ghassemian, M., et al., Expanding proteome coverage with orthogonal-specificity α-lytic proteases. *Mol. Cell. Proteomics* 2014, 13, 823–35.

[22] Swaney, D.L., Wenger, C.D., Coon, J.J., Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J. Proteome Res.* 2010, 9, 1323–9.

[23] Miller, R.M., Millikin, R.J., Hoffmann, C. V, Solntsev, S.K., et al., Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data. *J. Proteome Res.* 2019, 18, 3429–3438.

[24] Lobas, A.A., Karpov, D.S., Kopylov, A.T., Solovyeva, E.M., et al., Exome-based proteogenomics of HEK-293 human cell line: Coding genomic variants identified at the level of shotgun proteome. *Proteomics* 2016.

[25] Rappsilber, J., Ishihama, Y., Mann, M., Stop and Go Extraction Tips for Matrix-Assisted Laser Desorption/Ionization, Nanoelectrospray, and LC/MS Sample Pretreatment in Proteomics. *Anal. Chem.* 2003, 75, 663–670.

[26] Adusumilli, R., Mallick, P., Data conversion with proteoWizard msConvert. *Methods Mol. Biol.* 2017, 1550, 339–368.

[27] Levitsky, L.I., Ivanov, M. V, Lobas, A.A., Bubis, J.A., et al., IdentiPy: An Extensible Search Engine for Protein Identification in Shotgun Proteomics. *J. Proteome Res.* 2018, 17, 2249–2255.

[28] Ivanov, M. V, Levitsky, L.I., Bubis, J.A., Gorshkov, M. V, Scavager: A Versatile Postsearch Validation Algorithm for Shotgun Proteomics Based on Gradient Boosting. *Proteomics* 2019, 19, 1800280.

[29] Lobas, A.A., Pyatnitskiy, M.A., Chernobrovkin, A.L., Ilina, I.Y., et al., Proteogenomics of Malignant Melanoma Cell Lines: The Effect of Stringency of Exome Data Filtering on Variant Peptide Identification in Shotgun Proteomics. *J. Proteome Res.* 2018, 17, 1801–1811.

[30] Wingo, T.S., Duong, D.M., Zhou, M., Dammer, E.B., et al., Integrating Next-Generation Genomic Sequencing and Mass Spectrometry To Estimate Allele-Specific Protein Abundance in Human Brain. *J. Proteome Res.* 2017, 16, 3336–3347.

[31] Bubis, J.A., Levitsky, L.I., Ivanov, M. V, Gorshkov, M. V, Validation of Peptide Identification Results in Proteomics Using Amino Acid Counting. *Proteomics* 2018, 18, e1800117.

[32] Levitsky, L.I., Bubis, J.A., Gorshkov, M. V, Tarasova, I.A., AA_stat: Intelligent profiling of in vivo and in vitro modifications from open search results. *J. Proteomics* 2021, 248, 104350.

[33] Ivanov, M. V, Bubis, J.A., Gorshkov, V., Abdrakhimov, D.A., et al., Boosting MS1-only Proteomics with Machine Learning Allows 2000 Protein Identifications in Single-Shot Human Proteome Analysis Using 5 min HPLC Gradient. *J. Proteome Res.* 2021, 20, 1864–1873.

[34] Bubis, J.A., Gorshkov, V., Gorshkov, M. V, Kjeldsen, F., PhosphoShield: Improving Trypsin Digestion of Phosphoproteins by Shielding the Negatively Charged Phosphate Moiety. *J. Am. Soc. Mass Spectrom.* 2020, 31, 2053–2060.

[35] Klammer, A.A., MacCoss, M.J., Effects of modified digestion schemes on the identification of proteins from complex mixtures. *J. Proteome Res.* 2006, 5, 695–700.

[36] Tsiatsiani, L., Heck, A.J.R., Proteomics beyond trypsin. *FEBS J.* 2015, 282, 2612–2626.

[37] Guergues, J., Wohlfahrt, J., Zhang, P., Liu, B., Stevens, S.M., Deep proteome profiling reveals novel pathways associated with pro-inflammatory and alcohol-induced microglial activation phenotypes. *J. Proteomics* 2020, 220, 103753.

[38] Saei, A.A., Sabatier, P., Tokat, Ü.G., Chernobrovkin, A., et al., Comparative Proteomics of Dying and Surviving Cancer Cells Improves the Identification of Drug Targets and Sheds Light on Cell Life/Death Decisions. *Mol. Cell. Proteomics* 2018, 17, 1144–1155.

[39] Kuznetsova, K.G., Kliuchnikova, A.A., Ilina, I.U., Chernobrovkin, A.L., et al., Proteogenomics of Adenosine-to-Inosine RNA Editing in the Fruit Fly. *J. Proteome Res.* 2018, 17, 3889–3903.

[40] Sharma, K., Schmitt, S., Bergner, C.G., Tyanova, S., et al., Cell type- and brain region-resolved mouse brain proteome. *Nat. Neurosci.* 2015, 18, 1819–31.

[41] Levitsky, L.I., Ivanov, M. V, Lobas, A.A., Gorshkov, M. V, Unbiased False Discovery Rate Estimation for Shotgun Proteomics Based on the Target-Decoy Approach. *J. Proteome Res.* 2017, 16, 393–397.

[42] Sims, D., Sudbery, I., Ilott, N.E., Heger, A., Ponting, C.P., Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 2014, 15, 121–132.

[43] Chick, J.M., Kolippakkam, D., Nusinow, D.P., Zhai, B., et al., A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* 2015, 33, 743–749.

[44] Geiger, T., Wehner, A., Schaab, C., Cox, J., Mann, M., Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Mol. Cell. Proteomics* 2012, 11, M111.014050.

[45] Zahn-Zabal, M., Michel, P.-A., Gateau, A., Nikitin, F., et al., The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Res.* 2019.

[46] Kuznetsova, K.G., Levitsky, L.I., Pyatnitskiy, M.A., Ilina, I.Y., et al., Cysteine alkylation methods in shotgun proteomics and their possible effects on methionine residues. *J. Proteomics* 2021, 231, 104022.

[47] Kolbowski, L., Combe, C., Rappsilber, J., xiSPEC: web-based visualization, analysis and sharing of proteomics data. *Nucleic Acids Res.* 2018, 46, W473–W478.

[48] Zybailov, B.L., Florens, L., Washburn, M.P., Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. *Mol. Biosyst.* 2007, 3, 354–360.

[49] Wang, X., Codreanu, S.G., Wen, B., Li, K., et al., Detection of Proteome Diversity Resulted from Alternative Splicing is Limited by Trypsin Cleavage Specificity. *Mol. Cell. Proteomics* 2018, 17, 422–430.

[50] Lee, H., Martinez-Agosto, J.A., Rexach, J., Fogel, B.L., Next generation sequencing in clinical diagnosis. *Lancet. Neurol.* 2019, 18, 426.

[51] Polyakova, A., Kuznetsova, K., Moshkovskii, S., Proteogenomics meets cancer immunology: mass spectrometric discovery and analysis of neoantigens. *Expert Rev. Proteomics* 2015, 1–9.

[52] Rivero-Hinojosa, S., Grant, M., Panigrahi, A., Zhang, H., et al., Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. *Nat. Commun.* 2021, 12, 6689.

[53] Fesenko, I., Shabalina, S.A., Mamaeva, A., Knyazev, A., et al., A vast pool of lineage-specific microproteins encoded by long non-coding RNAs in plants. *Nucleic Acids Res.* 2021, 49, 10328–10346.