ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Multiaspect Examinations of Possible Alternative Mappings of Identified Variant Peptides: A Case Study on the HEK293 Cell Line

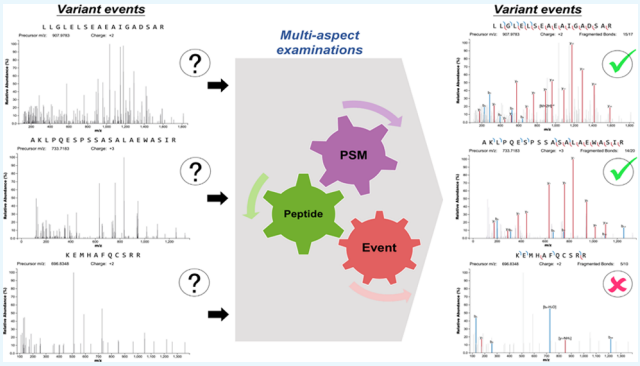
Wai-Kok Choong and Ting-Yi Sung*

Cite This: *ACS Omega* 2022, 7, 16454−16467

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Adopting proteogenomics approach to validate single nucleotide variation events by identifying corresponding single amino acid variant peptides from mass spectrometry (MS)-based proteomics data facilitates translational and clinical research. Although variant peptides are usually identified from MS data with a stringent false discovery rate (FDR), FDR control could fail to eliminate dubious results caused by several issues; thus, postexamination to eliminate dubious results is required. However, comprehensive postexaminations of identification results are still lacking. Therefore, we propose a framework of three bottom-up levels, peptide−spectrum match, peptide, and variant event levels, that consists of rigorous 11-aspect examinations from the MS perspective to further confirm the reliability of variant events. As a proof of concept and showing feasibility, we demonstrate 11 examinations on the identified variant peptides from an HEK293 cell line data set, where various database search strategies were applied to maximize the number of identified variant PSMs with an FDR <1% for postexaminations. The results showed that only FDR criterion is insufficient to validate identified variant peptides and the 11 postexaminations can reveal low-confidence variant events detected by shotgun proteomics experiments. Therefore, we suggest that postexaminations of identified variant events based on the proposed framework are necessary for proteogenomics studies.



1. INTRODUCTION

Single amino acid variations (SAVs) in proteins could affect protein folding, protein−protein interaction, and protein domain functionality that could cause diseases or cancers.[1−5] Moreover, somatic SAVs in tumor samples may become a neoantigen for developing personalized therapeutic vaccines for cancers.[6,7] To identify SAVs at the proteomic level, a proteogenomics approach is usually applied, which utilizes the single nucleotide variant (SNV) information derived from next-generation sequencing of genomic or transcriptomic data to generate SAV-harboring protein sequences and identify SAV variant peptides from mass spectrometry (MS)-based proteomics data.[8−11] To obtain confidently identified variant peptides, the target−decoy database search approach is commonly applied to estimate a false discovery rate (FDR) to filter out unconfident identification.

Although variant PSMs are obtained from a rigorous identification procedure passing an FDR of 1%, it is mentioned in the literature that we must still check their reliability to avoid false positives.[10−13] For example, database searches may yield false-positive PSMs, which still pass the FDR threshold because an incomplete protein sequence database is used for searches.[10] Or because possible modifications are neglected in the search parameters,[14−16] spectra of modified peptides can be incorrectly identified as other peptide sequences. Furthermore, considering 11 types of isobaric substitutions such as isoleucine/leucine substitution and delaminated-glutamine/glutamate substitution, more than 6% of variant peptides in neXtProt can be interpreted as wild-type peptides of PE1 proteins, i.e., human proteins with experimental evidence at the protein level as classified by neXtProt.[17] Thus, it is necessary to further examine the reliability of identified variant peptides.

Several criteria or methods for postexamination of identified variant peptides have been reported to eliminate dubious results. For instance, the claimed variant peptides should not appear in any major reference protein sequence database such as Ensembl and RefSeq[18,19] nor they match any modification or isobaric substitution of a wild-type peptide.[10,12,13,17] Three bioinformatics tools—SpectrumAI,[20] SAVControl,[21] and PepQuery[22]—have been developed to validate the quality of claimed variant peptides. SpectrumAI and SAVControl use different concepts to validate the reliability of a variant peptide at the site level. SpectrumAI examines whether the mass difference of the ions flanking both sides of the variant site on the MS/MS

Received: January 23, 2022
Accepted: April 20, 2022
Published: May 2, 2022

spectrum equals the mass of the amino acid variant. SAVControl, in contrast, adopts mass shift relocation to assess the reliability of the variant site, where the mass shift, defined as the variant peptide precursor mass minus the wild-type peptide mass, is relocalized on each position of the wild-type peptide sequence to confirm that the mass shift localizing at the variant site has the highest matching probability. PepQuery adopts a peptide-centric strategy, which takes the user's preselected variant peptide sequences as input to retrieve the highest-scoring variant peptide−spectrum matches (PSMs) and then performs unrestricted modification searching with all of the modifications from Unimod to confirm that the variant PSMs have no other interpretation.

Identifying and confirming variant events from MS data sets is a critical process for proving that SNVs or SAVs exist in biosamples. From the perspective of MS-based qualitative proteomics, the procedure of variant event identification is highly similar to the protein identification procedure, which is conducted at three hierarchical levels—the PSM, peptide, and protein levels—to filter out false-positive identification. Furthermore, when publishing protein identification results of large-scale MS-based proteomics data, journals generally require that the reliability of identification results at each of the aforementioned levels be reported; for example, for special issues of the Human Proteome Project in Journal of Proteome Research, relevant requirements are mentioned in MS-based data interpretation guidelines.[23,24] However, comprehensive guidelines that account for similar hierarchical levels to eliminate dubious results of variant event identification are still lacking.

In this paper, we propose a framework around three bottom-up levels—PSM, peptide, and variant event—for the post-examination of identified variant events. In this framework, rigorous 11-aspect examinations at the three levels, i.e., five PSM-level, two peptide-level, and four variant event-level examinations, are proposed to further confirm the reliability of identified variant peptides. To demonstrate these examinations, we first conducted a comprehensive variant peptide identification study on the HEK293 cell line to acquire the maximized number of identified variant peptides. Then, only the identified variant PSMs and their variant peptides, not the whole MS data set, were further evaluated for the reliability of the variant events by the postexaminations. Our results reveal that variant peptide identification that only passes the FDR threshold is insufficient for ensuring authenticity, and these examinations can reveal low-confidence variant events. Thus, we suggest using the proposed examination methods to further examine the reliability of variant events at the PSM, peptide, and variant event levels.

2. MATERIALS AND METHODS

2.1. MS/MS Data Set and Variant Information of the HEK293 Cell Line. An MS/MS data set (a total of 24 .raw data files, PXD001468) of the HEK293 cell line was downloaded from the PRIDE[25] (PRoteomics IDEntifications) database. The data set was acquired using a Q-Exactive Orbitrap spectrometer (Thermo Fisher Scientific, San Jose, CA) with higher-energy collisional dissociation for peptide fragmentation. Detailed information about sample collection, experimental preparation, and MS conditions was reported in Chick et al.[26] We converted the MS/MS files from .raw to .mgf using MSConvert (ProteoWizard 3.0.11110 64-bit).[27]

We obtained 1336 genome-annotated variant information of the HEK293 cell line from the supplementary data (Supplementary Data 2; sheet name: "Known common homozygous SNP") in Lin et al.[28] Since we used the Swiss-Prot database for sequence database searching but the variant information was provided with RefSeq identifiers, we filtered out any variants in the proteins with RefSeq identifiers unmatched to Swiss-Prot identifiers. This yielded 1123 single amino acid variants for this study. We henceforth use the variant to refer to "single amino acid variant". Note that this pair of genomic and proteomics data sets of HEK293 cell line was used for proof of concept of the proposed postexaminations on the identified variant peptides, regardless of the complexity of their source samples.

2.2. Construction of a Customized Target−Decoy Protein Database. The target protein sequence database used for database searches of the HEK293 cell line MS data set consisted of the following entries: (1) 42 197 human protein sequences, including isoforms, from the UniProt database (ver. 201707, human),[29] (2) 48 sequences of contaminants from the cRAP (common Repository of Adventitious Proteins) database,[30] (3) 35 sequences of the human adenovirus C serotype 5 (HAdV-5) proteome from UniProt (ver. 201707), and (4) 1123 sequences including HEK293 genome-annotated variants. The 1123 variant protein sequences were generated by integrating exactly one variant into its wild-type protein sequence, i.e., the 1123 variants yielded 1123 variant protein sequences. The resulting target sequence database is called RefP_V. All of the sequences in RefP_V were reversed to generate decoy protein sequences, which were then concatenated with all of the target sequences for target−decoy searches to estimate the FDR for variant peptide identification.[31]

2.3. Database Search Types and FDR Estimation Approaches for Variant Peptide Identification. Protein database searches are usually performed using search engine(s) for shotgun proteomics data analyses. Similarly, in proteogenomics analyses, either results from a single search engine (SSe) or combined search results from multiple search engines (CMSe) are coupled with global FDR (gFDR) or class-specific FDR (cFDR) estimation for variant peptide identification. Spectra matched to wild-type peptides of reference proteins, termed wild-type peptide−spectrum matches (PSMs), and spectra matched to variant peptides, termed variant PSMs, were combined for 1% gFDR estimation at the PSM level using a statistical validation tool for FDR estimation, called MAYU.[32] In contrast, cFDR estimation used only the variant PSMs and was defined as the number of decoy variant hits divided by the number of target variant hits above a discrimination score threshold to determine the 1% FDR at the PSM level. To be specific, all of the identified target and decoy variant PSMs were ranked in a decreasing order of their discrimination scores calculated by the search engine. The discrimination score threshold was determined based on the above-mentioned cFDR estimation that reaches 1% FDR. In this paper, we consider FDR at the PSM level only; for convenience, we simply term this FDR.

We implemented the above four different combinations of search types and FDR calculations, i.e., SSe-gFDR, SSe-cFDR, CMSe-gFDR, and CMSe-cFDR, using the Comet,[33] MS-GF+,[34] and X!Tandem[35] search engines on the HEK293 cell line data set for protein and variant peptide identification. The workflow is shown in Figure S1. For SSe searching, each of the three search engines was individually used to search against RefP_V database, and its search results were further processed by PeptideProphet,[36] followed by iProphet[37] and then by MAYU. Adopting the SSe-gFDR strategy, we manually extracted target variant PSMs from MAYU-validated PSMs that passed a gFDR

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

of 1% as the result of variant peptide identification. For the SSe-cFDR strategy, we used four scores, including the search score and *E*-value from the respective search engine, and PeptideProphet and iProphet probabilities of variant PSMs to evaluate a cFDR of 1% and then manually retrieved all of the identified variant peptides from variant PSMs with an FDR of 1% based on any of the four scores.

For CMSe searching, search results from each search engine were further processed by PeptideProphet, and the PeptideProphet results from the three search engines were combined by iProphet. Particularly for the CMSe-cFDR strategy, the iProphet probability of variant PSMs was used as the discrimination score for estimating cFDR to acquire PSMs passing a cFDR of 1%, from which we manually extracted the identified variant peptides. In contrast, adopting the CMSe-gFDR strategy, we further used MAYU to process the combined results from iProphet for validation at a gFDR of 1% and manually extracted variant peptides from the PSMs passing MAYU's validation. Identified variant PSMs with an FDR <1% are termed variant PSMs for convenience; similarly, the spectra in the PSMs are termed variant spectra.

2.4. Database Search Parameters. The following search parameters were used in the three search engines: a precursor mass tolerance of ±10 ppm, a fragment mass tolerance of ±0.01 Da, carbamidomethylation of cysteine as a fixed modification, and oxidation of methionine and acetylation of protein N-term as variable modifications. The parameters used for PeptideProphet and MAYU were "-OpdEAP-PPM" and "-P mFDR = 0.01:t -G 0.01 -H 51 -I 2", respectively. In this study, trypsin was considered the protease for digestion.

2.5. Overview of 11-Aspect Examinations of Identified Variant Events at Three Bottom-Up Levels. In this study, we propose a bottom-up trilevel framework of postexaminations of identified variant events from MS spectra at the variant PSM, peptide, and event levels, in which 11-aspect examinations are involved, as shown in Figure S2. First, the foundation PSM level enhances the peptide−spectrum match results via the following five examinations: (1) open modification search, (2) explosive search, (3) combined open modification and explosive search, (4) de novo peptide sequencing, and (5) similarity between a variant spectrum and the predicted spectrum of the corresponding variant peptide. With the first four examinations, we seek to detect dubious PSMs caused by searching an incomplete protein sequence database or neglecting possible modifications; they are designed with increasing search spaces of peptides and modifications, i.e., (1) < (2) < (3) < (4). The last examination compares the fragment ion peaks by checking the similarity between an identified variant spectrum and the predicted spectrum of the identified variant peptide obtained by an MS/MS peak intensity prediction tool because database search tools usually ignore fragment ion intensities in a spectrum.

The middle peptide level disambiguates variant peptide sequences by examining (1) isobaric substitutions and semitryptic cleavage and (2) spectral counting. Variant peptides without the possibility of isobaric substitution or semitryptic cleavage and with multiple PSMs are more reliable. Variant peptides with multiple PSMs are more reliable than those with single PSM.

Lastly, the variant event level at the top confirms variant events by (1) checking for the occurrence of two consecutive *b*-ions or *y*-ions identifying the amino acid variant, (2) checking for the existence of an identified wild-type counterpart peptide, (3) checking that its parental protein is identified, and (4) checking the variant peptide location in the protein when the SAV involves lysine (K) or arginine (R), or proline (P) after K or R.

2.6. PSM-Level Examination. *2.6.1. Open Modification Search Examination.* When using conventional database search tools for searches, to avoid explosions of search space and search time, proteomics researchers usually set only a few modifications in the database search process on an MS/MS data set. However, more than a hundred types of in vivo and in vitro modifications are reported in the Unimod database,[38] which may possibly exist in the samples of MS experiments. To account for false-positive variant PSMs caused by neglecting possible modifications in database searches, we utilized three open modification search (OMS) tools—PIPI,[39] MSFragger,[40] and SpecOMS[41]—with unrestricted modification types when searching against the RefP_V database. When a variant spectrum is identified by an OMS tool as the corresponding variant peptide, the variant PSM is regarded as reliable.

PIPI and SpecOMS mainly allocate the mass of an unexpected but possible modification to an amino acid of a peptide and thus can pinpoint the exact modification site on the peptide. MSFragger enlarges the precursor mass tolerance to 500 Da (default setting) to identify modified peptides but reports neither the modification site nor the modification type in the peptide sequence. To conduct the search, each tool was configured with the following parameters:

(1) PIPI: peptide tolerance = 10 ppm, fragment ion tolerance = 0.01 Da, PTM mass tolerance = ±500 Da, fixed modification = carbamidomethylation (+57), missed cleavages = 2, protease = trypsin, minimum peptide length = 7, maximum peptide length = 50, *q*-value = 0.01.

(2) MSFragger: peptide tolerance = 500 Da, fragment ion tolerance = 0.01 Da, fixed modification = carbamidomethylation (+57), variable modifications = methionine oxidation and acetylation of protein N-term, missed cleavages = 2, protease = trypsin.

(3) SpecOMS: minimum peptide length = 7, maximum peptide length = 50, threshold = 6, max masses count = 100, minimum peptide charge = 1, maximum peptide charge = 7, number of decimals = 2, decimal value = 4, decoy base = false.

2.6.2. Explosive Search Examination with the Human-Associated Tryptic Peptide Database (SuperPep_V). In contrast to open modification searching against RefP_V with large PTM or peptide tolerance, explosive search conducts searches against a huge human tryptic peptide database containing peptides from complete human proteome and variants; this database is called SuperPep_V and is much larger than the peptide space of RefP_V. We examine the consistency between variant PSMs by searching the RefP_V and the huge SuperPep_V databases.

To construct SuperPep_V for explosive search, we needed to select very comprehensive protein or peptide sequence databases and variant databases. To the best of our knowledge, the PeptideAtlas Mapping Database (PAmap)[42] and the dbSAP[43] database provide comprehensive human protein sequences and single amino acid variants, respectively. The PAmap database integrates proteomics-based protein sequences (such as sequences in Swiss-Prot, TrEMBL, and neXtProt[44]), genomics-based protein sequences (such as sequences in Ensembl and RefSeq), SAVs listed in neXtProt but excluding those in COSMIC,[45] as well as human-associated microbiome

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

and nonhuman contaminant protein sequences. The dbSAP database combines sequence variant annotations from public databases such as dbSNP,[46] COSMIC, UniProt, HPMD,[47] MS-CanProVar,[48] and Ensembl. Therefore, we collected the human and associated protein sequences from the PAmap and dbSAP databases and subsequently concatenated them with the RefP_V database. For this concatenated database, we performed in silico trypsin digestion, following the cleavage specificity rules (cleaving K or R not before P at the C-terminal) provided in Keil's rules[49] and allowing up to two missed cleavage sites. After in silico digestion, all nonredundant tryptic peptides, including tryptic variant peptides, of 7−50 amino acids were collected to construct the SuperPep_V database for explosive search. As a result, SuperPep_V contained 77,551,699 unique tryptic peptides, a file of 3.5 GB. Each tryptic peptide entry in the database was recorded in the FASTA format, which is suitable for database searches. Notably, compared with the peptide space of RefP_V, the SuperPep_V database includes an approximately 30-fold increase in the number of unique peptides. The peptide length distributions of RefP_V and SuperPep_V are shown in Figure S3; the number of peptides of each length in SuperPep_V is at least seven times as much as that in RefP_V. Thus, the SuperPep_V database is suitable for explosive search examination.

To perform explosive search on the variant spectra obtained by searching against RefP_V, we used Comet, MS-GF+, and X! Tandem to search the spectra against the SuperPep_V database. The search parameters of each search engine were the same as those described in Section 2.4. The search results of the explosive search were then compared with the variant PSMs obtained from searching against RefP_V. For the comparison, we used the original search score, instead of *E*-value, from each search engine for the following two reasons. First, the peptide search spaces of SuperPep_V and RefP_V can affect the *E*-values of a spectrum−peptide pair. Second, the original search score from a search engine represents the similarity between a variant spectrum and a theoretical spectrum generated from a specific peptide in the database and thus is unlikely to be changed for a given spectrum−peptide pair regardless of which database the peptide is from. For results from each search engine, we compared the search scores and the matched peptides of variant PSMs. Originally identified variant PSMs having explosive search supports were regarded as reliable.

2.6.3. Examination of Combined Open Modification and Explosive Search. We further propose an examination of OMS combined with explosive search, i.e., using the three OMS tools (MSFragger, PIPI, and SpecOMS) to search against the SuperPep_V database to examine the reliability of variant PSMs.

Due to the large file size of SuperPep_V (3.5 GB), some OMS tools could not successfully perform searches caused by insufficient memory when using a computing server with 64G RAM. We propose solving this problem using a strategy of searching specific scans against specific FASTA to reduce the FASTA file size but maintain a proper peptide search space. First, we classified the variant spectra into four spectral groups (denoted as SG) based on the length of their corresponding variant peptides: 7−15 amino acids (a.a.) (SG-A), 16−24 a.a. (SG-B), 25−35 a.a. (SG-C), and 36−44 a.a. (SG-D). Considering the 500 Da precursor tolerance (approximately ± 5 a.a.) for open modification searches, we divided SuperPep_V by peptide length into four FASTA sets (denoted as PepS) as search spaces for spectra groups as follows: 7−20 a.a. (PepS-A for SG-A searching), 11−30 a.a. (PepS-B for SG-B searching), 20−40 a.a. (PepS-C for SG-C searching), and 30−50 a.a. (PepS-D for SG-D searching). Table S1 in the Supporting Information shows the numbers of spectra and peptides in the four spectral groups and four peptide FASTA sets, respectively. The search parameters of the three OMS tools were the same as those described above in the OMS examination.

2.6.4. De Novo Peptide Sequencing Examination. De novo peptide sequencing is an alternative method to interpret the peptide sequence directly from an MS/MS spectrum without using a sequence database but based on the mass difference of consecutive peaks. It is a good basis by which to verify variant PSMs. To this end, we adopted three common de novo peptide sequencing tools—PepNovo+,[50] pNovo+,[51] and PEAKS[52]—to examine the interpretation consistency of variant spectra obtained from sequence database searching and from de novo peptide sequencing. The variant spectra with consistent interpretations between the two approaches are regarded as reliable identification of variant peptides.

The general parameters used in de novo peptide sequencing tools were a precursor mass tolerance of ±10 ppm, a fragment mass tolerance of ±0.01 Da, carbamidomethylation of cysteine as a fixed modification, oxidation of methionine as a variable modification, and output of the top 10 rank hits for each spectrum. Furthermore, we used DeNovoGUI's (version 1.15)[53] built-in "peptide matches" function to reversely match the top 10 hits of each spectrum from de novo sequencing tools to peptide sequences of the RefP_V database. This resulted in the peptide sequences in RefP_V yielded by de novo sequencing tools, which were used to examine the interpretation consistency between sequence database searches and de novo peptide sequencing.

2.6.5. Examination of Similarity between Variant Spectra and Predicted MS/MS Spectra. The peptide−spectrum matching of the conventional database search approach usually does not consider peak intensities in experimental spectra. We thus proposed an examination that considers peak intensity to evaluate the reliability of variant spectra. For each variant spectrum, we obtained its corresponding variant peptide and charge state and used MS^{2}PIP[54] and MS^{2}PBPI[55] to generate a predicted MS/MS spectrum for the variant peptide sequence and charge state. Then, we examined the cosine similarity between the variant and predicted spectra.

To determine the similarity score threshold for a variant-predicted spectral pair to be similar (i.e., a reliable variant spectrum supported by the predicted spectrum), we adopted a target−decoy strategy in which similarity scores were calculated between all pairs of a variant spectrum and a predicted spectrum (workflow illustrated in Figure S4). We defined the similarity score of a correct pair of spectra, i.e., having the same peptide sequence and charge state, as a matched-pair score, which represents the score of a target case; otherwise, it was defined as a mismatched-pair score, the score of a decoy case. Note that we excluded mismatched pairs of spectra that corresponded to a peptide and its truncated sequence because both peptide sequences have overlapping *b*- and *y*-ions and similar intensity patterns, possibly affecting the distribution of similarity scores; for example, ALMDEGMK and ALMDEGMKEK have at least 70% *b*- and *y*-ions in common. Then, all matched- and mismatched-pair scores form two respective distributions to estimate the threshold to justify the confidence of all of the matched pairs. For a similarity score *s*, we defined the corresponding FDR as the number of mismatched pairs with a similarity score less than or equal to *s* divided by the number of

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

matched pairs having a similarity score less than or equal to *s*. Then, we determined the similarity score threshold as the score *s** that yields an FDR of 5%. Thus, all of the matched pairs that pass the specified threshold are considered high-confidence results, i.e., the corresponding variant spectra identifying variant peptide sequences are reliable, as they are supported by the predicted MS/MS spectra from state-of-the-art tools.

2.7. Peptide-Level Examination. *2.7.1. Isobaric Substitution and Semitryptic Cleavage Checks.* As our previous study[17] showed that more variant PSMs can be interpreted as wild-type peptides of proteins when considering 11 types of isobaric substitution, for each variant PSM, we suggested verifying whether the variant peptide can be obtained by isobaric substitution of a wild-type peptide. If yes, the identified variant peptide is not reliable. In this study, we conducted an examination of isobaric substitution and semitryptic cleavage on the RefP_V database to verify the identified variant peptides. The detailed algorithm for checking the 11 isobaric substitutions is described in Choong et al.[17] We also performed semitryptic cleavage examination to determine whether a variant peptide can be derived from semicleavage or terminal truncation of any wild-type peptide. To check, we first performed in silico trypsin digestion on the RefP_V database with two missed cleavages and then examined whether each variant peptide sequence maps to a subsequence of any wild-type tryptic peptide.

2.7.2. Spectral Counting. Spectral counting is widely used in label-free quantification; it involves counting the number of identified PSMs of a given peptide and integrating all of the numbers for the identified peptides in a protein to represent the protein abundance. Popular tools such as Scaffold,[56] CRUX,[57] and Proteome Discoverer adopt this approach for label-free quantification. Some proteins in an experiment are possibly identified by a single PSM. However, such single PSMs may be false positives even though they pass the FDR control, casting doubt on the proteins of such one-hit wonders. Using a similar concept, we stratified the evidence levels of variant peptides based on their spectral counts as follows: one PSM, two PSMs, and multiple ($\geq$3) PSMs to classify variant peptides, denoted by 1_PSM, 2_PSM, and 3up_PSM, respectively. Variant peptides having more PSMs are regarded as having greater confidence.

2.8. Variant Event-Level Examination. *2.8.1. Checking Consecutive Fragment Ion Peaks.* Checking the mass difference of consecutive fragment ions in an MS/MS spectrum is common in proteomics data analysis, for instance, de novo peptide sequencing, PTM site localization, and glycopeptide identification.[50,58,59] In Ivanov et al.,[60] the analysis of consecutive *y*-ions in an MS/MS spectrum is used to confirm the existence of variant sites in the variant peptide sequence. Therefore, we adopted a similar concept to examine the existence of consecutive variant site-specific *b*- or *y*-ions in variant spectra. If a variant spectrum reveals two consecutive variant site-specific *b*- or *y*-ions with a mass difference corresponding to the variant residue, the variant event can be confirmed; otherwise, the event cannot be confirmed, although the variant PSM and peptide sequence can be still correct.

For each variant spectrum, we converted the spectrum information into a pNovo+ output format and loaded the converted spectrum to a spectrum viewer, a built-in function of DeNovoGUI to generate spectrum−peptide annotation. The spectrum viewer parameter settings only consider matching *b*- and *y*-ions of charge +1 to spectrum peaks that have intensities at least 10% of the most intense peak and are within 0.02 Da tolerance. Then, we manually examined the spectrum annotation of each variant peptide−spectrum to confirm the existence and mass difference of the two consecutive variant site-specific *b*- or *y*-ions of the variant peptide.

2.8.2. Checking the Existence of Wild-Type Counterparts in PeptideAtlas and the Identification of Its Parental Protein. To enhance the reliability of the variant events derived from variant PSMs, we inspected the existence of their wild-type counterparts and parental proteins. For each variant site, we adopted the LeTE-fusion pipeline proposed in Mamie Lih et al.[61] to check whether its wild-type counterpart peptides, fully digested or miscleaved, could be found in PeptideAtlas (build human Jan 2018),[62] a public repository of experimentally observed peptides. If a variant event could not find any wild-type counterpart peptide in PeptideAtlas, the identification of this variant event was doubtful. To check the existence of its parental protein, we used MAYU to perform validation with an FDR of 1% at the PSM, peptide, and protein levels on the CMSe-gFDR search results. If the parental protein of the variant event did not pass the protein-level FDR, identification of this variant event was dubious. Only variant events with evidence of wild-type counterparts and parental proteins were regarded as reliable identification results.

2.8.3. Variant Peptide Location in the Protein. Variant peptides are identified from variant-harboring protein sequences with protease digestion in sequence database searches. Notably, when an amino acid is mutated to K or R, in silico trypsin digestion of SAV-harboring protein may generate peptides, which are regarded as variant peptides by database search engines, but they do not include the actual variant site. For example, in Q9BUP3, the fully digested wild-type peptide containing position 197 is 186-KFFGSLPDSWA**S**GHSVPV-VTVVR-207. For its SAV S197R, the SAV-harboring protein contained in the customized database for database searches yields fully digested peptides 187-FFGSLPDSWA**R**-197 and 198-GHSVPVVTVVR-207, neither of which belong to the wild-type peptide space and thus are regarded as variant peptides by database search engines. The former peptide is indeed a variant peptide, but the latter peptide does not contain the variant site, and its identification cannot be used to support the identification of the S197R variant event because this peptide can come from peptide truncation or semitryptic digestion of a wild-type peptide. Similarly, because a proline (P) after K or R may render a missed cleavage, an SAV involving P, which is after K or R, mutated to another amino acid can also lead to a misclassified variant peptide. Thus, it is essential to exclude variant peptides that do not contain the variant event.

3. RESULTS AND DISCUSSION

3.1. Maximizing the Number of Identified Variant PSMs of the HEK293 Cell Line Data Set. To demonstrate 11 postevaluation examinations on identified variant peptides, we used the HEK293 cell line MS data set to obtain as many variant PSMs with an FDR of 1% as possible. In this proteogenomics study, we applied SSe-gFDR, SSe-cFDR, CMSe-gFDR, and CMSe-cFDR, the four combinations of database search type and FDR estimation approach, for database searches.

To explore the impact of the above four strategies, we compared the identification results of different search types under the same FDR estimation approach. To compare SSe-gFDR and CMSe-gFDR, we used Comet, MS-GF+, and X! Tandem for database searches and obtained 227, 252, and 272 variant PSMs passing a gFDR of 1%, respectively, that identified 78 (68 variant events), 85 (76), and 89 (80) variant peptides. In

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

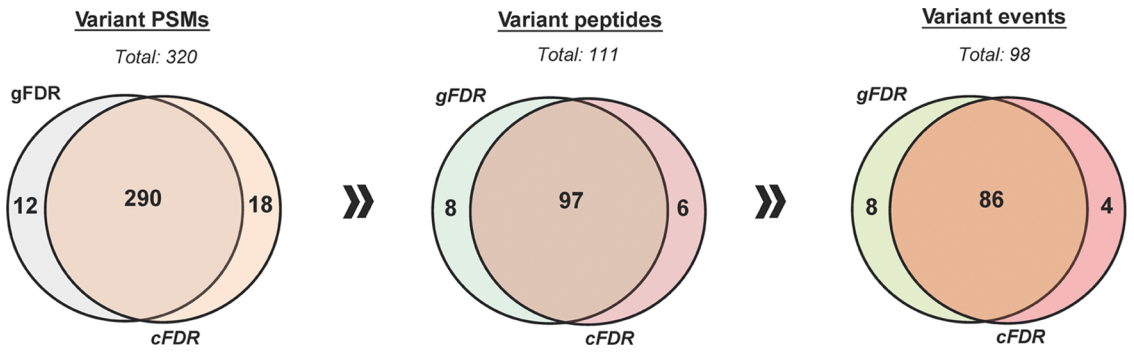


Figure 1. Variant peptide identification results obtained by using the four strategies of combining the type of database search and FDR estimation approach (SSe-gFDR, SSe-cFDR, CMSe-gFDR, and CMSe-cFDR) on the HEK293 cell line MS data set.

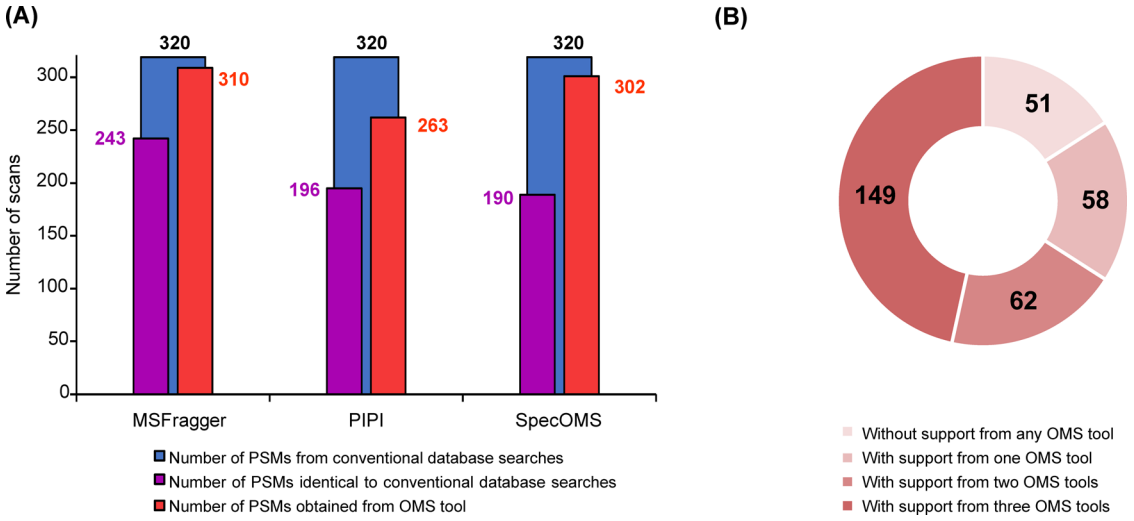


Figure 2. Results of open modification search (OMS) on 320 variant PSMs obtained from conventional database searches. We used three OMS tools to perform the search against the RefV_P database. (A) Comparison of the consistency of identified variant PSMs between OMS and conventional database searches. (B) Distribution of 320 variant PSMs supported by 0−3 OMS tools.

contrast, using the CMSe approach yielded 282 variant PSMs passing a gFDR of 1% that corresponded to 93 variant peptides with 82 variant events, more than using the SSe approach. Overall, a total of 302 variant PSMs corresponding to 105 variant peptides with 94 variant events passing a gFDR of 1% were obtained from the SSe-gFDR and CMSe-gFDR strategies. The results are summarized in Table S2. Furthermore, we used a Venn diagram to illustrate the relationship of variant PSMs, variant peptides, and variant events among CMSe-gFDR, Comet-gFDR, MS-GF+-gFDR, and X!Tandem-gFDR strategies in Figure S5A. Using the SSe-gFDR strategy, 5.3% (12/227), 4.4% (11/252), and 11.4% (31/272) variant PSMs were exclusively obtained by Comet, MS-GF+, and X!Tandem, respectively. Notably, 1.8% (4/227), 2.8% (7/252), and 4.0% (11/272) of the PSMs obtained by Comet, MS-GF+, and X!Tandem, respectively, were not included in the PSMs obtained using the CMSe-gFDR strategy. This shows that the variant PSMs resulting from the use of SSe-gFDR and CMSe-gFDR are slightly different, likewise for variant peptides and variant events. More detailed results about the diversity of SSe-gFDR versus CMSe-gFDR are provided in Table S3.

To compare the SSe-cFDR and CMSe-cFDR strategies, we followed the cFDR estimation methods described in Section 2.3. The numbers of identified PSMs, variant peptides, and variant events obtained from SSe-cFDR and CMSe-cFDR are listed in Table S2. Specifically, using CMSe-cFDR, Comet-cFDR, MS-GF+-cFDR, and X!Tandem-cFDR, we obtained 279, 227, 257, and 273 variant PSMs, respectively, and 308 variant PSMs in total, passing a cFDR of 1% from any score cutoff. Based on the 308 PSMs, 103 variant peptides and 90 variant events were identified. Furthermore, the differences in the identification results of SSe-cFDR and CMSe-cFDR are shown in Figure S5B and Table S4. The above results reflect a similar phenomenon to that for the comparison between SSe-gFDR and CMSe-gFDR and suggest that using different search engines and different discrimination scores for cFDR estimation yields slightly different results. In summary, observing the above comparisons of using the four strategies (Figure S5), CMSe is suggested for database searching. When adopting the CMSe strategy, using gFDR can obtain more variant PSMs than using cFDR.

Finally, using the four strategies, we identified a total of 320 unique variant PSMs, corresponding to 111 variant peptides and 98 variant events (Figure 1), where 302 PSMs and 308 PSMs were obtained from both search types using gFDR and cFDR, respectively. All 320 variant PSMs were used in this study to demonstrate 11-aspect examinations to evaluate their reliability at the PSM, peptide, and variant event levels (Table S5).

3.2. Examinations of Variant PSMs of the HEK293 Data Set at the PSM Level. *3.2.1. Examination by Open Modification Search (OMS).* To investigate the reliability of all 320 identified variant PSMs, we first compared them with the results of the OMS tools; the results are shown in Figure 2 and

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

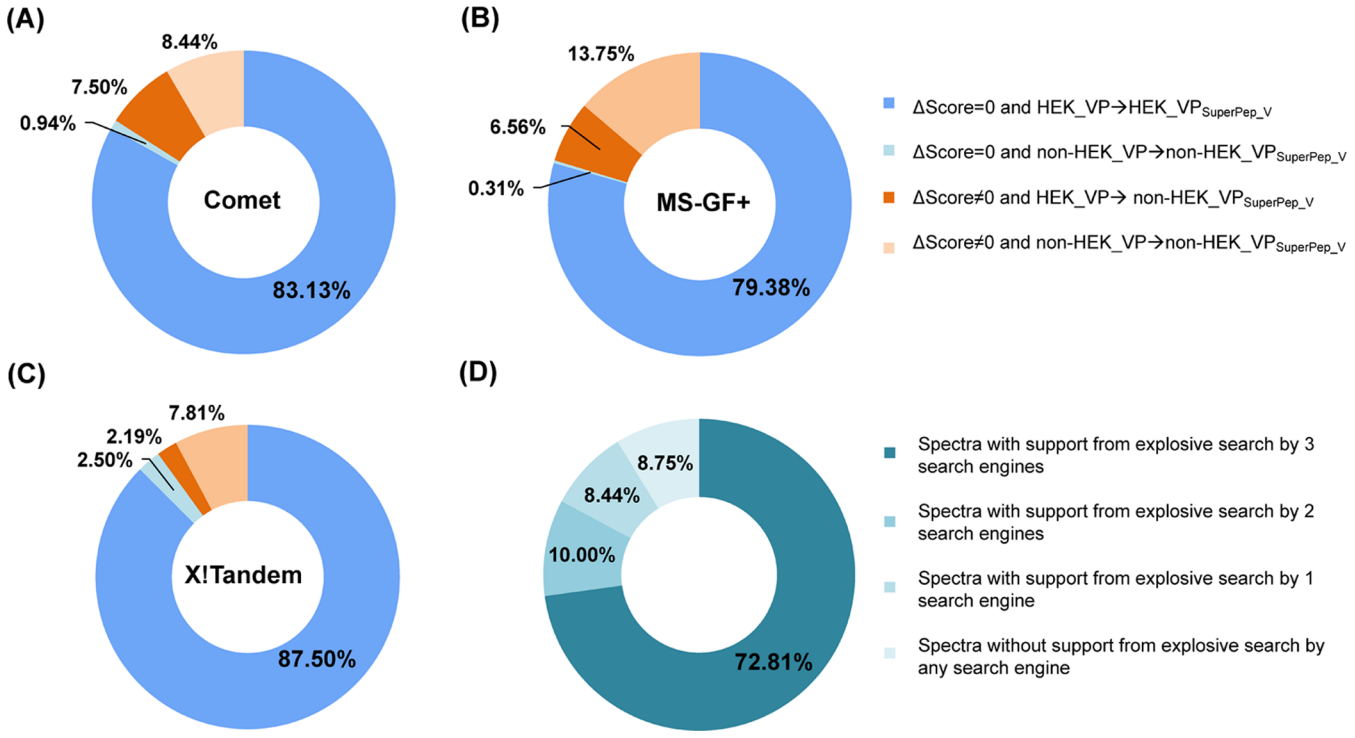


Figure 3. Results of explosive search on 320 identified variant PSMs. We used three conventional database search engines to search the 320 variant spectra against the huge SuperPep_V database. (A−C) Based on each search engine's results, classifying the 320 variant spectra into four groups. (D) Distribution of 320 variant PSMs with support from explosive search by 0−3 search engines.

Table S6. Of the 320 spectra, the three OMS tools reported 190−243 spectra having results consistent with conventional protein database searches (Figure 2A). In other words, 24.1−40.6% of the 320 variant PSMs were not supported by the three respective OMS tools. Notably, of the 320 variant PSMs, 149 (46.6%), 62 (19.4%), and 58 (18.1%) PSMs were supported by all three, two, and one OMS tools, respectively. However, the remaining 51 (15.9%) variant spectra were not supported by any OMS tool, i.e., they were unidentified or mostly identified as other wild-type peptides with modifications (Figure 2B). Thus, these 51 variant PSMs were very likely unreliable identification results, as verified by the three OMS tools.

3.2.2. Examination of Explosive Search against the Huge SuperPep_V Database. After using Comet, MS-GF+, and X! Tandem to search the 320 variant spectra against SuperPep_V, we extracted the top PSMs and their search scores of the 320 spectra from the search results (.t.xml, pep.xml, .mzid) of each search engine. For a search engine, each spectrum had two PSM results and their search scores obtained from the original search (searching against RefP_V) and the explosive search, respectively. For each spectrum, we compared its identified peptides of the two PSMs and the respective search scores. If both search scores were the same, then it was denoted as ΔScore = 0, and ΔScore ≠ 0 otherwise. The two associated identified peptides could be variant peptide(s) of the HEK293 cells, termed HEK_VP, or any peptide(s) other than the HEK293 variant peptides from RefP_V or SuperPep_V, termed non-HEK_VP. Then, we classified the 320 spectra according to the comparison results into the following four groups:

(1) Both PSMs have the same search score and identify the same variant peptide: denoted as ΔScore = 0 and HEK_VP → $HEK\_VP_{SuperPep\_V}$;

(2) Both PSMs have the same search score and identify non-HEK_VP peptides: denoted as ΔScore = 0 and non-HEK_VP → $non\text{-}HEK\_VP_{SuperPep\_V}$;

(3) Both PSMs have different search scores, and one identifies HEK_VP but the other does not: denoted as ΔScore ≠ 0 and HEK_VP → $non\text{-}HEK\_VP_{SuperPep\_V}$;

(4) Both PSMs have different search scores and identify different non-HEK_VPs: denoted as ΔScore ≠ 0 and non-HEK_VP → $non\text{-}HEK\_VP_{SuperPep\_V}$.

Note that the 320 variant spectra were obtained by integrating results from the three single search engines and those using multiple search engines. The second and fourth groups with non-HEK_VP → $non\text{-}HEK\_VP_{SuperPep\_V}$ could occur when the search engine identified the spectrum as a non-HEK_VP but at least one of the other two search engines identified it as HEK_VP.

Classifications of the 320 variant spectra using Comet, MS-GF+, and X!Tandem are shown in Figure 3A−C and Table S7. Of the 320 spectra, 266 (83.13%), 254 (79.38%), and 280 (87.50%) spectra are in Group 1, ΔScore = 0, and HEK_VP → $HEK\_VP_{SuperPep\_V}$, as determined by using Comet, MS-GF+, and X!Tandem for searches, respectively. In other words, the spectra in group 1, the majority of the 320 spectra, still matched the same variant peptide even when searching against the huge SuperPep_V database, and thus their identified variant peptides were high confidence. In addition, the spectra in the second and fourth groups (light blue and bisque color in Figure 3A−C) accounted for 9.38−14.06% of the 320 spectra for each search engine. Such spectra were not assigned to variant peptides by the specific search engine when searching against RefP_V and against SuperPep_V, but they were assigned to variant peptides by other search engine(s) and could represent medium-confidence variant PSMs. In contrast, 7−24 (2.19−7.50%)

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

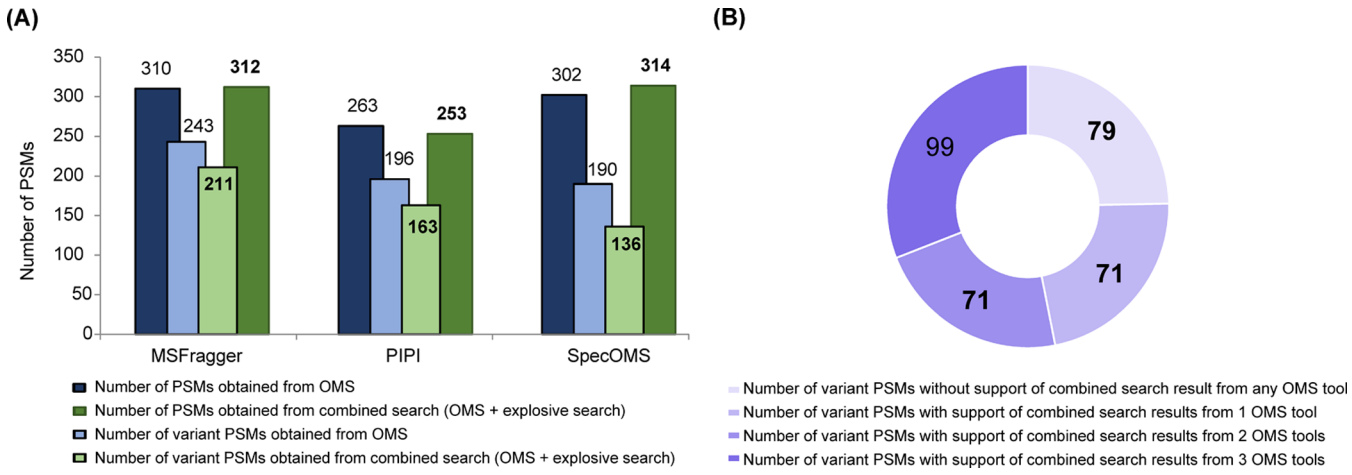


Figure 4. Results of combined open modification and explosive search on 320 variant PSMs. Three OMS tools were used to search the 320 variant spectra against the huge SuperPep_V database. (A) Comparison of identified PSMs and variant PSMs between OMS and combined search (OMS + explosive search) results for each OMS tool. (B) Distribution of 320 variant PSMs with the support of combined open modification and explosive search results from 0 to 3 OMS tools.

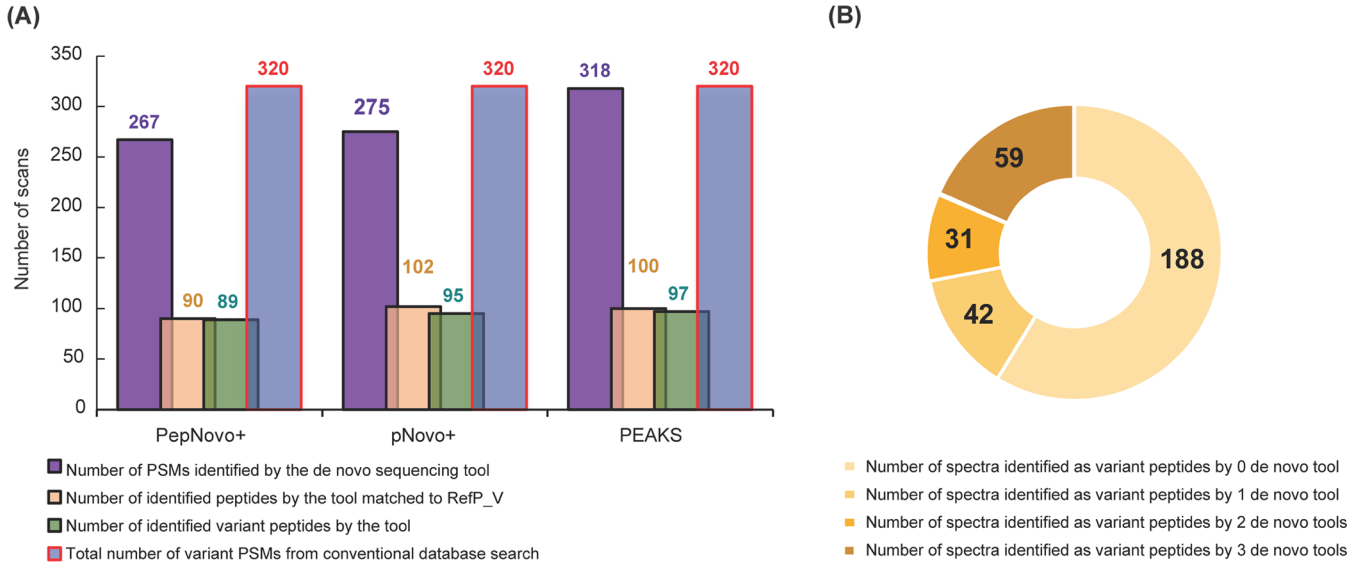


Figure 5. De novo peptide sequencing on the 320 variant spectra by PepNovo+, pNovo, and PEAKS. (A) Identification results of three de novo sequencing tools based on the top 10 hits of each of the 320 spectra. (B) Distribution of 320 variant PSMs with support from de novo sequencing by 0−3 de novo sequencing tools.

spectra in the third group (dark orange color in Figure 3A−C), ΔScore ≠ 0 and HEK_VP → non-HEK_VP$_{SuperPep_V}$, of the respective search engine, were identified as variant peptides of HEK293 cells but not when searching against SuperPep_V. Notably, this also suggests that some of these spectra may be misinterpreted as a variant peptide due to an incomplete search space.

Furthermore, a total of 292 spectra were classified as group 1 by any of the search engines; the confidence levels of their identified variant peptides are further summarized as follows (Figure 3D). Of the 320 variant spectra, 233 (72.81%) spectra were identified as variant peptides when searching SuperPep_V by all of the three search engines, and these variant peptides were considered to have the highest confidence. Moreover, 32 (10%) and 27 (8.44%) spectra were identified as variant peptides in explosive search by two and one search engines, respectively; these variant peptides are regarded to have high and medium confidence, respectively. The remaining 28 (8.75%) spectra were consistently annotated as non-HEK_VP in explosive search by the three search engines, although they were identified as variant peptides in the original search by some search engines. The results show that the explosive search examination is an effective method to inspect possible false-positive variant peptides caused by using an incomplete database for searches.

3.2.3. Examination by Combined Open Modification and Explosive Search. Using OMS tools, MSFragger, PIPI, and SpecOMS to search the 320 variant spectra obtained from the original search against SuperPep_V yielded 312, 253, and 314 PSMs, respectively, as shown in Figure 4A and Table S8. Of the PSMs of the combined open modification and explosive search results, 211, 163, and 136 spectra were identified as variant peptides by the three OMS tools (light green bar), respectively, consistent with the original search, and were considered high-confidence variant PSMs. Moreover, OMS against SuperPep_V (light green bar), i.e., the combined search, had a decrease of 13.2−28.4% variant PSMs compared with OMS against RefP_V

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

(light blue bar). Note that a number of spectra were not matched to variant peptides in this combined search with open modifications and a huge peptide database as the search space. Thus, the variant peptide identification of these spectra was insufficiently confirmed.

Furthermore, we integrated the combined search results of the 320 spectra to classify the reliability of the originally identified variant PSMs (Figure 4B). Of the 320 spectra, 99, 71, and 71 spectra were assigned to variant peptides by three, two, and only one OMS tools, respectively. The remaining 79 spectra were unidentified or assigned to nonvariant peptides by all of the three OMS tools. Thus, these 79 spectra of variant PSMs could be considered to be less reliable results caused by possible modifications and incomplete peptide search space.

3.2.4. De Novo Peptide Sequencing Examination. We used three de novo peptide sequencing tools—PepNovo+, pNovo+, and PEAKS—to identify the 320 variant spectra obtained from database searches (Table S9). Of the 320 spectra, the top 10 hits of 89, 95, and 97 spectra obtained from PepNovo+, pNovo+, and PEAKS (green bar), respectively, contained the corresponding variant peptides (Figure 5A), of which more than 79.8% belonged to rank-1 or -2 hits from the three tools (Figure S6). Integrating the results of the three tools to examine the interpretation consistency of the identified spectra, we observed that of the 320 spectra, 59 (18.4%), 31 (9.7%), and 42 (13.1%) spectra had support from three, two, and one tools, respectively, as shown in Figure 5B. However, 188 (58.8%) spectra had no support from any tool and could be considered to be unreliable identification results.

3.2.5. Similarity Examination by Predicted MS/MS Spectra. We used the prediction tools MS^{2}PIP and MS^{2}PBPI to generate the simulated MS/MS spectra of 111 identified variant peptide sequences while considering charge states and modifications. We then used the target−decoy approach described in Section 2.6.5 to determine the similarity score threshold between variant spectra and simulated spectra for the confidence results. As a result, the similarity score thresholds between a variant spectrum and the corresponding predicted spectrum to be considered similar at an FDR less than 5% were set to 0.5 and 0.4 for MS^{2}PIP and MS^{2}PBPI, respectively, as shown in Figure S7; i.e., a variant spectrum yielding a similarity score with the predicted spectrum higher than the threshold was regarded as a high-confidence variant PSM. Of the 320 variant spectra obtained from sequence database searches, 179 and 111 of the variant spectra passed the similarity score thresholds of MS^{2}PIP and MS^{2}PBPI, respectively, and were considered to be high-confidence identification results (Table S10). Furthermore, 108 (33.75%) and 74 (23.125%) spectra were verified by both and only one tools, respectively (Figure 6). The remaining 138 (43.125%) spectra had no evidence support from any tool and were likely to be less reliable results.

3.3. Examinations of Variant PSMs of the HEK293 Data Set at the Peptide Level. *3.3.1. Isobaric Substitution and Semitryptic Check.* For the 111 variant peptides identified in 320 variant PSMs, we conducted examinations of isobaric substitution and semitryptic cleavage on the RefP_V database to further verify the variant peptides. Of the 111 variant peptides, 3 variant peptides could be derived from reference peptides with isobaric substitutions, as shown in Table S11. For example, the PSM of the variant peptide "SLVQESLSTNSSDLVAPSPDAFR" of protein Q12888 with the D353E variant could be interpreted as "SLVQD[methylated]SLSTNSSDLVAPSPDAFR" because the masses of glutamate and methylated aspartate are equivalent. Also, we found a variant peptide "TQDLLNQNHSANAVR.L" of protein Q9H040 with the P296L variant possibly belonging to a reference tryptic peptide "TQDLLNQNHSANAVR.PNSK" of Q9H040 by semitryptic cleavage checking (Table S11). However, this case is not an actual semicleavage or terminal-truncation case because the P is mutated to L although trypsin retains low cleavage specificity at K or R before P.[63] Therefore, these 4 out of the 111 variant peptides were ambiguous results at the peptide sequence level but not at the PSM level.

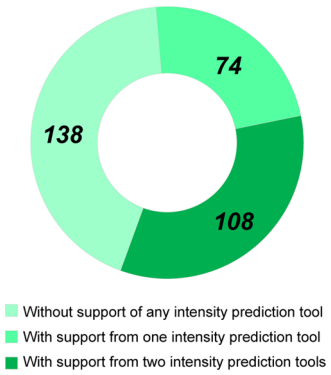


Figure 6. Distribution of 320 variant spectra with support by spectral similarity with predicted MS/MS spectra generated by 0−2 intensity prediction tools. MS^{2}PIP and MS^{2}PBPI were used to generate simulated spectra based on variant peptide sequence and charge state. A variant spectrum has the support of a prediction tool if the spectrum yields a similarity score with the predicted spectrum higher than the threshold.

3.3.2. Spectral Counting Evaluation. Following the spectral counting described in Section 2.7.2, we grouped the 111 variant peptides in the 320 variant PSMs into three peptide classes: 1_PSM, 2_PSM, and 3up_PSM, which contained 54 (48.65%), 21, and 36 variant peptides, respectively (Figure 7A). Note that 48.65% of the variant peptides in the 1_PSM class were less confident than the other two peptide classes. Next, we grouped identified variant events into three event classes based on the numbers of peptides and PSMs detecting the event, defined as follows: a variant event detected by a single peptide and a single PSM (denoted as 1P_1PSM), a single peptide and multiple (≥2) PSMs (1P_mPSM), and multiple (≥2) peptides and multiple PSMs (mP_mPSM). Of the 98 variant events, 46, 40, and 12 variant events were grouped in the 1P_1PSM, 1P_mPSM, and mP_mPSM classes, respectively (as shown in Figure 7B), where 46.94% of variant events were less confident than the other two event classes. The hierarchical visualization of the variant event classification based on peptide and spectral counting is shown in Figure S8; detailed results are provided in Table S12.

3.4. Variant Event Examination on the HEK293 Data Set via Four Different Checks. *3.4.1. Checking Consecutive Fragment Ion Peaks.* Following the procedure described in Section 2.8.1, we manually examined the spectrum annotation of each of the 320 variant spectra to confirm the existence and mass difference of the two consecutive variant site-specific *b*- or *y*-ions. The detailed information is listed in Table S13.

Of the 320 variant PSMs, 161 spectra contained two consecutive variant site-specific fragment ions and more *y*-ion pairs, as shown in Figure 7C. Moreover, of the 98 variant events derived from the 320 PSMs, 47 variant events were confirmed by consecutive fragment ions in MS/MS spectra (Figure 7D).

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

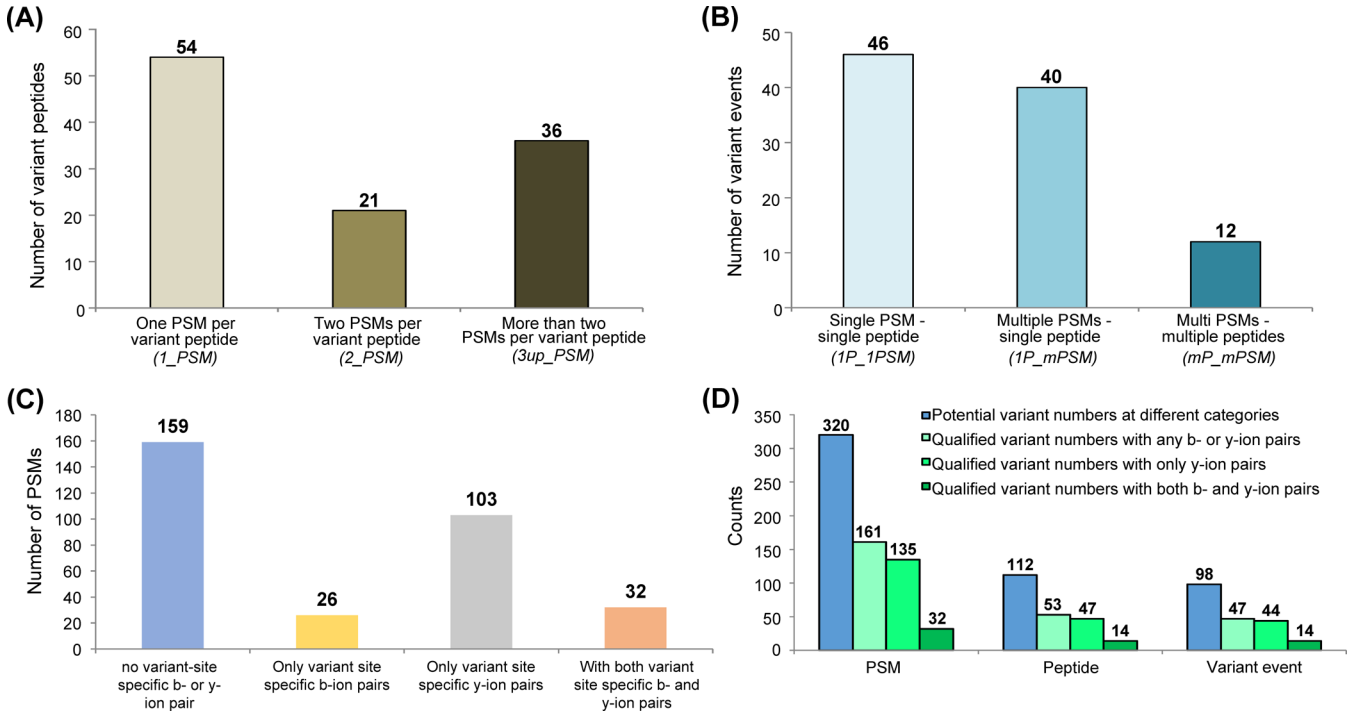


Figure 7. Evaluation of 320 variant PSMs at the peptide and variant event levels. The 320 variant PSMs corresponded to 111 variant peptides and 98 variant events. Spectral counting evaluation of (A) 111 variant peptides and (B) 98 variant events. Checking two consecutive variant site-specific fragment ions (C) in 320 variant spectra and (D) from PSM, peptide, and variant event perspectives.

These results show that 48% of variant events are more reliable, as confirmed by the consecutive variant site-specific *b*- or *y*-ions.

3.4.2. Checking Variant Peptide Sequence Location. We examined whether the 111 variant peptides in the 320 variant PSMs contained the SAVs. As explained in Section 2.8.3, 198-GHSVPVVTVVR-207 cannot be regarded as a variant peptide with S197R in Q9BUP3 and is indeed a wild-type peptide, although it is not a digested peptide of the reference protein. Of these variant peptides, we found that four peptides do not contain the variant sites because the variant sites are trypsin cleavage sites or occur after the C-terminal of the trypsin cleavage site, as shown in Table S14. Of the 98 variant events identified in the 320 variant PSMs, 96 variant events were supported by variant-site-containing peptides and thus were more reliable than the remaining two variant events, each of which was reported in only one peptide but without containing the variant site.

3.4.3. Checking the Existence of Wild-Type Counterparts in PeptideAtlas and the Identification of Parental Proteins. Of the 98 variant events in the 320 variant PSMs, for 70 events, their wild-type counterpart peptides were found in PeptideAtlas; for the remaining 28 events, no such peptides were found (Table S15). Based on the CMSe-gFDR results, the parental proteins of 93 variant events passed an FDR of 1% at the protein level (Table S16). By inspecting the existence of wild-type counterpart peptides and parental proteins, we found that the 98 variant events in the 320 variant PSMs reflected different levels of reliability.

4. DISCUSSION

4.1. Applying the 11-Aspect Examinations to Extract the Highest-Confidence Variant Events. In the HEK293 cell line MS data set, using both search strategies, 375 variant PSMs corresponding to 135 variant peptides and 121 variant events were reported. After applying FDR filtering, 320 variant PSMs passed an FDR <1%. It showed that FDR filtering removed 55 (14.9%) PSMs, 24 (17.8%) peptides, and 23 (19.0%) events. To identify variant events with the highest confidence, we applied all of the 11-aspect examinations to the 320 variant PSMs containing 98 variant events (Table S17). Of these, 111 PSMs passed all of the five-aspect examinations at the PSM level, where each aspect examination results from at least one tool supporting the reliability of variant PSMs. When further applying the two-aspect examinations at the peptide level on the 111 variant PSMs, 25 variant peptides satisfied both evaluations, i.e., ambiguous variant peptides caused by isobaric substitutions or semitryptic cleavage, and variant peptides belonging to 1P_1PSM class were filtered out. Finally, 14 variant events of the 17 variant peptides passed the four evaluations at the variant event level, where each variant peptide was verified by at least one pair of site-specific *b*- or *y*-ions, the existence of a wild-type counterpart peptide and parental protein, and its sequence containing the variant site. With such stringent examinations at the PSM, peptide, and variant event levels, 14 variant events contained in 17 variant peptides of 71 variant PSMs in the HEK293 data set achieved the highest confidence.

We noted that 249 (77.8%) out of 320 variant PSMs were filtered out by the 11 postexaminations. Such a high filtration rate was also observed in the nine deep proteome data sets of cancer cell lines after applying two filtering strategies—filtration against reference proteome and chemical modification filter—to filter out unreliable identified variant peptides, as reported by Alfaro et al.[12] based on their Additional File 4 (CL9_exom_snv). To be specific, in their nine deep proteome data sets, on average, 668 variant PSMs were reported. After filtering, on average, 587 (87.9%) variant PSMs were filtered out and only 81 (12.1%) variant PSMs remained. It again emphasized the necessity of checking the reliability of identified variant PSMs

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

with 1% FDR. Furthermore, to validate the results after 11 postexaminations, we used Lobas et al.'s[64] results of three HEK293 MS data sets from three different sources using two search engines. The authors classified identified variant peptides into four levels of confidence by the number of their confirmations. All of our 17 variant peptides passing the 11 postexaminations were found in their identification results. Notably, 10 (71.4%) variant events were found in at least two MS data sets and regarded as more confident identification. The high overlapping rate of variant events showed that our postexaminations are reliable and effective. In addition, our postexamination methods can remedy the situation without sufficient technical replicate data due to limited sample amount for verifying the reliability of identified variant peptides.

4.2. Flexibility of the Proposed Methodology. However, performing the 11 postexaminations is a very stringent evaluation of the variant events at the three bottom-up levels that can facilitate detecting the most reliable variant peptide identification. We consider that the PSM-level examination is the fundamental examination. Because the examinations at the PSM level greatly affect the ultimate outcomes of verification, researchers can select performing specific examinations or adopting a "voting" strategy to determine reliable variant PSMs as passing at least k (say, $k = 2$ or 3) out of five examinations depending on their stringency requirements. Among the five PSM-level examinations, we suggest that examination by open modification search (OMS) is essential. Variant spectra passing the PSM-level examination(s) will then proceed for examinations at variant peptide and event levels. Researchers can also select specific examinations to check. For peptide-level examinations, we consider that isobaric substitution check is a must because identified variant peptides that can be obtained by isobaric substitutions of wild-type peptides are unreliable, and this examination can be easily done. Event-level examinations are quite easy to perform and can detect unreliable variant events.

5. CONCLUSIONS

In this paper, we present a framework for the postexamination of variant peptide identification results to verify the reliability of identified variant events. The framework consists of 11 examinations at the PSM, peptide, and variant event levels based on MS proteomic knowledge. Each examination is performed on identified variant PSMs and their variant peptides, not on the whole MS data set, and can be done using public software tools or in-house programs. As a proof of concept and showing feasibility, we demonstrate the 11 examinations on the identified variant peptides by sequence database searching of a public MS data set of the HEK293 cell line. Although identified variant peptides pass an FDR of 1%, the results of 11 examinations reveal that the essential FDR criterion requirement is not sufficient to validate identified variant peptides. These rigorous examinations can serve to reveal low-confidence variant events from the shotgun proteomics experiment. Moreover, in this framework, researchers can replace some of the proposed examinations with other examinations or add different examinations. We suggest that postexaminations of identified variant events are essential and can be considered as additional guidelines to evaluate the reliability of variant events in proteogenomics studies.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c00466.

Numbers of spectra and peptides in the four spectral groups and four peptide FASTA sets, respectively (Table S1); numbers of identified variant PSMs, variant peptides, and variant events obtained from the SSe-cFDR, CMSe-cFDR, SSe-cFDR, and CMSe-cFDR (Table S2); detailed information of 302 variant PSMs obtained from adopting gFDR estimation, i.e., SSe-gFDR and CMSe-gFDR (Table S3); detailed information of 308 variant PSMs obtained from adopting cFDR estimation, i.e., SSe-cFDR and CMSe-cFDR (Table S4); a total of 320 variant PSMs obtained from four strategies used in this study (Table S5); examination of 320 spectra by open modification search (OMS) using three OMS tools (Table S6); examination of 320 spectra by explosive search against the huge SuperPep_V database (Table S7); examination of 320 spectra by combined open modification and explosive search (Table S8); examination of 320 spectra by de novo peptide sequencing tools (Table S9); examination of 320 spectra by similarity of variant spectra and simulated spectra generated by two intensity prediction tools (Table S10); four variant peptides determined as ambiguous results by isobaric substitution and semitryptic check (Table S11); classification of 98 variant events by the numbers of associated peptides and PSMs (Table S12); examination of 320 spectra by checking consecutive fragment ion peaks (Table S13); four peptides without containing the variant sites determined as nonvariant peptide by checking variant peptide sequence location (Table S14); examination of 98 variant events by checking the existence of wild-type counterparts in the PeptideAtlas database (Table S15); examination of 98 variant events by checking the identification of parental proteins (Table S16); and applying all 11-aspect examinations to the 320 variant PSMs (Table S17) (XLSX)

Workflow for variant peptide identification (Figure S1); bottom-up trilevel framework to further verify the reliability of variant events, consisting of 11-aspect examinations (Figure S2); comparison of the tryptic peptide distribution between the SuperPep_V and RefP_V databases (Figure S3); the workflow for the examination of similarity between variant spectra and predicted MS/MS spectra (Figure S4); comparison of variant peptide identification results that pass an FDR of 1% between global FDR (gFDR) and class-specific FDR (cFDR) (Figure S5); distribution of hit rank of 89, 95, and 97 spectra that identified corresponding variant peptide sequences by de novo peptide sequencing tools, PepNovo +, pNovo+, and PEAKS, respectively (Figure S6); determining the threshold of similarity score between experimental variant spectra and simulated spectra that were generated from MS^2PIP and MS^2PBPI for confident results (Figure S7); and hierarchical visualization of variant event classification based on variant peptide and spectral counting (Figure S8) (PDF)

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

AUTHOR INFORMATION

Corresponding Author

Ting-Yi Sung − *Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan;* orcid.org/0000-0002-6028-0409; Email: tsung@iis.sinica.edu.tw

Author

Wai-Kok Choong − *Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan;* orcid.org/0000-0001-6883-6865

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c00466

Funding

This work was supported, in part, by the Next-generation Pathway of Taiwan Cancer Precision Medicine Program (Grant No. AS-KPQ-107-TCPMP) in Academia Sinica and by the Ministry of Science and Technology, Taiwan (Grant No. MOST110-2221-E-001-023).

Notes

The authors declare no competing financial interest.

ABBREVIATIONS

SAVs:single amino acid variants; SNVs:single nucleotide variants; MS:mass spectrometry; CMSe:combined multiple search engines; gFDR:global FDR; cFDR:class-specific FDR; SSe:single search engine; PSMs:peptide−spectrum matches

REFERENCES

(1) Alfalah, M.; Keiser, M.; Leeb, T.; Zimmer, K. P.; Naim, H. Y. Compound heterozygous mutations affect protein folding and function in patients with congenital sucrase-isomaltase deficiency. *Gastroenterology* **2009**, *136*, 883−892.

(2) Dogan, S.; Shen, R.; Ang, D. C.; Johnson, M. L.; D'Angelo, S. P.; Paik, P. K.; Brzostowski, E. B.; Riely, G. J.; Kris, M. G.; Zakowski, M. F.; Ladanyi, M. Molecular epidemiology of EGFR and KRAS mutations in 3,026 lung adenocarcinomas: higher susceptibility of women to smoking-related KRAS-mutant cancers. *Clin. Cancer Res.* **2012**, *18*, 6169−6177.

(3) Gu, X.; Xing, L.; Shi, G.; Liu, Z.; Wang, X.; Qu, Z.; Wu, X.; Dong, Z.; Gao, X.; Liu, G.; Yang, L.; Xu, Y. The circadian mutation PER2(S662G) is linked to cell cycle progression and tumorigenesis. *Cell Death Differ.* **2012**, *19*, 397−405.

(4) Jones, R.; Ruas, M.; Gregory, F.; Moulin, S.; Delia, D.; Manoukian, S.; Rowe, J.; Brookes, S.; Peters, G. A CDKN2A mutation in familial melanoma that abrogates binding of p16INK4a to CDK4 but not CDK6. *Cancer Res.* **2007**, *67*, 9134−9141.

(5) Prior, I. A.; Lewis, P. D.; Mattos, C. A comprehensive survey of Ras mutations in cancer. *Cancer Res.* **2012**, *72*, 2457−2467.

(6) Ott, P. A.; Hu, Z.; Keskin, D. B.; Shukla, S. A.; Sun, J.; Bozym, D. J.; Zhang, W.; Luoma, A.; Giobbie-Hurder, A.; Peter, L.; Chen, C.; Olive, O.; Carter, T. A.; Li, S.; Lieb, D. J.; Eisenhaure, T.; Gjini, E.; Stevens, J.; Lane, W. J.; Javeri, I.; Nellaiappan, K.; Salazar, A. M.; Daley, H.; Seaman, M.; Buchbinder, E. I.; Yoon, C. H.; Harden, M.; Lennon, N.; Gabriel, S.; Rodig, S. J.; Barouch, D. H.; Aster, J. C.; Getz, G.; Wucherpfennig, K.; Neuberg, D.; Ritz, J.; Lander, E. S.; Fritsch, E. F.; Hacohen, N.; Wu, C. J. An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **2017**, *547*, 217−221.

(7) Schumacher, T. N.; Scheper, W.; Kvistborg, P. Cancer Neoantigens. *Annu. Rev. Immunol.* **2019**, *37*, 173−200.

(8) Dou, Y.; Kawaler, E. A.; Cui Zhou, D.; Gritsenko, M. A.; Huang, C.; Blumenberg, L.; Karpova, A.; Petyuk, V. A.; Savage, S. R.; Satpathy, S.; Liu, W.; Wu, Y.; Tsai, C. F.; Wen, B.; Li, Z.; Cao, S.; Moon, J.; Shi, Z.; Cornwell, M.; Wyczalkowski, M. A.; Chu, R. K.; Vasaikar, S.; Zhou, H.; Gao, Q.; Moore, R. J.; Li, K.; Sethuraman, S.; Monroe, M. E.; Zhao, R.; Heiman, D.; Krug, K.; Clauser, K.; Kothadia, R.; Maruvka, Y.; Pico, A. R.; Oliphant, A. E.; Hoskins, E. L.; Pugh, S. L.; Beecroft, S. J. I.; Adams, D. W.; Jarman, J. C.; Kong, A.; Chang, H. Y.; Reva, B.; Liao, Y.; Rykunov, D.; Colaprico, A.; Chen, X. S.; Czekanski, A.; Jedryka, M.; Matkowski, R.; Wiznerowicz, M.; Hiltke, T.; Boja, E.; Kinsinger, C. R.; Mesri, M.; Robles, A. I.; Rodriguez, H.; Mutch, D.; Fuh, K.; Ellis, M. J.; DeLair, D.; Thiagarajan, M.; Mani, D. R.; Getz, G.; Noble, M.; Nesvizhskii, A. I.; Wang, P.; Anderson, M. L.; Levine, D. A.; Smith, R. D.; Payne, S. H.; Ruggles, K. V.; Rodland, K. D.; Ding, L.; Zhang, B.; Liu, T.; Fenyo, D.; Clinical Proteomic Tumor Analysis Consortium. Proteogenomic characterization of endometrial carcinoma. *Cell* **2020**, *180*, 729−748.e26.

(9) Gillette, M. A.; Satpathy, S.; Cao, S.; Dhanasekaran, S. M.; Vasaikar, S. V.; Krug, K.; Petralia, F.; Li, Y.; Liang, W. W.; Reva, B.; Krek, A.; Ji, J.; Song, X.; Liu, W.; Hong, R.; Yao, L.; Blumenberg, L.; Savage, S. R.; Wendl, M. C.; Wen, B.; Li, K.; Tang, L. C.; MacMullan, M. A.; Avanessian, S. C.; Kane, M. H.; Newton, C. J.; Cornwell, M.; Kothadia, R. B.; Ma, W.; Yoo, S.; Mannan, R.; Vats, P.; Kumar-Sinha, C.; Kawaler, E. A.; Omelchenko, T.; Colaprico, A.; Geffen, Y.; Maruvka, Y. E.; da Veiga Leprevost, F.; Wiznerowicz, M.; Gumus, Z. H.; Veluswamy, R. R.; Hostetter, G.; Heiman, D. I.; Wyczalkowski, M. A.; Hiltke, T.; Mesri, M.; Kinsinger, C. R.; Boja, E. S.; Omenn, G. S.; Chinnaiyan, A. M.; Rodriguez, H.; Li, Q. K.; Jewell, S. D.; Thiagarajan, M.; Getz, G.; Zhang, B.; Fenyo, D.; Ruggles, K. V.; Cieslik, M. P.; Robles, A. I.; Clauser, K. R.; Govindan, R.; Wang, P.; Nesvizhskii, A. I.; Ding, L.; Mani, D. R.; Carr, S. A.; Clinical Proteomic Tumor Analysis Consortium. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell* **2020**, *182*, 200−225.e35.

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

(11) Wen, B.; Li, K.; Zhang, Y.; Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat. Commun.* **2020**, *11*, No. 1759.

(12) Alfaro, J. A.; Ignatchenko, A.; Ignatchenko, V.; Sinha, A.; Boutros, P. C.; Kislinger, T. Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines. *Genome Med.* **2017**, *9*, No. 62.

(13) Choong, W. K.; Wang, J. H.; Sung, T. Y. MinProtMaxVP: Generating a minimized number of protein variant sequences containing all possible variant peptides for proteogenomic analysis. *J. Proteomics* **2020**, *223*, No. 103819.

(14) Bogdanow, B.; Zauber, H.; Selbach, M. Systematic errors in peptide and protein identification and quantification by modified peptides. *Mol. Cell. Proteomics* **2016**, *15*, 2791−2801.

(15) Chen, Y.; Zhang, J.; Xing, G.; Zhao, Y. Mascot-derived false positive peptide identifications revealed by manual analysis of tandem mass spectra. *J. Proteome Res.* **2009**, *8*, 3141−3147.

(16) Duncan, M. W.; Aebersold, R.; Caprioli, R. M. The pros and cons of peptide-centric proteomics. *Nat. Biotechnol.* **2010**, *28*, 659−664.

(17) Choong, W. K.; Lih, T. M.; Chen, Y. J.; Sung, T. Y. Decoding the effect of isobaric substitutions on identifying missing proteins and variant peptides in human proteome. *J. Proteome Res.* **2017**, *16*, 4415−4424.

(18) Cunningham, F.; Achuthan, P.; Akanni, W.; Allen, J.; Amode, M. R.; Armean, I. M.; Bennett, R.; Bhai, J.; Billis, K.; Boddu, S.; Cummins, C.; Davidson, C.; Dodiya, K. J.; Gall, A.; Giron, C. G.; Gil, L.; Grego, T.; Haggerty, L.; Haskell, E.; Hourlier, T.; Izuogu, O. G.; Janacek, S. H.; Juettemann, T.; Kay, M.; Laird, M. R.; Lavidas, I.; Liu, Z.; Loveland, J. E.; Marugan, J. C.; Maurel, T.; McMahon, A. C.; Moore, B.; Morales, J.; Mudge, J. M.; Nuhn, M.; Ogeh, D.; Parker, A.; Parton, A.; Patricio, M.; Abdul Salam, A. I.; Schmitt, B. M.; Schuilenburg, H.; Sheppard, D.; Sparrow, H.; Stapleton, E.; Szuba, M.; Taylor, K.; Threadgold, G.; Thormann, A.; Vullo, A.; Walts, B.; Winterbottom, A.; Zadissa, A.; Chakiachvili, M.; Frankish, A.; Hunt, S. E.; Kostadima, M.; Langridge, N.; Martin, F. J.; Muffato, M.; Perry, E.; Ruffier, M.; Staines, D. M.; Trevanion, S. J.; Aken, B. L.; Yates, A. D.; Zerbino, D. R.; Flicek, P. Ensembl 2019. *Nucleic Acids Res.* **2019**, *47*, D745−D751.

(19) 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.; DiCuccio, M.; Kitts, P.; Murphy, T. D.; Pruitt, K. D. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733−D745.

(20) Zhu, Y.; Orre, L. M.; Johansson, H. J.; Huss, M.; Boekel, J.; Vesterlund, M.; Fernandez-Woodbridge, A.; Branca, R. M. M.; Lehtio, J. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat. Commun.* **2018**, *9*, No. 903.

(21) Yi, X.; Wang, B.; An, Z.; Gong, F.; Li, J.; Fu, Y. Quality control of single amino acid variations detected by tandem mass spectrometry. *J. Proteomics* **2018**, *187*, 144−151.

(22) Wen, B.; Wang, X.; Zhang, B. PepQuery enables fast, accurate, and convenient proteomic validation of novel genomic alterations. *Genome Res.* **2019**, *29*, 485−493.

(23) Deutsch, E. W.; Lane, L.; Overall, C. M.; Bandeira, N.; Baker, M. S.; Pineau, C.; Moritz, R. L.; Corrales, F.; Orchard, S.; Van Eyk, J. E.; Paik, Y. K.; Weintraub, S. T.; Vandenbrouck, Y.; Omenn, G. S. Human proteome project mass spectrometry data interpretation guidelines 3.0. *J. Proteome Res.* **2019**, *18*, 4108−4116.

(24) Paik, Y. K.; Jeong, S. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Cho, S. Y.; Lee, H. J.; Na, K.; Choi, E. Y.; Yan, F.; Zhang, F.; Zhang, Y.; Snyder, M.; Cheng, Y.; Chen, R.; Marko-Varga, G.; Deutsch, E. W.; Kim, H.; Kwon, J. Y.; Aebersold, R.; Bairoch, A.; Taylor, A. D.; Kim, K. Y.; Lee, E. Y.; Hochstrasser, D.; Legrain, P.; Hancock, W. S. The chromosome-centric human proteome project for cataloging proteins encoded in the genome. *Nat. Biotechnol.* **2012**, *30*, 221−223.

(25) Perez-Riverol, Y.; Csordas, A.; Bai, J.; Bernal-Llinares, M.; Hewapathirana, S.; Kundu, D. J.; Inuganti, A.; Griss, J.; Mayer, G.; Eisenacher, M.; Perez, E.; Uszkoreit, J.; Pfeuffer, J.; Sachsenberg, T.; Yilmaz, S.; Tiwary, S.; Cox, J.; Audain, E.; Walzer, M.; Jarnuczak, A. F.; Ternent, T.; Brazma, A.; Vizcaino, J. A. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **2019**, *47*, D442−D450.

(26) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **2015**, *33*, 743−749.

(27) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M. Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30*, 918−920.

(28) Lin, Y. C.; Boone, M.; Meuris, L.; Lemmens, I.; Van Roy, N.; Soete, A.; Reumers, J.; Moisse, M.; Plaisance, S.; Drmanac, R.; Chen, J.; Speleman, F.; Lambrechts, D.; Van de Peer, Y.; Tavernier, J.; Callewaert, N. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. *Nat. Commun.* **2014**, *5*, No. 4767.

(29) UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480−D489.

(30) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **2004**, *3*, 1234−1242.

(31) Choong, W. K.; Sung, T. Y. Comparison of different variant sequence types coupled with decoy generation methods used in concatenated target-decoy database searches for proteogenomic research. *J. Proteomics* **2021**, *231*, No. 104021.

(32) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8*, 2405−2417.

(33) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13*, 22−24.

(34) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, No. 5277.

(35) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466−1467.

(36) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74*, 5383−5392.

(37) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10*, No. M111.007690.

(38) Creasy, D. M.; Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **2004**, *4*, 1534−1536.

(39) Yu, F.; Li, N.; Yu, W. PIPI: PTM-Invariant Peptide Identification Using Coding Method. *J. Proteome Res.* **2016**, *15*, 4423−4435.

(40) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **2017**, *14*, 513−520.

(41) David, M.; Fertin, G.; Rogniaux, H.; Tessier, D. SpecOMS: A full open modification search method performing all-to-all spectra comparisons within minutes. *J. Proteome Res.* **2017**, *16*, 3030−3038.

(42) Deutsch, E. W.; Sun, Z.; Campbell, D. S.; Binz, P. A.; Farrah, T.; Shteynberg, D.; Mendoza, L.; Omenn, G. S.; Moritz, R. L. Tiered human integrated sequence search databases for shotgun proteomics. *J. Proteome Res.* **2016**, *15*, 4091−4100.

(43) Cao, R.; Shi, Y.; Chen, S.; Ma, Y.; Chen, J.; Yang, J.; Chen, G.; Shi, T. dbSAP: single amino-acid polymorphism database for protein variation detection. *Nucleic Acids Res.* **2017**, *45*, D827−D832.

(44) Zahn-Zabal, M.; Michel, P. A.; Gateau, A.; Nikitin, F.; Schaeffer, M.; Audot, E.; Gaudet, P.; Duek, P. D.; Teixeira, D.; Rech de Laval, V.; Samarasinghe, K.; Bairoch, A.; Lane, L. The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Res.* **2020**, *48*, D328−D334.

(45) Tate, J. G.; Bamford, S.; Jubb, H. C.; Sondka, Z.; Beare, D. M.; Bindal, N.; Boutselakis, H.; Cole, C. G.; Creatore, C.; Dawson, E.; Fish, P.; Harsha, B.; Hathaway, C.; Jupe, S. C.; Kok, C. Y.; Noble, K.; Ponting, L.; Ramshaw, C. C.; Rye, C. E.; Speedy, H. E.; Stefancsik, R.; Thompson, S. L.; Wang, S.; Ward, S.; Campbell, P. J.; Forbes, S. A. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **2019**, *47*, D941−D947.

(46) Sherry, S. T.; Ward, M. H.; Kholodov, M.; Baker, J.; Phan, L.; Smigielski, E. M.; Sirotkin, K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **2001**, *29*, 308−311.

(47) Mathivanan, S.; Ji, H.; Tauro, B. J.; Chen, Y. S.; Simpson, R. J. Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. *J. Proteomics* **2012**, *76*, 141−149.

(48) Schandorff, S.; Olsen, J. V.; Bunkenborg, J.; Blagoev, B.; Zhang, Y.; Andersen, J. S.; Mann, M. A mass spectrometry-friendly database for cSNP identification. *Nat. Methods* **2007**, *4*, 465−466.

(49) Keil, B. *Specificity of Proteolysis*; Springer-Verlag: Berlin, New York, 1992; p ix, 336 p.

(50) Frank, A.; Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77*, 964−973.

(51) Chi, H.; Chen, H.; He, K.; Wu, L.; Yang, B.; Sun, R. X.; Liu, J.; Zeng, W. F.; Song, C. Q.; He, S. M.; Dong, M. Q. pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. *J. Proteome Res.* **2013**, *12*, 615−625.

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467

(52) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337−2342.
(53) Muth, T.; Weilnbock, L.; Rapp, E.; Huber, C. G.; Martens, L.; Vaudel, M.; Barsnes, H. DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. *J. Proteome Res.* **2014**, *13*, 1143−1146.
(54) Degroeve, S.; Martens, L. MS2PIP: a tool for MS/MS peak intensity prediction. *Bioinformatics* **2013**, *29*, 3199−3203.
(55) Dong, N. P.; Liang, Y. Z.; Xu, Q. S.; Mok, D. K.; Yi, L. Z.; Lu, H. M.; He, M.; Fan, W. Prediction of peptide fragment ion mass spectra by data mining techniques. *Anal. Chem.* **2014**, *86*, 7446−7454.
(56) Searle, B. C. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* **2010**, *10*, 1265−1269.
(57) McIlwain, S.; Tamura, K.; Kertesz-Farkas, A.; Grant, C. E.; Diament, B.; Frewen, B.; Howbert, J. J.; Hoopmann, M. R.; Kall, L.; Eng, J. K.; MacCoss, M. J.; Noble, W. S. Crux: rapid open source protein tandem mass spectrometry analysis. *J. Proteome Res.* **2014**, *13*, 4488−4491.
(58) Lynn, K. S.; Chen, C. C.; Lih, T. M.; Cheng, C. W.; Su, W. C.; Chang, C. H.; Cheng, C. Y.; Hsu, W. L.; Chen, Y. J.; Sung, T. Y. MAGIC: an automated N-linked glycoprotein identification tool using a Y1-ion pattern matching algorithm and in silico MS(2) approach. *Anal. Chem.* **2015**, *87*, 2466−2473.
(59) Taus, T.; Kocher, T.; Pichler, P.; Paschke, C.; Schmidt, A.; Henrich, C.; Mechtler, K. Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **2011**, *10*, 5354−5362.
(60) Ivanov, M. V.; Lobas, A. A.; Levitsky, L. I.; Moshkovskii, S. A.; Gorshkov, M. V. Brute-force approach for mass spectrometry-based variant peptide identification in proteogenomics without personalized genomic data. *J. Am. Soc. Mass Spectrom.* **2018**, *29*, 435−438.
(61) Mamie Lih, T. S.; Choong, W. K.; Chen, Y. J.; Sung, T. Y. Evaluating the possibility of detecting variants in shotgun proteomics via LeTE-Fusion analysis pipeline. *J. Proteome Res.* **2018**, *17*, 2937−2952.
(62) Deutsch, E. W. The PeptideAtlas Project. In *Methods in Molecular Biology*; Humana Press, 2010; Vol. *604*, pp 285−296.
(63) Rodriguez, J.; Gupta, N.; Smith, R. D.; Pevzner, P. A. Does trypsin cut before proline? *J. Proteome Res.* **2008**, *7*, 300−305.
(64) Lobas, A. A.; Karpov, D. S.; Kopylov, A. T.; Solovyeva, E. M.; Ivanov, M. V.; Ilina, I. Y.; Lazarev, V. N.; Kuznetsova, K. G.; Ilgisonis, E. V.; Zgoda, V. G.; Gorshkov, M. V.; Moshkovskii, S. A. Exome-based proteogenomics of HEK-293 human cell line: Coding genomic variants identified at the level of shotgun proteome. *Proteomics* **2016**, *16*, 1980−1991.

https://doi.org/10.1021/acsomega.2c00466
ACS Omega 2022, 7, 16454−16467