Computational and Structural Biotechnology Journal 23 (2024) 711–722

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/web server article

PgxSAVy: A tool for comprehensive evaluation of variant peptide quality in proteogenomics – catching the (un)usual suspects

Anurag Raj [a,b,1], Suruchi Aggarwal [c,d,e,2,3], Prateek Singh [a,b,4], Amit Kumar Yadav [c,d,e,*,5], Debasis Dash [a,b,**,6,7]

[a] G. N. Ramachandran Knowledge Centre for Genomics Informatics, CSIR – Institute of Genomics and Integrative Biology, New Delhi, India
[b] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India
[c] Computational and Mathematical Biology Centre (CMBC), 3rd Milestone, Faridabad-Gurgaon Expressway, Faridabad, Haryana 121001, India
[d] Centre for Drug Discovery (CDD), 3rd Milestone, Faridabad-Gurgaon Expressway, Faridabad, Haryana 121001, India
[e] Centre for Microbial Research (CMR), Translational Health Science and Technology Institute, NCR Biotech Science Cluster, 3rd Milestone, Faridabad-Gurgaon Expressway, Faridabad, Haryana 121001, India

ARTICLE INFO

Keywords:
Proteogenomics
PgxSAVy
Variant peptides
Proteoforms
SAVs
SAPs
Mass spectrometry
Proteomics
Mutations
SNPs
False discovery rate
Quality assessment

ABSTRACT

Variant peptides resulting from single nucleotide polymorphisms (SNPs) can lead to aberrant protein functions and have translational potential for disease diagnosis and personalized therapy. Variant peptides detected by proteogenomics are fraught with high number of false positives, but there is no uniform and comprehensive approach to assess variant quality across analysis pipelines. Despite class-specific FDR along with ad-hoc filters, the problem is far from solved. These protocols are typically manual and tedious, and thus not uniform across labs. We demonstrate that variant peptide rescoring, integrated with intensity, variant event information and search result features, allows better discrimination of correct variant peptides. Implemented into PgxSAVy – a tool for quality control of variant peptides, this method can tackle the high rate of false positives. PgxSAVy provides a rigorous framework for quality control and annotations of variant peptides on the basis of (i) variant quality, (ii) isobaric masses, and (iii) disease annotation. PgxSAVy demonstrated high accuracy by identifying true variants with 98.43% accuracy on simulated data. Large-scale proteogenomic reanalysis of ~2.8 million spectra (PXD004010 and PXD001468) resulted in 12,705 variant peptide spectrum matches (PSMs), of which PgxSAVy evaluated 3028 (23.8%), 1409 (11.1%) and 8268 (65.1%) as confident, semi-confident and doubtful respectively. PgxSAVy also annotates the variants based on their pathogenicity and provides support for assisted manual validation. The analysis of proteins carrying variants can provide fine granularity in discovering important pathways. PgxSAVy will advance personalized medicine by providing a comprehensive framework for quality control and prioritization of proteogenomics variants. PgxSAVy is freely available at https://pgxsavy.igib.res.in/ as a webserver and https://github.com/anuragraj/PgxSAVy as a stand-alone tool.

* Corresponding author at: Computational and Mathematical Biology Centre (CMBC), Centre for Drug Discovery (CDD), Centre for Microbial Research (CMR), Translational Health Science and Technology Institute, NCR Biotech Science Cluster, 3rd Milestone, Faridabad-Gurgaon Expressway, Faridabad, Haryana 121001, India.
** Corresponding author at: G. N. Ramachandran Knowledge Centre for Genomics Informatics, CSIR – Institute of Genomics and Integrative Biology, New Delhi, India.
E-mail addresses: amit.yadav@thsti.res.in (A.K. Yadav), ddash@igib.res.in (D. Dash).
[1] http://orcid.org/0000–0001–9656–8167
[2] Present Address: Division of Biomedical Informatics, Indian Council of Medical Research, Ansari Nagar, New Delhi–110029, India
[3] http://orcid.org/0000–0002–3921–321X
[4] http://orcid.org/0000–0001–9053–4249
[5] http://orcid.org/0000–0002–9445–8156
[6] http://orcid.org/0000–0002–5647–3785
[7] Present Address: Institute of Life Sciences, Bhubaneswar, Odisha, 751023, India

https://doi.org/10.1016/j.csbj.2023.12.033
Received 4 October 2023; Received in revised form 19 December 2023; Accepted 23 December 2023
Available online 26 December 2023

1. Introduction

Single nucleotide polymorphisms (SNPs) play an important role in defining the health and disease status of an organism. Single amino acid variations (SAVs) caused due to SNPs can alter the structure, interactions or activity of the corresponding proteins [1,2]. Proteogenomics is a powerful approach to mine mass spectrometry data for identification and characterization of variants. Many variants of biological importance have been discovered in diseases like cancers, Alzheimer's, Parkinson's, diabetes and cardiovascular diseases [3–8]. Integrated analysis of multi-omics data can immensely benefit personalized medicine by proteogenomics analysis [9,10], which can directly connect patient-specific genomic variants and their protein level translation products. Thus, proteogenomics can provide important information for personalized medicine, by connecting disease susceptibility, diagnosis and response to drugs based on their associated variants [11,12].

In shotgun proteomics, the experimental MS/MS spectra are generated from the fragmentation of enzymatically digested peptides, using proteases such as trypsin, chymotrypsin, lys-C, etc. These experimental MS/MS spectra are matched against theoretical MS/MS generated from in silico digested peptides from a reference protein database. The database search tools such as X!Tandem [13], OMSSA [14], MSGF+ [15] etc. can identify peptides and their known modifications. The single amino acid variants (SAVs) are usually not present in the database and therefore not identified. The proteogenomics search against 3-frame translated transcriptome or 6-frame translated genome are best suited for variant peptide detection [16–20], but often contain many false positives. The statistical validation is performed by a target-decoy (TD) based false discovery rate (FDR) control, in which the data is searched simultaneously in a reference (target) database and its reversed-sequence (decoy) database. The FDR is estimated by ratio of decoy/target at any given threshold to control error rates [21–23].

Since the proteogenomics databases tend to be humongous in size, there is a higher-than-usual chances of false hits. A custom variant peptide database is created representing all combinations of known variants at every site from neXtProt database. To identify variants, a custom variant database is used which is ten times or larger in size compared to a standard proteogenomic search database created from six/three frame translated genome or transcriptome data[24]. Thus, it is also fraught with high rates of error that the target-decoy (TD) based FDR often fails to control [23,25]. This reduced sensitivity and increased rate of false positives pose major challenges in variant identification and are highlighted in recent articles [25–29]. Most proteogenomic studies calculate a global false discovery rate for all peptide identifications, including known, novel and variant peptides [30–32]. Some studies have recommended a higher threshold of FDR for novel peptides and implemented stringent filters or class-specific FDR (cFDR) [33]. Even with stricter thresholds, the problem remains recalcitrant as demonstrated in several studies [34–36]. Therefore, a comprehensive identification of the sequence variants of the proteins is still a formidable challenge [35].

The process of confidence assessment requires additional evaluation and filtering criteria after the database search and FDR estimation. Li et al. developed a modified FDR estimation workflow for evaluation of variant peptides, but also acknowledged that many false variants still remained in the data [37]. It is also required that the variant peptides should score higher than their wild type sequences for a given MS/MS spectrum. The variant peptide should be able to score better than the wild type to provide higher confidence, otherwise we can safely assume that wild type is the likely peptide as there is no extra information that can point towards the presence of the variant peptide. Another study proposed SpectrumAI tool to verify SAV peptides by requiring variant-event ions to directly support the residue substitution in MS/MS spectra [38]. However, it only checks variant amino acid peak but does not use intensity information, neither does it compare against the wildtype peptides. A comprehensive set of recommendations was proposed for variant peptide search and evaluation, in which the authors employed multi-algorithm searches and a split target-decoy approach (same as class-specific FDR) coupled with various post-search filters to remove false identifications [36]. This included isobaric checks to test if variant can be explained by other isobaric amino acids, post-translation modifications (PTM) mass shift, reference proteome mapping to detect indistinguishable peptides and protein abundance. Similarly, the post-examination of variant peptides by applying filters and checking criteria like isobaric substitutions has been proposed recently [35] with the demonstration that standard FDR estimation was not sufficient in controlling variant false positives.

However, these assessment criteria and recommendations, although useful, are not amenable for automated verification of variant peptides. Also, since these recommendations are not adopted in a uniform manner across different labs, the implementations can differ widely. Due to lack of a standard tool, it either requires advanced bioinformatics skills or manual verification, neither of which is complete or comprehensive. Variant quality evaluation, thus requires robust and objective scoring system implemented as an automated pipeline. SAVControl [34] is a promising tool that filters variant peptides based on subgroup FDR. It additionally checks for variant amino acid position in the peptide by mapping given mass shift and residue against UniMod counterparts. It subsequently provides a classification based on mapping of site determined by PTMiner [39] with UniMod but does not perform in-depth check of variant quality.

Using stage-specific searches for the wild type peptides followed by variant peptides for unassigned spectra, helps in reducing false variants [40]. Some studies have also used multiple search engines to improve PSM identification rate by harnessing complementarity between search engines [36]. Measures like spectral match quality, intensity coverage and b/y ion counts are often employed to weed out false variants in manual verification. An easy to calculate variant quality measure that is implemented in a search-engine-agnostic pipelines is an important requirement or design goal to facilitate quality control in variant proteogenomics.

With these design goals and challenges in mind, we developed PgxSAVy tool that first rescores the variant peptides using a modified version of MassWiz score [41], and calculates *variant ambiguity score* (VAS) as a universally applicable variant scoring method. Here, we posited that variant peptide rescoring and integration of various features for PSM quality, variant event and global search results (i.e., PSMs count per peptides and/or search engine count) together into VAS can reduce false variants. These features add value in assessing the quality of variant peptide above and beyond TD based FDR, multiple search engines and cFDR. We have integrated the tedious step of manual validation, isobaric checks and available pathogenic/disease related information for human variant peptides into a seamless, standardized and easy-to-use tool. We have demonstrated the accuracy and utility using simulated spectra as well as large-scale data from Alzheimer's disease and HEK293 cell lines. This will advance variant proteogenomics from method development to biological applications.

2. Methods

Fig. 1 shows the overall workflow used in this study and the PgxSAVy method.

2.1. Development of PgxSAVy

We developed PgxSAVy incorporating match quality, variant events and result features. We utilized various PSM and variant event features for segregating true and false variant peptides for the development of a variant ambiguity score (VAS). Features like spectrum intensity coverage, b and y ion peak matches and their continuity score are well-known for evaluation of PSM quality and part of MassWiz score (MW). In

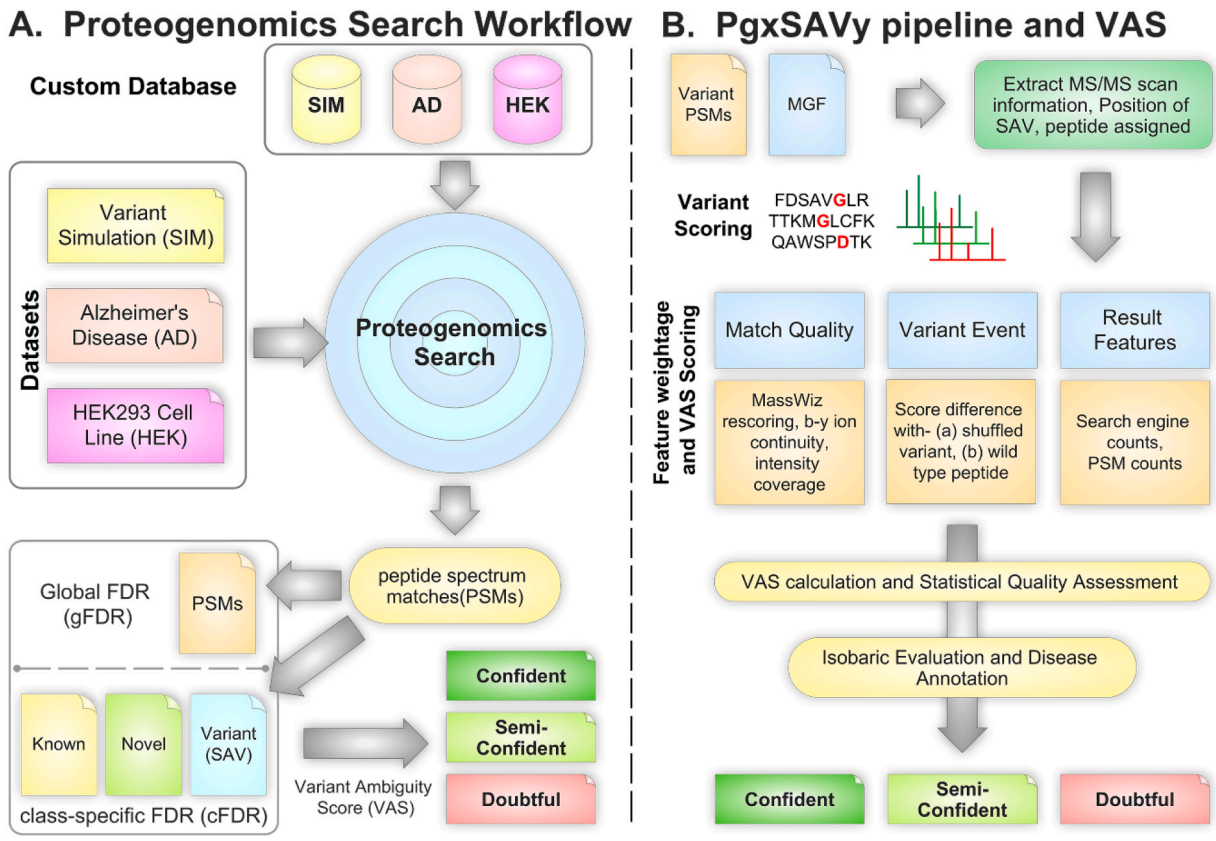

Fig. 1. Workflow for identification and post-search validation of variant peptides from proteogenomics searches using PgxSAVy. (A) Simulation (SIM), Alzheimer's disease (AD) and HEK293 Cell line (HEK) data were searched against respective databases- simDB, AD and HEK search database using EuGenoSuite, followed by global and class-specific (gFDR and cFDR) FDR estimation. Variant Ambiguity Score (VAS) is calculated for the variant PSMs that pass 1%FDR threshold through PgxSAVy. (B) In PgxSAVy tool, VAS is calculated using variant search results and MS/MS spectra, utilizing match quality, variant event and result features. This is followed by statistical quality assessment (z-score and p-value calculation) for quality evaluation and classification. Isobaric assessment and annotation is also performed within PgxSAVy framework for the variant PSMs.

addition, the continuity of a series (b/y) significantly enhances confidence in matched ions even when fragmentation is incomplete due to partially mobile or non-mobile proton-containing peptides. Since these features are generally not used together and also not part of search output, it makes the post-search evaluation on these features impossible. The variant evaluation features or descriptors can be grouped in three main categories: match quality, variant events and result features. In PSM matching, we measure the goodness-of-fit of experimental and theoretical spectra and therefore, it is easier to measure the match as being "good" or "bad" than being correct or incorrect. These features have the ability to discriminate between true (good) and false (bad) variant hits. The PgxSAVy tool evaluates variant PSMs individually and does not conflate them into variant peptides before evaluation, as different scans matching to the same variant sequence may have a variable degree of quality.

2.2. Match quality (Variant Rescoring)

2.2.1. Fractional intensity coverage (FIC)

Fractional intensity coverage refers to the ratio of the summed intensity for the matched peaks for the peptide sequence, against total summed intensity of the respective spectrum.

$$FIC = \frac{\sum_{i=1}^{k} I_i}{\sum_{i=1}^{n} I_i}$$

where,

n = total experimental peaks,
k = matched peaks.
$I_i = i^{th}$ matched peak intensity.

A peptide that covers most of the high-quality MS/MS peaks for a spectrum gets a good fractional intensity coverage (at least half to two-third of the spectrum). Thus, this feature can distinguish between good and bad PSM quite well and is part of the variant scoring.

2.2.2. b and y ions continuity count (byCC)

Occurrence of continuous b and y ions in a spectrum provides information about the quality of spectrum match and can distinguish a true variant from false one. More the number of continuous b and y ion ladders, more the indication of good fragmentation and quality variant match. The byCC is a sum total of the b-ion continuity score and y-ion continuity score. The byCC provides higher weightage to variant matches that have better coverages of continuous b and y ions.

$$bCC = \sum_{i=1}^{k} P_{b_i} \quad \begin{cases} P = 1, \& P_i \quad and \quad P_{i+1} \quad are \quad both \quad matched \\ P = 0, \& otherwise \end{cases}$$

$$yCC = \sum_{i=1}^{k} P_{y_i} \quad \begin{cases} P = 1, \& P_i \quad and \quad P_{i+1} \quad are \quad both \quad matched \\ P = 0, \& otherwise \end{cases}$$

$$byCC = bCC + yCC$$

where,

k = matched peaks,
P_{bi} = b-ion flag (0 if absent, 1 if present),

P_{yi} = y-ion flag (0 if absent, 1 if present).

2.2.3. Complementary ions

Complementary ions such as- (i) immonium ions, (ii) neutral losses of water, and (iii) neutral losses of ammonia, can also aid in distinguishing between borderline true and false hits when b and y ions counts are closely similar, which can immensely benefit the identification of variant peptides.

2.2.4. Variant peptide rescoring

All the above features and complementary neutral loss ions are used in MassWiz score (MW) [41], which we used to rescore the variant peptides for additional confidence. However, it was shown to have a slight bias for longer peptides [42], which was normalized by the theoretical b & y ion counts. The normalized MassWiz Score (nMW) thus obtained, is calculated as: -

$$nMW = \frac{MW}{(2 \times L - 3) + byCC}$$

where,

L = Peptide length (2 L-3 denotes theoretical b/y ions and assumes b1 to be absent),

byCC = b and y ions continuity count (explained before).

This normalization ensures that the nMW represents the strength of match per ion matched, and better intensity matches are still scored higher.

2.3. Variant event features

2.3.1. Score difference with shuffled variant decoy peptides (ΔSV)

The variant event needs to be correctly localized at the appropriate amino acid to distinguish from other possible variant positions in the peptide sequence. Previously, post translational modifications studies have used this strategy to build a rescoring to correctly localize the modification site [43–45]. This variant decoy approach takes in a single/double amino acid variant peptide, and then iteratively exchange/shuffle the variant amino acid at a given position with the amino acids present in that peptide at other positions as long as the exchanged amino acids are not the same. These shuffled variants effectively act as decoys and VAS is calculated for the original variant and these decoy candidates. The hypothesis is that a decoy variant peptide (created by swapping the variant amino acid with other amino acid, one at a time) in the sequence, should always score less than a correct variant peptide. This ΔSV score highlights the gap between a correct variant and a random decoy variants, as exemplified in a previous study [34].

2.3.2. Score difference with wild type peptide (ΔWT)

For the correct variant peptide match, the corresponding score for the wild type (WT) peptide should ideally be lesser for the given MS/MS spectrum. This spectrum-centric argument assumes that in a narrow accurate mass search (10 ppm usually), it is highly improbable to match different allelic peptide variants for the same MS/MS spectrum. Similarly, we also do not expect the corresponding wild type peptides to match within that tolerance of 10 ppm, for that MS/MS spectrum. If parent mass errors are ignored, the variant peptide should still be able to score better than the wild type to be able to provide higher confidence. We can safely assume that the variant peptide must have a better match than the wild type. Therefore, the score difference between these two types of peptides matches to the same spectrum can be exploited to establish if the variant is likely to be correct. The higher the score difference between variant peptide and WT peptide, the higher the chances of variant match to be correct.

2.4. Result features

2.4.1. Number of search engines identifying the variant peptide

The higher the number of search engines that identify a given MS/MS spectrum with the same variant peptide sequence, the higher the probability of the PSM to be correct. During a database search, it is plausible that different search engines report different peptide sequences for the same scan. For variants, search engines are more likely to agree for true variants, while disagreements may be due to: - (i) false hits, (ii) low information content in MS/MS, or (iii) one search engine reporting incorrect variant. Thus, using search engine count as a weightage can add additional discriminatory power.

2.4.2. PSM count per peptide

For any identified variant, higher the number of PSMs identifying the same variant sequence, higher are the chances of the variant being correct. A majority of incorrect hits are identified with a single PSM which requires greater scrutiny. The chances of random matching of variant peptides to single PSMs in proteogenomics is very high, and such "one-hit proteogenomics-wonders" occur at higher frequency than in proteomics [46]. So, PSM counts per variant peptide can prove to be a useful indicator for correct variants.

2.5. Variant ambiguity score (VAS)

The VAS calculates a score that can assess the quality of the variant peptides which are divided into three classes – confident, semi-confident and doubtful.

2.5.1. VAS calculation

For a variant peptide spectrum match i, the VAS is defined as –

$$VAS_i = (nMW_i + \Delta SV_i + \Delta WT_i)/3) * \log_{10}(P_i + 1) * \log_{SE}(SE_i + 1)$$

where,

nMW_i = normalized MassWiz Score for i^{th} variant match,

ΔSV_i = delta shuffled variant score for i^{th} variant match,

ΔWT_i = delta wildtype score for i^{th} variant match,

P_i = PSM counts for i^{th} variant match,

SE = number of search engines used for the proteogenomics search (this remains constant for a study and is used as the logarithmic base for calculating SE_i weight),

SE_i = number of search engines that identified the i^{th} variant PSM.

Although we know that the number of search engines and PSM counts can identify good hits, these are not exclusive criteria and should not be given too much weightage as they have the capacity to skew the results. The two logarithms used in the formula are there to prevent these factors from skewing resultant VAS too much.

2.5.2. P-value calculation

We model the random matching variants to follow a normal distribution centred at VAS score of zero (Supplementary Fig. S1). Using the minVAS (minimum value of VAS) as the left tail, we assume the corresponding modulus as the right tail of this distribution, hereby referred to as the null distribution. From this, we calculate standard scores (z-scores) and corresponding p-values. The non-random matches will get a low p-value. The null hypothesis H_0, assumes that the scores following this null distribution are false. If a variant scores significantly higher, it is likely to be correct and incorrect otherwise. Using the z-score, p-value is calculated and variant PSMs are classified as (i) confident (p-value <= 0.01), (ii) semi-confident (p-value <=0.1) and (iii) doubtful (p-value >0.1).

2.5.3. Isobaric analysis on identified variant peptides through proteogenomics

Isobaric (or near isobaric) amino acids and modifications are

indistinguishable if these masses are within the instrument mass error [47]. To tag such peptides, we implemented an isobaric assessment module, that evaluates the mass difference between variant amino acids or modifications within the applied MS/MS tolerance range. This check is performed for both single and double variant types and can result in three broad classes and their respective subclasses (in parentheses)– (i) single variant (var, iso-var, mod-var), (ii) double variant (var-var), and (iii) isobaric (iso, iso-iso, iso-mod).

2.5.4. *Annotating identified variants through PgxSAVy*

For the identified variants, PgxSAVy performs disease annotation using biological information present in the UniProt database that describes their pathogenicity. These biological annotations include their clinical significance and phenotypes. If the annotation is missing in curated databases culled by UniProt, it could be a novel variant discovery (not observed before).

2.5.5. *Implementation as a stand-alone tool and webserver*

All the above steps are implemented in the tool PgxSAVy, that reads the input, rescores variants in an automated manner, calculates VAS scores and classifies the variants with isobaric checks and perform biological annotation. The tool evaluates variant PSMs individually and does not conflate them into variant peptides before evaluation, as different scans matching the same sequence may have variable quality. We implemented an isobaric assessment module, that evaluates isobaric amino acids and modifications [47], and can result in three broad classes (i) single variant, (ii) double variant, and (iii) isobaric. PgxSAVy also performs disease annotation using UniProt database that includes clinical significance and phenotype. A Perl script is also provided to create pLabel batch file for PSM annotation and visualization.

PgxSAVy reads the input, calculates VAS scores and classifies the variants with isobaric checks and perform biological annotation in automated manner. PgxSAVy is available at https://github.com/anuragraj/PgxSAVy.

To enable a more convenient access to the PgxSAVy tool, a webserver has been created at https://pgxsavy.igib.res.in. The aim is to provide a user-friendly web-based application of the tool. The webserver contains two parts: the frontend management portal and the backend scheduler. The frontend management portal is developed using the Python library Flask which creates and manages the projects submitted by the users. The backend scheduler has the ability to execute jobs in parallel for the confidence assessment and annotation of variant peptides (supplementary Fig. S13). Overall, combining the frontend and the scheduler creates a simple yet fully functional implementation of the PgxSAVy tool.

2.6. *Datasets*

A simulated data was generated using MaSS-Simulator [48] using various parameters stated in Supplementary Table S1 and S2. For one and two amino acid variants, 5000 peptides each were provided to the MS/MS spectra simulator, along with 40,000 wildtype peptides (Fig. 2). Two publicly available MS/MS datasets- one from Alzheimer's disease (AD) (PXD004010) containing 1.7 million spectra [49] and another from HEK293 cell line data (HEK) containing 1.1 million spectra

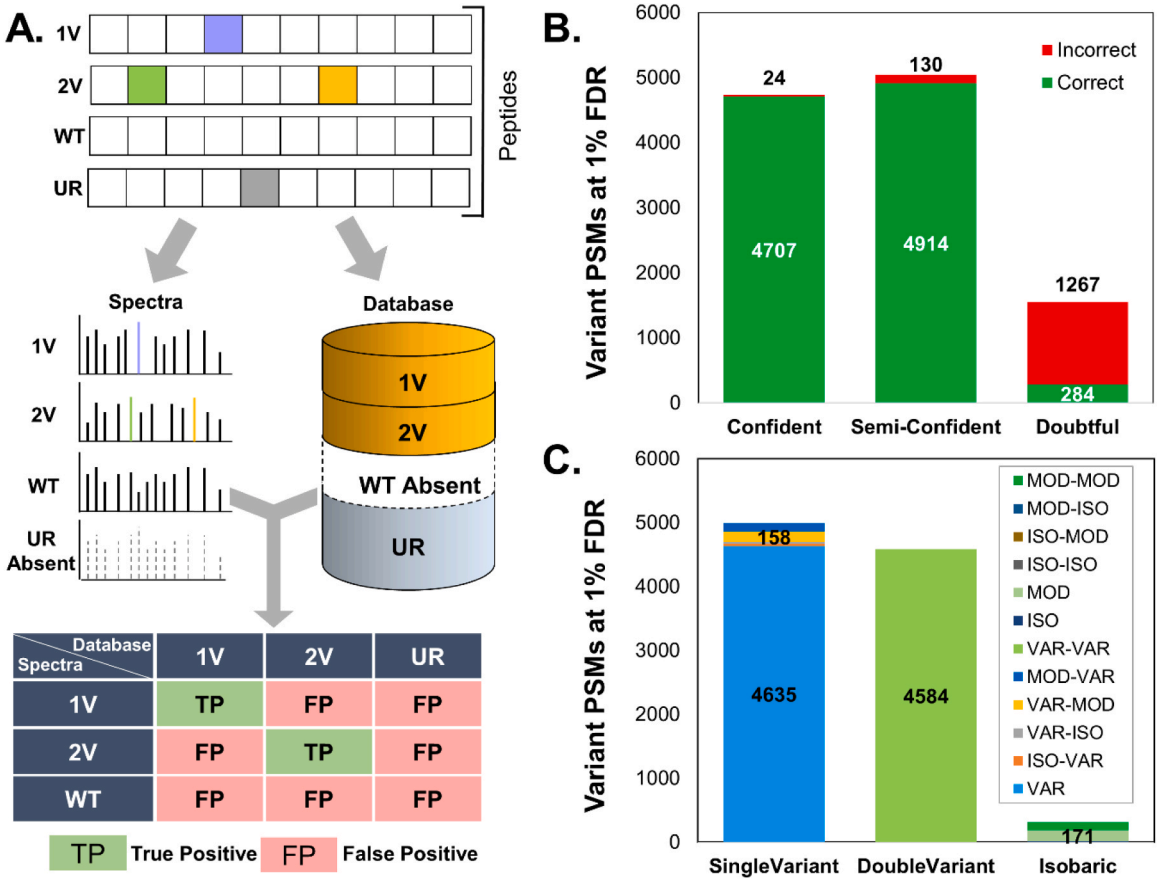

Fig. 2. (A) Workflow for generation of simulated spectra and the database construction with true and false positive evaluation matrix. Simulated spectra (5000 each) were created for single variants (1 V) and double variants (2 V), along with 40000 wild type peptides (WT). 500,000 unrelated variants (UR) were added to FASTA database but spectra were not simulated for these. The 1 V and 2 V peptides were included in FASTA but not WT peptides. Based on this, the evaluation table shows the definition of true and false variant peptides. (B) The proportion of correct and incorrect variants is shown in the PgxSAVy classified variants. (C) Proportion of single, double and isobaric PSMs after isobaric evaluation by PgxSAVy.

(PXD001468) were downloaded from PRIDE repository [50] and converted to MGF using MSConvert tool [51].

2.7. Database creation

Simulated data was search against SimDB containing 5000 single and 5000 double amino acid variant peptides with 500,000 unrelated (UR) variant peptides that had no overlap with the variant peptides used in simulation. For 40,000 wildtype peptides, spectra were generated but sequences were not added to database. The reference database (RefDB) downloaded from SwissProt (May 2022) contained 42,362 protein sequences. The human GENCODE database (version 40) [52] containing 246,624 transcripts were translated in three-frames, resulting in 1676, 122 ORFs (TranscriptDB) using an in-house Perl script. Human protein variant information from neXtProt database were downloaded (PEFF format) [53] to create VariantDB by tryptic digestion, allowing a maximum of two variant amino acid residues per peptide. The VariantDB thus created contained 93,669,236 variant peptides. To each of these databases, contaminant proteins downloaded from cRAP website (https://www.thegpm.org/crap/) were added to the FASTA before search.

2.8. Database search

2.8.1. Simulation data search

Simulation data was searched against SimDB using multiple database search engines (X!Tandem, OMSSA and MSGF+) with EuGenoSuite tool [19]. The search parameters were: 0.1 Da precursor tolerance, trypsin cleavage enzyme, no missed cleavage and carbamidomethylation at cysteine as fixed modification. The FDRscore method was used at 1% FDR [54].

2.8.2. Single search engine

One fraction from AD data (ad_pl01, hereafter referred as F1) was searched with MSGF+ search engine (version v2019.07.03). Searches were performed in 3-stage manner iteratively in ReferenceDB, TranscriptDB, and VariantDB, in which every subsequent search was conducted on the unassigned spectra from previous search. The parameters for the searches were: - 6 ppm precursor mass tolerance, 2 missed cleavages and trypsin as the proteolytic enzyme, fixed modification of carbamidomethylation at cysteine and variable modification of methionine oxidation and FDR 1%. However, missed cleavages were not allowed for third search.

2.8.3. AD data search

The 10 fractions (F1-F10) from AD dataset were searched with EuGenoSuite [19] in a stage-wise manner against ReferenceDB, TranscriptDB, and VariantDB in which every subsequent search was conducted on the unassigned spectra from previous search. The parameters for the searches were: - 6 ppm precursor tolerance, 2 missed cleavages, trypsin proteolytic enzyme, fixed modification of carbamidomethylation at cysteine, variable modification of methionine oxidation and 1% FDR. For third stage search, no missed cleavage was allowed.

2.8.4. HEK data search with multiple search engines

All fractions of HEK data were also searched in multi stage manner using EuGenoSuite in similar manner as described for AD dataset. The parameters were: - 5 ppm parent mass tolerance, 0.01 Da fragment mass tolerance, trypsin with one missed cleavage, Carbamidomethylation at cysteine and oxidation at methionine as fixed and variable modifications and 1% FDR.

2.9. Global and class-specific false discovery rate (gFDR & cFDR) estimation

For gFDR estimation, all PSMs from the three stage-searches were combined together before FDR using combinedFDR method [54] for integrating multiple search algorithms. For cFDR, the identified peptides were classified into – known, novel and variant peptides, and FDR estimated separately for each class. For both gFDR and cFDR, the formula remained the same although the input is different:

$$FDR_x = \frac{2 * D_x}{T_x + D_x}$$

where,

x = category of FDR input (all peptides, or class –i.e. known, novel or variants),

FDR_x = false discovery rate,

D_x = number of decoy PSMs above threshold,

T_x = number of target PSMs above threshold.

The variant peptides passing the 1% FDR threshold are selected for PgxSAVy analysis for quality assessment and annotation.

2.10. Manual annotation of spectra for creating gold standard for VAS evaluation

We performed rigorous manual annotation of variant PSMs identified in F1 fraction of AD through multiple search engines. For the identified 797 variant PSMs at 1% FDR, the study authors independently annotated each PSM as “good” (acceptable) or “bad” (unacceptable) matches. Majority-voted “good” and “bad” PSM annotations were accepted, while the rest (ties) were labelled as “average” (ambiguous) matches. The annotated PSM images for the spectra were generated through PLabel [55] using batch mode.

2.11. Variant evaluation through PgxSAVy

The variants in different searches were analyzed through PgxSAVy using tab-separated input format with the search-specific parameters.

2.12. STRING analysis

For STRING (version 12) analysis [56], all proteins with variants were compared against selected proteins with reliable variants (confident and semi-confident). The networks were created with high confidence threshold of 0.9.

3. Results and discussion

3.1. PgxSAVy Workflow

The overview of current study is shown in Fig. 1A. Three different datasets from simulation, AD and HEK were used in this study to evaluate variants through PgxSAVy tool, the workflow for which is explained in Fig. 1B. In PgxSAVy tool, we have integrated variant peptide rescoring, scoring variant PSM quality through VAS, isobaric assessment and disease related annotation of the variants. We tested the method using multiple datasets of increasing complexity. We used a simulation dataset, another dataset with manually annotated variant quality, and large-scale datasets for examining the variant quality through PgxSAVy.

The simulation data represents ideal scenario with near perfect separation of good and bad PSMs. The databases searches were conducted in multi-stage manner using EuGenoSuite [19] for search and integrated ProteoStats for FDR estimation [57]. The combined FDR method [54] was then used to estimate the integrated FDR from multi-search algorithms at 1% gFDR as well as cFDR. Eventually, the results were analyzed through PgxSAVy, and from poor-quality hits that follow a normal distribution centred at zero (Supplementary Fig. S1), z-score and p-values were calculated.

We have implemented three types of annotation in PgxSAVy for

variant peptides: (i) quality annotation (described above), (ii) variant class annotation (single variant, double variant, or isobaric), and (iii) disease annotation. Quality annotation includes assigning variant peptide quality such as - confident, semi-confident and doubtful based on p-values calculated from VAS. The confident, and semi-confident variants are considered true variants and doubtful are considered as false. Isobaric analysis determines if any variant amino acid in a peptide has a mass shift which can be explained with post translational modification shift. For peptides with two amino acid variants, it is classified based on whether both amino acids are variants, or a single variant with isobaric or modification mass (Supplementary Fig. S2). The disease annotation includes annotation of these variant peptides with biological information present in the UniProt database about its clinical significance and phenotype.

3.2. Evaluation of VAS on simulated data

Simulated data for 10,000 variant peptides containing one or two variants per peptide sequence (5000 each) was generated using MaSS-Simulator (Supplementary Table S1 and Fig. 2A). Additionally, 40,000 spectra were simulated from unrelated wild type peptides. All 50,000 spectra were generated with 20% noise. The database contained all sequences except the 40,000 WT peptides to simulate false variant matches. We added 500,000 variant sequences in database for which no spectra were generated, and any spectra matching to these sequences will also be false. This enabled us to rigorously test the PgxSAVy pipeline with known true and false hits. The data was searched using EuGenoSuite and variant PSMs at 1% FDR were evaluated with PgxSAVy.

Since we already know the correct and incorrect variants in the simulation data, we can easily calculate the number of false positive and false negative variants at the FDR threshold. When incorrect peptides were assigned to the variant spectra, these were termed as false positives. When the variant spectra could not be identified above FDR, despite being present in the data, they were termed as false negatives. The density plots of various features used in PgxSAVy on simulation results is shown in Supplementary Fig. S3, and the density distribution of correct and incorrect variants show good separation by VAS (Supplementary Fig. S4). After removing decoy and contaminants from 1% FDR results, a total of 11326 PSMs remained, of which the true and false positives were 9905 (87.45%) and 1421 (12.54%) respectively, while 95 (0.95%) were false negative variants. The real-world mass spectrometry data has more limitations than simulated data, due to several issues such as variable degree of ionization, missing or unknown peaks, poor or incomplete fragmentation, variable degree of chemical noise and intensities etc. This may lead to higher false positives in real-world compared to an ideal scenario (simulation). The false positive variants (12.54%) in the simulation results suggest that real-world data might have more false identifications. PgxSAVy classified 9775 variant PSMs as true (4731 confident and 5044 semi-confident), while the remaining 1551 PSMs as doubtful (Fig. 2B). From these 9775 PSMs, only 154 (1.57%) incorrect PSMs were misjudged as either confident or semi-confident, while 9621 (98.43%) were correctly identified as true. The true identification percentage in final results improved from 87.45% to 98.43% which shows that PgxSAVy enhances the true identification rates. Furthermore, only 284 (1.83%) of correct PSMs were misclassified as doubtful. Also, 1267 (89.16%) false variants were correctly labeled and removed. This shows PgxSAVy was able to effectively evaluate majority of true and false variants correctly. PgxSAVy found that 4994 and 4584 PSMs were correctly annotated from the simulated single and double variants (5000 each) respectively. Of this, 4635 (92.81%) and 4584 (91.68%) were correctly classified as single and double variants in isobaric analysis. A small number of variant PSMs were mis-annotated from double to single variant class or in isobaric class (Fig. 2C).

3.3. Evaluation on the manually validated dataset

Real world scenario has deviations due to sample complexity, technical variations, noise etc. However, it is a challenge to evaluate proteogenomics methods on real-world data, due to lack of well-established ground truth data. For this purpose, we manually annotated the variant spectra as *Good*, *Average* or *Bad* quality (see methods) from F1 fraction of the AD data to establish ground truth. The F1 data contains 172,337 spectra on which a proteogenomics search was conducted using EuGenoSuite, followed by variant evaluation by PgxSAVy.

The proteogenomics search led to identification of 60589 known, 185 novel and 797 variant PSMs, of which we selected the 797 variants for further analysis. Through manual validation, 221, 363 and 213 PSMs were categorized as good, average and bad category, respectively. The density and scatter plots of various features against VAS is shown in Supplementary Fig. S5 & S6 respectively.

Out of 797 variant PSMs, 270, 99 and 428 were identified as confident, semi-confident and doubtful PSMs respectively. The density distribution of manual annotation with PgxSAVy quality annotations reveals that their VAS distributions (Fig. 3A & B) as well as their means are somewhat similar (Fig. 3C & D) despite some disagreements in the number of variant PSMs in each class-wise comparison (Supplementary Table S3). We observed that VAS was slightly stricter than manual annotation (especially for average quality PSMs). Thus, it is possible that some average labeled PSMs were either classified as confident or as doubtful by VAS. The manual assignment of PSM quality is rather subjective and therefore, these categories may not fully agree with VAS. The distribution of semi-confident variants is much closer to the distribution of doubtful ones which suggests that semi-confident variant hits should not be trusted as true variant peptide identifications. However, since semi-confident hits only partially overlap with doubtful ones but overlap more with confident ones (Fig. 3B), we have kept them for user assessment so that VAS is not overtly stringent. Also, if any variant seems important to the user and lies in semi-confident category, the semi-confident label will allow the user to reassess the variant.

3.4. Can multi-algorithm searches and cFDR remove false variants?

We have demonstrated that PgxSAVy is highly effective in quality control of variant PSMs. For manually annotated evaluation, we had selected 1% cFDR results from multiple search engines based on recommendations from previous variant proteogenomic studies [35,36]. Alfaro et al. suggested employing multiple algorithms, the cFDR method, and various post-search filters help us to identify confident variant peptides [36]. However, these aspects have not been systematically evaluated. Here, we aimed to evaluate these approaches to control variant false positives by using (i) multiple algorithms, (ii) cFDR, and (iii) multiple algorithms with cFDR to find the most suitable approach.

We sought to compare the degree of false removal in single vs multiple algorithm search, as well as gFDR vs cFDR, using F1 data. The F1 fraction of AD data was searched with the MSGF+ Search algorithm and EuGenoSuite in a three-stage search with gFDR and cFDR estimation. The complete results obtained are provided in Table 1. We identified 1710 variant PSMs from MSGF+ search, of which 287 are classified as confident, 239 as semi-confident and 1184 as doubtful. From multiple algorithms, we obtained 728 variant PSMs, of which 232 are classified as confident, 98 as semi-confident and 398 as doubtful (Fig. 4A & Table 1). While a single algorithm identifies more variants, a large proportion of them (786 PSMs) is false. It is important to note that neither method is able to completely remove false variants in gFDR. Thus, we recommend multiple algorithms searches for variant proteogenomics. We also asked if cFDR method can improve these results. Applying cFDR to MSGF+ search results identified 847 variant PSMs, of which 246 were classified as confident, 95 as semi-confident and 473 as doubtful by PgxSAVy (Fig. 4B). This shows that cFDR performs better than gFDR for single search engine. The figure also shows that while cFDR with

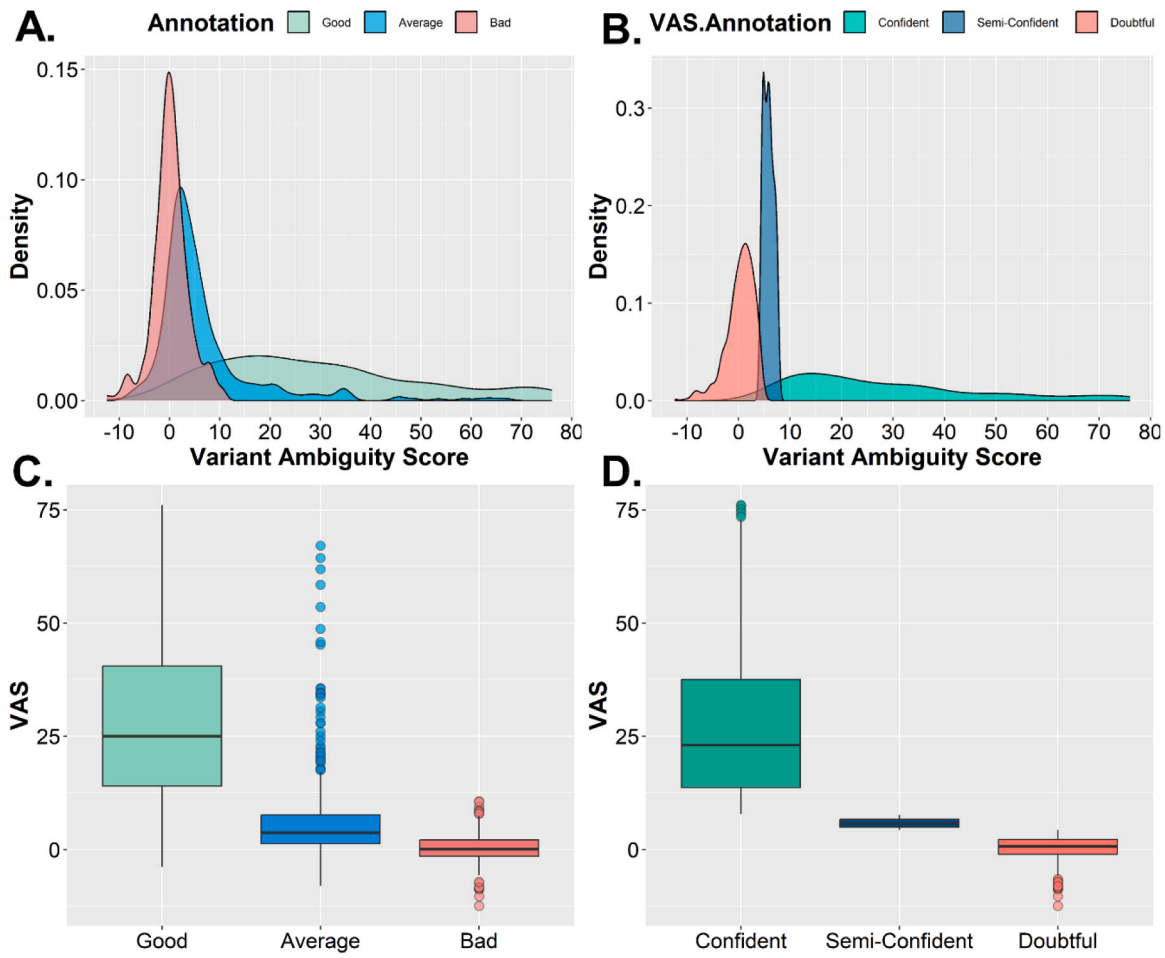

Fig. 3. (A) Distribution of all variant PSMs identified in brain AD dataset (first fraction) with manually annotation. (B) Distribution of all variant PSMs identified in brain AD dataset (first fraction) with VAS quality annotation. (C) Boxplot showing the average number of variant PSMs in each manual annotation class for brain AD database (first fraction). (D) Boxplot showing the average number of variant PSMs in each VAS annotation class for brain AD database (first fraction).

Table 1
Single and multiple search engine results with global and class specific FDR and VAS classification on variant PSMs.

		Known	Novel	Variant			
				Total	Confident	Semi-Confident	Doubtful
Single Search Algorithm	gFDR	57962	286	1710	287	239	1184
	cFDR	60743	243	814	246	95	473
Multiple Search Algorithm	gFDR	60634	53	728	232	98	398
	cFDR	60589	185	797	270	99	428

multiple search engines increases false variants slightly, the overall gain in correct variants is higher than gFDR and PgxSAVy can remove the false hits. The analysis of gFDR and cFDR on AD and HEK data are provided in supplementary Fig. S7 and supplementary tables S5 and S6. This demonstrates that multiple algorithms and cFDR combination is more potent in reducing false variants while keeping the sensitivity high. We recommend that this combination is best for variant identification.

The gFDR and cFDR results were compared to evaluate whether the VAS annotated PSMs also agree with manual annotations for multiple algorithms (Fig. 4C). It was observed that all PSMs in gFDR were encompassed within cFDR set. Out of 69 variants exclusive to cFDR, the exclusive manual (15 good, 31 average and 23 bad) annotations depict that although 46 variant PSMs were identified as confident/semi-confident, it also allowed 23 doubtful PSMs to pass (Fig. 4C). Even though multiple algorithms with cFDR perform better, even this combination is not fully capable of removing false variants. The Fig. 4D shows that similar VAS is scored for the three categories in both gFDR and cFDR. It should not escape notice that despite these FDR methods, the false variants are still looming large in data, and thus tools like PgxSAVy will be indispensable for variant proteogenomics.

Thus, we have established the utility of PgxSAVy through simulation and manual annotation data. We have also established that multiple search engine workflow combined with cFDR method is best suited for variant proteogenomics followed by PgxSAVy evaluation of the variants thus identified.

3.5. *Analysis of variants in AD and HEK datasets*

We applied PgxSAVy to two independent datasets (AD and HEK) to evaluate the effect of PgxSAVy on real world scenario with its implications on final biological outcome.

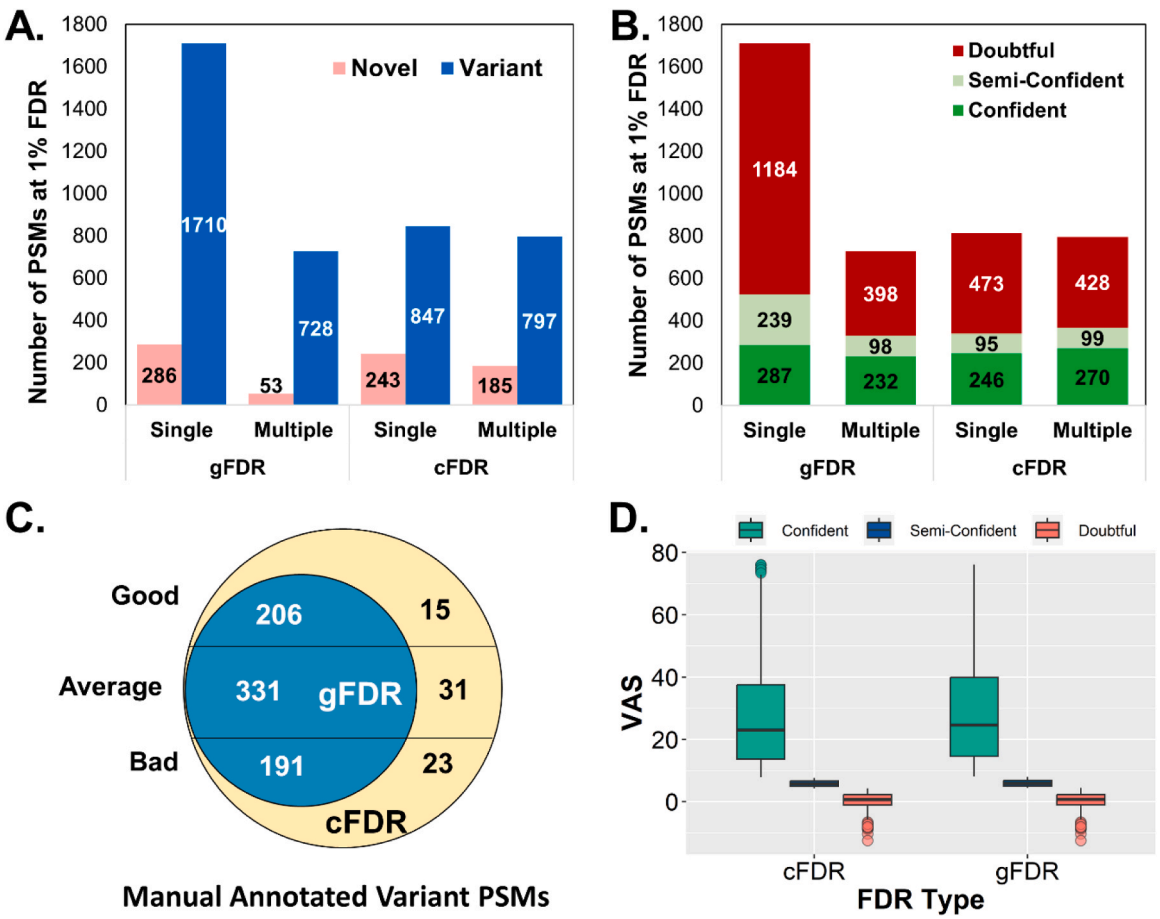

Fig. 4. (A) Novel and Variant PSMs identification in single and multiple search engine results passing 1% FDR (B) PgxSAVy classification on variant PSMs showing higher ratio of classified confident PSMs with multiple search algorithm and cFDR (C) cFDR results encompass gFDR results and both methods have non-negligible number of PSMs categorized as bad by manual annotation. (D) VAS rescoring showing similar scores in for cFDR and gFDR for multiple search engine results but both methods are prone to false positives in variant PSMs which VAS can control better than the two methods of FDR.

The AD dataset, containing 1718,768 spectra from Alzheimer's disease patients, was searched using EuGenoSuite, from which a total of 595,892 known, 2671 novel, and 8883 variant PSMs were identified at 1% cFDR. After removing contaminants and decoy hits, the identified PSMs were 582986, 2671, and 8777 for the known, novel and variant PSMs respectively.

We selected the 8777 variant PSMs at 1% cFDR threshold for further analysis. PgxSAVy classified 2084 variant PSMs as confident, 987 as semi-confident and 5706 as doubtful (Table 2 and Fig. 5A & B). From the reliable (confident and semi-confident) 3071 variant PSMs (1891 non-isobaric), we identified 782 unique variant peptides (500 non-isobaric) (Fig. 5C).

In the HEK data containing 1.12 million spectra, a total of 471,823 known, 1658 novel and 4015 variant PSMs were identified at 1% cFDR, of which 3928 remained after removing contaminants and decoys. PgxSAVy analysis classified 945 variant PSMs as confident, 424 as semi-confident and 2559 as doubtful (Table 2 and Fig. 5D & E). From the reliable (confident and semi-confident) 1369 PSMs (929 non-isobaric), we identified 626 unique variant peptides (434 non-isobaric) (Fig. 5F). After removing doubtful PSMs, 1159 single variant, 732 double variants and 1180 isobaric PSMs were left in AD data; while 644 single variant, 285 double variants and 440 isobaric PSMs remained in HEK data (Fig. 5 C & F). PgxSAVy allowed detection of variant peptides in the HEK data, enabling the investigation of their role in cellular processes and disease mechanisms.

Table 2
VAS classification for Brain AD dataset and HEK Cell line.

VAS	AD dataset	HEK Cell line
Total	8777	3928
Confident	2084	945
Semi-confident	987	424
Doubtful	5706	2559

The variants accepted after analysis are represented as sunbursts in Supplementary Fig. S8 and S9 that depict the frequency of original to variant amino acids conversion in the datasets respectively. Notably, in both independent datasets, even after several algorithms and strict cFDR, many false positives can sneak in, but PgxSAVy effectively removed those false variants, despite maintaining higher identification rate than previous study, which did not apply such strict filter. Approximately one-third of the variant peptides remained after stringent filtering by PgxSAVy and denotes its ability to identify the most reliable variant peptides. Furthermore, Supplementary Fig. S10 displays the matched spectra of some identified variant peptides from different classes.

To investigate the biological significance of the findings, we analyzed the variants discovered in AD and HEK data in more detail. Fig. 6 shows some variants identified in AD and HEK datasets. Neuromodulin protein (P17677–2) that plays a role in nerve growth was identified with mutations A214G (rs182766826) and T217P (rs200008838) in variant peptide QADVPAAVTAAGATTPVAEDAAAK supported with 20 reliable PSMs. Another protein Synapsin-2 (Q92777–2) was identified with mutation A127S (rs978216738) in variant peptide VLLVVDEPHADWSK (13 confident PSMs). Myelin

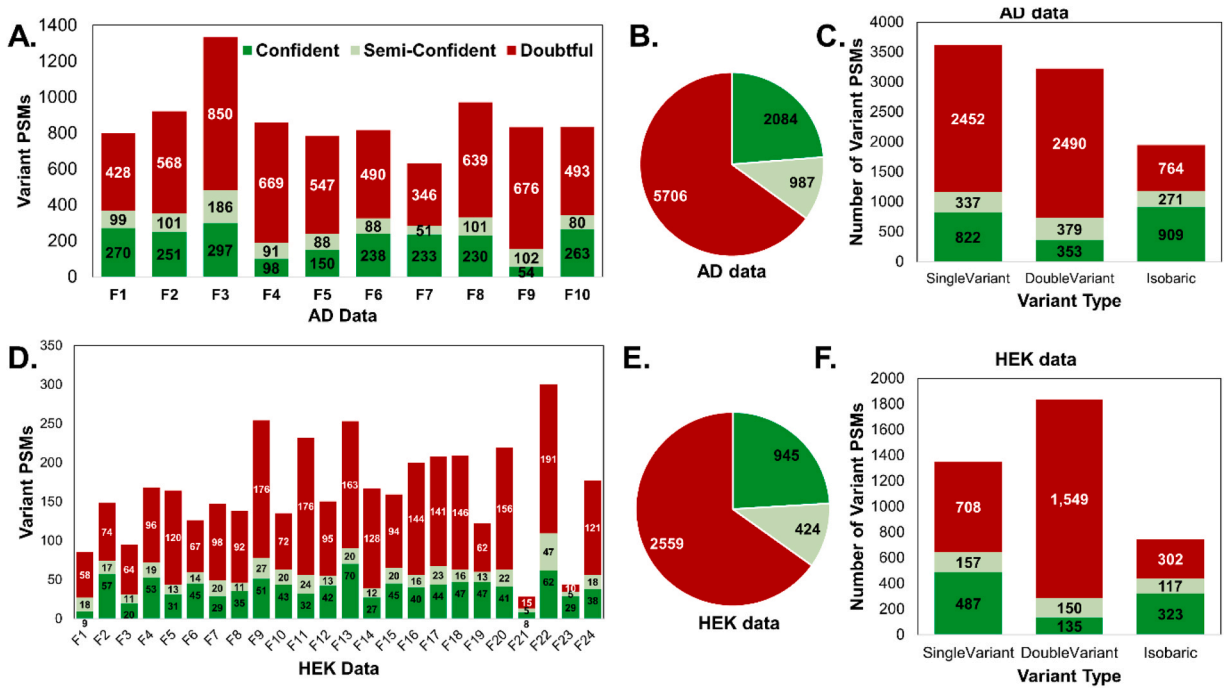

Fig. 5. Distribution of identified PSMs (A) Variant PSMs classified by PgxSAVy for AD dataset for F1-F10 (B) Total Variant PSMs classified by PgxSAVy for AD data. (C) Isobaric analysis of reliable variants for AD data. (D) Variant PSMs classified by PgxSAVy for HEK dataset for F1-F24. (E) Total Variant PSMs classified by PgxSAVy for HEK data. (F) Isobaric analysis of reliable variants for HEK data.

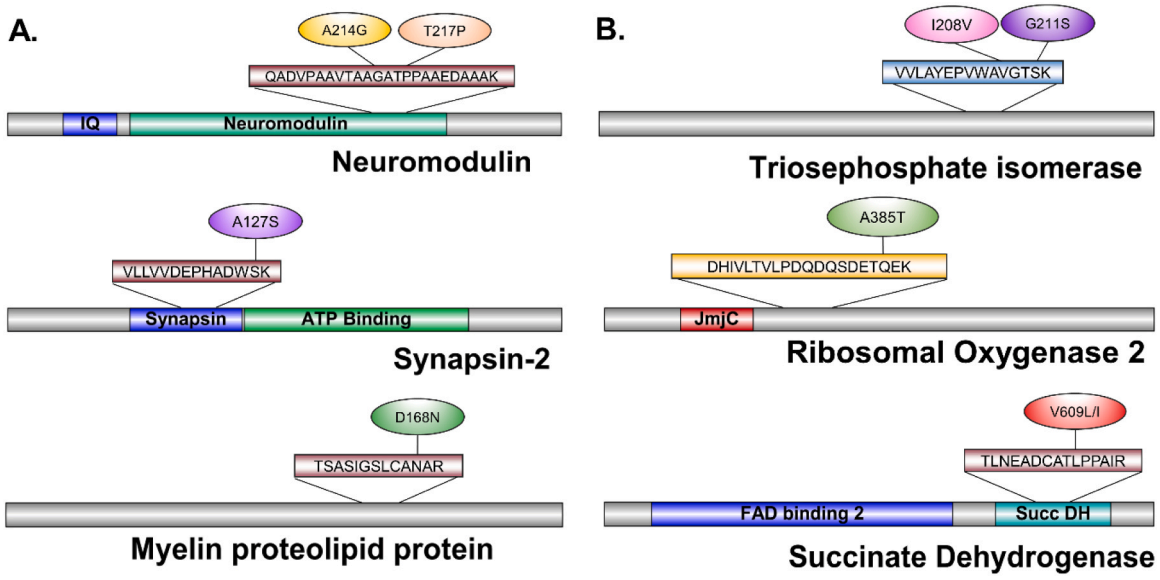

Fig. 6. Some identified variants in (A)AD data (B) HEK data.

proteolipid protein (P60201–2) was identified with one pathogenic mutation D168N (rs132630284) in variant peptide TSASIGSLCANAR (5 confident PSMs). In HEK data, a high-confidence peptide, TLNEADCAT [L/I]PPAIR with V609L/V609I mutation (rs6962) in succinate dehydrogenase (P31040–2) protein was identified with 88 confident PSMs and 3 semi-confident PSMs, which was also reported in original study [50]. For the same protein, we also found another variant peptide HTLSFVDVGTGK with Y581F mutation (rs6960) for the same protein which was not found earlier. Another new finding was Glucosidase 2 subunit beta (P14314–2), which has role in glycan metabolism, was also identified with variant peptide LGGSPTSLGTWSSWVGPDHDK having two mutations, G447S (rs761355123) and I450V (rs34351170). Ribosomal oxygenase 2 protein (Q8IUF8–4) was also identified with variant alanine-to-threonine, A385T (rs2172257) which plays an important role in cell growth and survival. We also discovered Triosephosphate isomerase (P60174–3) carrying a pathogenic mutation I208V, and another mutation G211S in peptide VVLAYEPVWAVGTSK with 8 reliable PSMs (7 confident, 1 semi-confident).

3.6. *Disease annotation of variants peptides in PgxSAVy*

The quality control of variant peptides is not sufficient for establishing the functional implication of the identified variant peptides and thus, mapping information about their functional biological or disease

impact is important. There are several rich sources of genomic variant assessment which need to be manually sifted for finding relevant information, a time consuming and tedious necessity. UniProt is a rich aggregated source of such information that we have leveraged towards an automated annotation of pathogenicity information for the identified variants.

PgxSAVy uses a downloaded annotation file from UniProt (https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/variants/homo_sapiens_variation.txt.gz) and maps the identified variant peptides for their annotations compiled in the UniProt database. This helps the biologists to make informed decisions about variant selection for further enquiry or planning validation experiments.

Using these annotations, we analyzed the important proteins using protein-protein interaction (PPI) database STRING. Here, we observed that while the PPI analysis of all proteins identifies all major pathways, PPI derived from a selective proteome set (only reliable variant peptides) leads to discovery of important pathways with mutated proteins. For example, glycolysis/gluconeogenesis mutations were highlighted in AD data; while spliceosome and actin folding in HEK data (Supplementary Fig. S11 and S12).

4. Conclusion

Controlling false identifications in variant peptides identified through proteogenomics is a daunting task. The cFDR approach and various ad-hoc filters have limitations of scale and are applied non-uniformly across labs, which makes it challenging to standardize or automate such approaches. We demonstrated that multiple algorithms perform better than a single algorithm for curtailing the false variant effect, but cannot completely remove it. Furthermore, even though the cFDR method provides better results than gFDR, it remains ineffective in reducing false variants. PgxSAVy performs an automated, comprehensive variant quality control, as well as provides annotations for the confidently identified variants to highlight their biological implications. We demonstrated the utility of PgxSAVy on a simulated variant dataset and a manually validated dataset. Applying PgxSAVy on variant peptides that were identified in proteogenomic reanalysis of large-scale public datasets, we show that a large number of variants were likely to be false and it is critical to employ quality control measures before interpretation. After a comprehensive scrutiny, we discovered several variants described in the original studies as well as discovering ones not described earlier.

In summary, our study underscores the effectiveness of PgxSAVy in controlling false positives in variant proteogenomics. We provide insights into its performance across different datasets and emphasize the biological significance of the findings. PgxSAVy is a handy tool that can perform post search variant quality control. It signifies the need for removing false variants that inadvertently use up limited critical resources for chasing incorrect findings. It helps in biological interpretation of variants by proving facile integration with UniProt knowledgebase for variant annotation and highlighting their role in diseases. Thus, PgxSAVy is available as an open-source, accessible, comprehensive and automated tool as well as an easy-to-use webserver, for performing variant quality control and post-search-filters in variant proteogenomics.

Funding

Authors acknowledge funding support from DST-INSPIRE (DST/INSPIRE Fellowship/2016/IF160360) to AR and MK Bhan – Young Researcher Fellowship Programme (MKB-YRFP) grant (HRD-12/4/2020-AFS-DBT-Part(1)(13815)) to SA, AKY is supported by Translational Research Program (TRP) at THSTI funded by DBT, SERB-SUPRA (SPR/2020/000315), and THSTI Intramural grants. DD is supported by DBT-BIC grant (BT/PR40260/BTIS/137/33/2021). DD and AKY acknowledge support from DBT-NNP grant (BT/PR40269/BTIS/137/62/2023).

CRediT authorship contribution statement

Anurag Raj: Formal analysis; Data curation; Investigation; Methodology; Writing – original draft. **Suruchi Aggarwal**: Formal analysis; Methodology; Investigation; Writing – review & editing. **Prateek Singh**: Webserver development **Amit Kumar Yadav**: Conceptualization; Formal analysis; Data Curation; Investigation; Methodology; Writing – review & editing; Funding Acquisition; Supervision. **Debasis Dash**: Conceptualization; Formal analysis; Data Curation; Investigation; Methodology; Writing – review & editing; Funding Acquisition; Supervision.

Declaration of Generative AI and AI-assisted technologies in the writing process

The authors declare that the manuscript is written completely by the authors and no help of generative AI was used in the process.

Competing Interests

The authors declare no competing interests.

Data availability statement

PgxSAVy can be accessed through https://pgxsavy.igib.res.in/ as a webserver and https://github.com/anuragraj/PgxSAVy as a stand-alone tool. All data are incorporated into the article and its online supplementary material. Search Database (FASTA) files are uploaded to Figshare and download links are provided in supplementary note 1.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.12.033.

References

[1] Giri K, Maity S, Ambatipudi K. In silico data mining of human body fluids to unravel the immunomes in breast cancer. J Proteins Proteom 2021;12:45–62.
[2] Li Y, Zhang Y, Pan T, et al. Shedding light on the hidden human proteome expands immunopeptidome in cancer. Brief Bioinform 2022;23.
[3] Anurag M., Jaehnig E.J., Krug K., et al. Proteogenomic Markers of Chemotherapy Resistance and Response in Triple-Negative Breast Cancer. Cancer Discov. 2022; OF1–OF20.
[4] Mertins P, Mani DR, Ruggles KV, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature 2016;534:55–62.
[5] Johnson ECB, Carter EK, Dammer EB, et al. Large-scale deep multi-layer analysis of Alzheimer's disease brain reveals strong proteomic disease-related changes not observed at the RNA level. Nat Neurosci 2022;25:213–25.
[6] Patrie SM. Modern proteomics – sample preparation. Anal Pract Appl 2016:919.
[7] Gonzalez-Teran B, Pittman M, Felix F, et al. Transcription factor protein interactomes reveal genetic determinants in heart disease. Cell 2022;185:794–814.e30.
[8] Lin Y-Y, Gawronski A, Hach F, et al. Computational identification of micro-structural variations and their proteogenomic consequences in cancer. Bioinformatics 2018;34:1672–81.
[9] Yadav AK, Banerjee SK, Das B, et al. Editorial: systems biology and omics approaches for understanding complex disease biology. Front Genet 2022;13:12–4.
[10] Tolani P, Gupta S, Yadav K, et al. Big data. Integr omics Netw Biol 2021:127–60.
[11] Barbieri R, Guryev V, Brandsma C-A, et al. Proteogenomics. Key Driv Clin Discov Pers Med 2016:21–47.
[12] Karimi MR, Karimi AH, Abolmaali S, et al. Prospects and challenges of cancer systems medicine: from genes to disease networks. Brief Bioinform 2021;00:1–31.
[13] Craig R, Beavis RC. TANDEM: Matching proteins with tandem mass spectra. Bioinformatics 2004;20:1466–7.
[14] Geer LY, Markey SP, Kowalak JA, et al. Open mass spectrometry search algorithm. J Proteome Res 2004;3:958–64.
[15] Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun 2014;5:1–10.
[16] Kelkar DS, Kumar D, Kumar PP, et al. Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. Mol Cell Proteom 2011;10:1–13.

[17] Kumar D, Mondal AK, Yadav AK, et al. Discovery of rare protein-coding genes in model methylotroph methylobacterium extorquens AM1. Proteomics 2014;14:2790–4.
[18] Kumar D, Yadav AK, Kadimi PK, et al. Proteogenomic analysis of bradyrhizobium japonicum USDA110 using genosuite, an automated multi-algorithmic pipeline. Mol Cell Proteom 2013;12:3388–97.
[19] Kumar D, Yadav AK, Jia X, et al. Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation. Mol Cell Proteom 2016;15:329–39.
[20] Wang X, Zhang B, Wren J. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics 2013;29:3235–7.
[21] Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007;4:207–14.
[22] Käll L, Storey JD, MacCoss MJ, et al. Posterior error probabilities and false discovery rates: Two sides of the same coin. J Proteome Res 2008;7:40–4.
[23] Aggarwal S, Yadav AK. False discovery rate estimation in proteomics. Methods Mol Biol 2016;1362:119–28.
[24] Salz R, Bouwmeester R, Gabriels R, et al. Personalized Proteome: Comparing Proteogenomics and Open Variant Search Approaches for Single Amino Acid Variant Detection. J Proteome Res 2021;20:3353–64.
[25] Aggarwal S, Raj A, Kumar D, et al. False discovery rate: the Achilles' heel of proteogenomics. Brief Bioinform 2022:1–15.
[26] Woo S, Cha SW, Na S, et al. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. Proteomics 2014;14:2719–30.
[27] Woo S, Cha SW, Bonissone S, et al. Advanced proteogenomic analysis reveals multiple peptide mutations and complex immunoglobulin peptides in colon cancer. J Proteome Res 2015;14:3555–67.
[28] Noble WS. Mass spectrometrists should search only for peptides they care about. Nat Methods 2015;12:605–8.
[29] Menschaert G, Fenyö D. Proteogenomics from a bioinformatics angle: a growing field. Mass Spectrom Rev 2017;36:584–99.
[30] Borchert N, Dieterich C, Krug K, et al. Proteogenomics of Pristionchus pacificus reveals distinct proteome structure of nematode models. Genome Res 2010;20:837–46.
[31] Chaerkady R, Kelkar DS, Muthusamy B, et al. A proteogenomic analysis of Anopheles gambiae using high-resolution Fourier transform mass spectrometry. Genome Res 2011;21:1872–81.
[32] Merrihew GE, Davis C, Ewing B, et al. Use of shotgun proteomics for the identification, confirmation, and correction of C. elegans gene annotations. Genome Res 2008;18:1660–9.
[33] Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods 2014;11:1114–25.
[34] Yi X, Wang B, An Z, et al. Quality control of single amino acid variations detected by tandem mass spectrometry. J Proteom 2018;187:144–51.
[35] Choong WK, Sung TY. Multiaspect examinations of possible alternative mappings of identified variant peptides: a case study on the HEK293 cell line. ACS Omega 2022:1–9.
[36] Alfaro JA, Ignatchenko A, Ignatchenko V, et al. Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines. Genome Med 2017;9.
[37] Li J, Su Z, Ma Z-Q, et al. A bioinformatics workflow for variant peptide detection in shotgun proteomics. Mol Cell Proteom 2011;10:M110.006536.
[38] Zhu Y, Orre LM, Johansson HJ, et al. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nat Commun 2018;9:903.
[39] An Z, Zhai L, Ying W, et al. PTMiner: localization and quality control of protein modifications detected in an open search and its application to comprehensive post-translational modification characterization in human proteome*. Mol Cell Proteom 2019;18:391–405.
[40] Li Y, Wang X, Cho J-HH, et al. JUMPg: an integrative proteogenomics pipeline identifying unannotated proteins in human brain and cancer cells. J Proteome Res 2016;15:2309–20.
[41] Yadav AK, Kumar D, Dash D. MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. J Proteome Res 2011;10:2154–60.
[42] Yadav AK, Kumar D, Dash D. Learning from decoys to improve the sensitivity and specificity of proteomics database search results. PLoS One 2012;7:1–10.
[43] Fermin D, Walmsley SJ, Gingras AC, et al. LuciPHOr: algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach. Mol Cell Proteom 2013;12:3409–19.
[44] Aggarwal S, Tolani P, Gupta S, et al. Posttranslational modifications in systems biology. Proteom Syst Biol 2021;127:93–126.
[45] Aggarwal S, Gupta P, Dhawan U, et al. The language of posttranslational modifications and deciphering it from proteomics data. Transcr Transl Heal Dis 2023:109–36.
[46] Jeong K, Kim S, Bandeira N. False discovery rates in spectral identification. BMC Bioinforma 2012;13:S2.
[47] Deutsch EW, Lane L, Overall CM, et al. Human proteome project mass spectrometry data interpretation guidelines 3.0. J Proteome Res 2019;18:4108–16.
[48] Awan MG, Saeed F. MaSS-simulator: a highly configurable simulator for generating MS/MS datasets for benchmarking of proteomics algorithms. Proteomics 2018;18:1–4.
[49] Wang H, Yang Y, Li Y, et al. Systematic optimization of long gradient chromatography mass spectrometry for deep analysis of brain proteome. J Proteome Res 2015;14:829–38.
[50] Chick JM, Kolippakkam D, Nusinow DP, et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol 2015;33:743–9.
[51] Chambers MC, MacLean B, Burke R, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 2012;30:918–20.
[52] Frankish A, Diekhans M, Jungreis I, et al. GENCODE 2021. Nucleic Acids Res 2021;49:D916–23.
[53] Zahn-Zabal M, Michel PA, Gateau A, et al. The neXtProt knowledgebase in 2020: Data, tools and usability improvements. Nucleic Acids Res 2020;48:D328–34.
[54] Jones AR, Siepen JA, Hubbard SJ, et al. Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. Proteomics 2009;9:1220–9.
[55] Wang L, Li D-Q, Fu Y, et al. pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. Rapid Commun Mass Spectrom 2007;21:2985–91.
[56] Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res 2023;51:D638–46.
[57] Yadav AK, Kadimi PK, Kumar D, et al. ProteoStats - A library for estimating false discovery rates in proteomics pipelines. Bioinformatics 2013;29:2799–800.