Quality control and annotation of variant peptides identified through Proteogenomics

Anurag Raj[a,b], Suruchi Aggarwal[c], Amit Kumar Yadav[c*], Debasis Dash[a,b*]

[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), Translational Health Science and Technology Institute, NCR Biotech Science Cluster, 3rd Milestone, Faridabad, Haryana, India*

Correspondence:

Debasis Dash, PhD (ddash@igib.res.in)

G. N. Ramachandran Knowledge Centre for Genomics Informatics,

CSIR – Institute of Genomics and Integrative Biology, New Delhi, India

And

Amit Kumar Yadav, PhD (amit.yadav@thsti.res.in)

Computational and Mathematical Biology Centre (CMBC), Translational Health Science and Technology Institute, NCR Biotech Science Cluster, 3rd Milestone, Faridabad, Haryana, India

Abstract

The importance of single nucleotide polymorphisms (SNPs) is well known in diseases but which ones are translated to form single amino acid variant (SAV or variant) is not studied in detail. There may be translational potential of such information but this necessitates integrated analysis of genomics and proteomics data. Identification of novel and variant peptides using proteogenomics has paved the way to understand their unique and diverse phenotypic relationships in health and disease. Proteogenomic studies on cancer, neurological and cardiovascular diseases have revealed many variants of clinical importance. However, false positives owing to large database are a major challenge in proteogenomics, and even at strict FDR, it is a challenge to segregate the true from false variant hits. Some approaches have been suggested to circumvent this problem, such as class-specific FDR estimation and filtering workflows for better sensitivity. Implementation of these methods and workflows on database search results is challenging for biologists and requires advanced bioinformatics skills. Owing to their computational complexity, such methods are not readily accessible to biologists. The goal of this study was to develop an accessible tool for quality control of variant peptides. We analyzed many descriptors that pertain to PSM match quality, variant events and peptide matches to evaluate their ability to distinguish true variants from false positives in proteogenomics. These features were used to develop a variant ambiguity score (VAS) which was implemented into the tool PgxSAVy, which has the framework to re-score the proteogenomics variants resulting from single or multiple search algorithms, to classify the SAVs based on their quality of evidence. To evaluate VAS, we tested it on a simulated data with true and false SAVs and observed that it was able to segregate true and false variants effectively (true hits 86.6% and sensitivity 95.93%). Manual annotation of identified variant PSMs from one fraction of a large-dataset (PXD004010) also demonstrated that VAS was highly effective in segregating true and false variants in automated manner. We also used

large public data with approximately 2.8 million spectra (PXD004010 and PXD001468) for a comprehensive evaluation of PgxSAVy. Using these datasets, PgxSAVy identified and filtered ~50% false variants which suggests that proteogenomics variants be rigorously tested before making biological conclusions. We also integrated current knowledge on variants using an annotation framework in PgxSAVy to annotate the variants based on their known role in diseases. PgxSAVy provides a rigorous framework for quality control and annotations of variant peptide results from one or more search algorithms and helps the researchers to prioritize these variants for further studies.

Keywords:

Variant Proteogenomics, Single amino acid variations, Mass spectrometry, Peptide identification, False discovery rate, Quality Assessment, PgxSAVy

1. Introduction

Identification and characterization of non-synonymous variations in proteins is important for understanding diverse phenotypic changes to cells in health and diseases. Non-synonymous variants present in protein-coding regions can alter amino acids in the protein sequences which in turn, can alter the structure, interactions or activity of the corresponding proteins. Even with 93 % of human proteome being identified (Omenn *et al.*, 2022), not all sequence variants of these proteins are known. Mass spectrometry data can be used for large scale identification of peptides and variant peptides in proteogenomics approach, which has emerged as a powerful tool for identification and characterization of variants in various systemic disorders like cancer, Alzheimer's, Parkinson's, diabetes and cardiovascular diseases (Mertins *et al.*, 2016; Patrie, 2016; Anurag *et al.*, 2022; Gonzalez-Teran *et al.*, 2022; Johnson *et al.*, 2022).

In a proteogenomics search, a customized search database is created to identify variant or novel peptides (Kelkar *et al.*, 2011; Kumar *et al.*, 2014), which can be accomplished in several ways (Woo, Cha, Merrihew, *et al.*, 2014; Zickmann and Renard, 2015; Kumar, Yadav and Dash, 2017; Agosto *et al.*, 2019; Lau *et al.*, 2019). Some tools create the proteogenomic database by six frame translation of genomics sequence of the organism for which the search is to be conducted. However, the database search on this customized database becomes too exhaustive so as to accommodate all possible reading frames. Another popular approach is using the proteomics and transcriptomics data from the same sample, and using the three-frame translation of the transcriptome to create the proteogenomics database (Kumar *et al.*, 2013, 2016a; Wang, Zhang and Wren, 2013). This method enables identification of all possible variants that are expressed in the sample.

Even with the three-frame translation, the database search is still quite exhaustive and a well-known computational challenge. Moreover, proteogenomics databases tend to be humongous in size. This results in higher chances of false hits even after a strict false discovery rate (FDR) (Aggarwal and Yadav, 2016; Aggarwal *et al.*, 2022). Some tools that assist in performing large-scale proteogenomics searches are Spritz (Cesnik *et al.*, 2020), JUMPg (Li, Wang, J.-H. H. Cho, *et al.*, 2016), IPAW (Zhu *et al.*, 2018), EuGenoSuite (Kumar *et al.*, 2016b) and PPLine (Krasnov *et al.*, 2015) etc. A brute force variant search will require creation of all possible amino acid changes at every position for all peptides of the already huge database. This will be computationally intractable and large database size will lead to reduced sensitivity and higher rate of false positives. Some of the major challenges in variant identification by proteogenomics are highlighted in recent articles (Woo, Cha, Na, *et al.*, 2014; Noble, 2015; Woo *et al.*, 2015; Menschaert and Fenyö, 2017; Aggarwal *et al.*, 2022). Most proteogenomic studies calculate a global false discovery rate for all peptide identifications, including both known and novel peptides (Borchert et al. 2010; Chaerkady et

al. 2011; Merrihew et al. 2008). Some studies have employed a higher threshold of FDR for novel peptides and implemented more stringent filters, such as posterior error probability (Brosch et al. 2011) or separate FDRs for annotated and novel peptides (Branca et al. 2014). Even with stricter thresholds, the problem remains recalcitrant.

Though novel peptides from proteogenomics are of high interest, the major biological contributors are mutant peptides resulting from non-synonymous SNPs. Identifying and assessing variant events from MS data sets is therefore a critical process for confidence assessment and subsequent biological interpretation. The process of confidence assessment for variant peptides is critical and may require additional analysis. Choong *et al* suggested the post-examination of variant peptides by applying filters and checking criteria like isobaric substitutions. They had also shown that standard FDR estimation is not sufficient in controlling false positives in variant proteogenomics (Choong and Sung, 2022). However, their assessment criteria are deeply subjective and not amenable for automated verification of variant peptides. Due to lack of a tool, it also requires bioinformatics skills for suggested methods to implement on any proteogenomics search results. A robust and objective scoring system implemented as an automated method/pipeline can allow a better assessment of the quality of variants identified. There have been examples of new methods like PMD-FDR which used precursor mass discrepancy to detect potentially false PSM hits which were earlier assigned with high score but do not address false positives in variant identifications (Hubler *et al.*, 2019). SAVControl is a promising tool that filters variant peptides based on subgroup FDR and subsequently performs variant mass shift relocalization from its isomers (Yi *et al.*, 2018) and thus fails to fully capitalize on variant event information. These tools also have other limitations. The PMD-FDR method requires large number of annotated good and bad hits to learn from data and recalculate the confidence scores. SAVControl can only

process the results from Myrimatch and Mascot, and it also requires MATLAB runtime, thus limiting its use in general proteogenomics workflows.

Using stage-specific searches for the wild type peptides followed by variant peptides for unassigned spectra, helps in reducing false positive variants to some extent. Some studies also use multiple search engines to improve PSM identification rate and quality but it has not been applied to variant identifications in proteogenomics. Additional measures like spectral match quality, fractional intensity covered and manual verification are often employed to weed out false variants. Some studies have calculated spectral features or match quality descriptors to determine goodness of a variant PSM match which were used to calculate a score to select true variant peptide hits (Ivanov *et al.*, 2014). A good MS feature should have capability to discriminate between low scoring PSMs from better ones and these features may include measures like precursor mass difference, spectrum quality and retention time. The usage of retention time prediction based on machine learning has been demonstrated to be a good MS feature for quality control in variant peptides (Wen *et al.*, 2020; Bouwmeester *et al.*, 2021). However, retention time prediction requires a large training dataset for training the model limiting its global evaluation & implementation in variant proteogenomics. Therefore, an easily applicable variant quality measure that is easily calculated and implemented widely in tool agonistic pipelines is an important requirement for variant proteogenomics.

We have developed a universally applicable variant scoring method Variant Ambiguity Score (VAS), for rescoring and post processing the identified results from any proteogenomics pipelines. This method is based on an empirical scoring derived from various salient MS features. It re-evaluates the identified variant peptide and provides an integrated score based on PSM quality, search engine match quality and several other features like fractional intensity coverage and b/y-ions continuity count. These features add value in assessing the quality of variant peptide above and beyond search results. To demonstrate the effectiveness

of our method, we performed proteogenomic searches on various proteogenomic scenario – comprising Alzheimer's brain data, HEK293 cell line data and simulated variant mass spectrometry data. Our results reveal that around 50% false variant peptide hits were present above 1% FDR cut-off in variant proteogenomics which our method could successfully identify and remove. We have implemented this scoring method in an automated tool PgxSAVy which is easy to use and free for academic use. It can be used on tabular results of any search engine for rescoring of variants and their classification into confident, semi confident and ambiguous. The variant peptides that are confidently assigned, are also annotated for their function and role in disease by integrating information from dbSNP, cosmic etc. as provided through neXtProt.

2. Methods

2.1. Variant Ambiguity Score (VAS)

We analysed and evaluated different MS features for the identified variant peptides on literature based empirical knowledge and importance of MS features. Features like fractional intensity and its ratio with total intensity, peak counts, b and y ions and their continuity score has capability for evaluation of PSM quality. Complementary ions, such as neutral losses and immonium ions (depending on the type of instrument used), can aid in distinguishing between true and false hits when b and y ions counts are closely similar which is very useful for variant peptides. In addition, the continuity of a series (b/y/a) significantly enhances confidence in matched ions even when fragmentation is incomplete due to partially mobile or nonmobile proton-containing peptides. Some of these features are used by few search engines but not all features are taken together to estimate score of PSM match and are also not reported in final search results making post-search evaluation impossible. The descriptors can be grouped in three categories: MS/MS Scoring, Spectral quality and MS/MS features. These features have capability to discriminate between good and bad variant hits. Each feature can

depict the chances of PSMs coming from good or bad match through their density distribution. A distribution of the various MS features was analysed using real MS datasets.

Fractional intensity coverage

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. A peptide that covers most of the good quality MS/MS peaks for a spectrum gets a good fractional intensity coverage. Thus, this feature can distinguish between good and bad PSM quite well and is part of the variant scoring.

b and y ions continuity count

Occurrence of continuous b and y ions in a spectrum also provide information about the quality of spectrum and can establish a variant over wild-type peptide. More the number of continuous b and y ion ladders, more the indication of good fragmentation of peptide sequence and thus higher chances of identification of correct variant peptide. Using the counts of theoretical b & y ions to normalize the bias of scores due to variable peptide length, can help to determine the accuracy of variant peptide. Using this feature to normalize variant scores can provide a VAS score that can assess the presence of variant peptides.

Number of search engines identifying the variant peptide

Higher the number of search engines that identify a given scan with the same peptide sequence, 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. It could be the same primary peptide sequence with isomeric/different modification position or peptide sequence with slightly different variants. For variants, search engines are more likely to agree for true variants, while they may all report different peptides/variants for false hits.

Thus, using search engine count as a weightage factor for variant identification can add additional discriminatory power.

PSM count per peptide

For any identified peptide sequence, the chances of being a good hit increases with a greater number of identified PSMs. A majority of bad hits are identified with a single PSM and cannot be considered reliable and thus rightly discarded. The chances of random matching of peptides to single PSMs in proteomics is high, and for variant proteogenomics is much higher. So, PSMs count per peptide can prove to be a useful indicator for correct variant identification.

Score difference with decoy variant peptides

Another method for assigning confidence to variant peptide is by deriving the correct location of variant amino acid from all the other possible positions in the variant peptide. Previously, post translational modifications studies have used this strategy to build a rescoring to correctly identify the modification site (Fermin *et al.*, 2013; Aggarwal *et al.*, 2021). The hypothesis behind this strategy is that a decoy variant peptide created by swapping the variant amino acid identified with all other amino acids in variant peptide should always score less than a true identified variant peptide. This localization score builds confidence by highlighting a gap between random decoy variants and a correct variant peptide.

Score difference with wild type peptide

For the correct variant peptide matches, the corresponding score for the wild type (WT) peptide should be lesser. If the score difference between variant peptide and WT peptide is high, higher the chances of variant match to be correct. If the variant peptide scores lesser or equal to the wild type peptide, the chances of the variant being false enhances.

Variant Ambiguity Score (VAS)

Based on the empirical MS descriptors combined with PSM validation, an empirical score called variant ambiguity score (VAS) was developed. Some of the features which need to be included in VAS calculation are used in our previously developed tool MassWiz (Yadav et al., 2011). It has a very effective scoring system and provides a stringent score for every PSM match calculating empirical scoring weight for different ion types. We built upon these strengths and developed VAS for variant quality assessment. The VAS combines the weightage of different features like ion abundances and intensities as implemented in normalized MassWiz Score (normalized for length and b/y continuity) and also uses factors such as search engine counts and PSM counts per peptide to calculate a score which can assess the quality of the variant peptides which are divided into three classes – confident, ambiguous and poor. For variant peptide spectrum match, VAS is defined as -

$$VAS = (nMW * \log(P + 1) * \log(SE + 1) + Delta\ Shuffled\ Variant\ Score + Delta\ Wildtype\ Score)/3 \quad \text{(eq 1)}$$

Where nMW is normalized MassWiz Score, P is PSMs count per peptide and SE is search engine counts.

The VAS is implemented in the tool called PgxSAVy, to rescore variants in an automated manner by taking the proteogenomics search result as input to calculate VAS scores. For benchmarking, VAS was compared with the search output of a simulated spectra dataset and on real experimental but manually annotated dataset.

2.2. VAS evaluation strategy

To evaluate the scoring method for variant quality assessment, the identified variant peptides from proteogenomic analysis of MS data were rescored and the results were re-evaluated. The method was benchmarked tested on public MS data along with a simulation data. Simulated data represents ideal PSM features with near perfect separation of good and bad hits which can contribute in segregation of good from bad variant PSMs. But, since the simulated data differs from real-life spectra in terms of noise, we also used publicly available mass spectrometry data to evaluate the scoring method. The databases searches were conducted in multi-stage manner using a modified version of EuGenoSuite (Kumar *et al.*, 2016a) which is a multi-algorithm proteogenomic search pipeline, and integrated ProteoStats to implement multiple false discovery rate (FDR) methods (Yadav *et al.*, 2013) for further evaluation. Multi stage searches help to control database size to relevant proteins of interest and somewhat reduce the chances of incorrectly assigned non-variant PSMs by confidently assigning the known proteins at the first stage search and removing those from competition for the variant search stage. The standard method of combined FDR (Jones *et al.*, 2009) was used to estimate the integrated FDR from multi search algorithms at 1% global FDR threshold (gFDR). In addition, a class-specific FDR approach (csFDR) was calculated for specific peptide classes i.e., known, novel and variant peptides after mapping the identified hits on the reference proteome.

Our variant scoring method (VAS) is based on the hypothesis that multiple features of mass spectra can help us define the quality of a peptide spectrum match. Although such information is available to search engines, either it is ignored for the sake of search speed, or most search engines do not fully capitalize on this information. VAS also uses some information not considered by search engines and thus this integrated feature set can allow us

to distinguish between good and bad variant hits. VAS uses MS features like peaks intensity, continuous b and y ions match, search engines confidence and PSMs count, the importance of which has been reported in many studies (Ivanov *et al.*, 2014; Salz *et al.*, 2021; Choong and Sung, 2022).

2.3. Database Search

Simulation data

To evaluate the accuracy of VAS, a simulated data was generated using MaSS-Simulator (Awan and Saeed, 2018). The spectra were simulated using various parameters like dissociation strategies, noise content, relative intensities of signal peaks, peptide coverage, peptide length and peptide modifications. The input provided to the simulator contained 5000 peptides with single amino acid variation and 5000 peptides with two amino acids variation along with 40,000 wildtype peptides containing no variation (Figure 2). These 10,000 single and double amino acid variant peptides were added to the search database and if found, were considered as true positives. The 40,000 spectra simulated from wildtype peptides were not added to search database and were therefore considered false positives. Additionally, 500,000 unrelated variant peptides were added to the search database which had no overlap with the variant peptides used in simulation. This custom search database is used to search the simulated spectra. The search parameters used were – 0.1 Da precursor mass tolerance, trypsin as cleavage enzyme, no missed cleavage and carbamidomethylation at cysteine as fixed modification. The search was conducted using multiple database search engines with EuGenoSuite proteogenomics suite on a computer with 64GB RAM using 24 processing cores (Kumar *et al.*, 2016b). The FDR cut-off was set at 1% using FDR score method (Jones *et al.*, 2009).

PXD004010 (Brain AD dataset)

A publicly available data (Wang *et al.*, 2015) from post mortem brain samples of Alzheimer disease patients was also used to test our method in real world scenario. The samples were processed for transcriptomics and deep proteomics experiment and the authors identified approximately 16000 genes and 10000 proteins in the study (Wang *et al.*, 2015). The authors observed around 500 peptides with variants, frameshift, alternative splicing and non-coding gene translation, using an iterative search process for the validation of search strategy used by the tool JUMPg (Li, Wang, J.-H. H. Cho, *et al.*, 2016). The dataset had 1.7 million raw MS/MS spectra, available at the PRIDE database (PXD004010). All 10 fractions of the dataset in raw format were downloaded and converted to de facto data standard- "MGF" format using MSConvert tool from ProteoWizard (Chambers *et al.*, 2012). The database search was conducted in an iterative stage wise manner as suggested in Li et al with a notable difference. The database was created by authors from a combination of UniProt and transcript study conducted for the same sample. However, we created a new custom database using the subset database approach (explained in section 2.4). The data was first searched in UniProt human database downloaded on (May 2022) containing 42168 protein sequences (proteins + isoforms). The parameters for this first search were: - 6ppm precursor mass tolerance, 0.5Da fragment ion mass tolerance, 2 missed cleavages and trypsin as the proteolytic enzyme, fixed modification of carbamidomethylation at cysteine and variable modification of methionine oxidation were used. The search was conducted on multiple database search engines using EuGenoSuite proteogenomics suite on a computer with 64GB RAM using 24 processing cores (Kumar *et al.*, 2016b). The FDR cut-off was set at 1%. The second stage search was performed on the GENCODE database (version 40 with 246,624 transcripts) enabling the identification of novel peptides, junction peptides, alternative splicing and frameshifts (Frankish *et al.*, 2021). The unassigned spectra from the first stage search were used as input spectra for the second stage search. The third stage search was conducted on the unassigned

spectra from previous search, again the known variants from neXtProt database, for subset of proteins identified in previous two searches (Zahn-Zabal *et al.*, 2020) but no missed cleavages were allowed.

PXD001468 (HEK293 Cell line Dataset)

Another dataset used in the study was downloaded from Chick et al [PXD001468] (Chick *et al.*, 2015). The data was generated from HEK293 human embryonic kidney cell lines for demonstrating the feasibility for ultra-mass tolerant search to discover unanticipated modifications using Sequest. Chick et al analysed this data using open modification search through Sequest search engine by increasing the mass tolerance to ±500Da to identify modified spectra either due to PTMs or sequence variants (Chick *et al.*, 2015). This data was later used by Choong and Sung *et al* to demonstrate the guidelines and filtering criteria for selecting high confidence variant identification using proteogenomics search (Choong and Sung, 2022). This dataset contained 1.1 million MS/MS spectra acquired on Q-Exactive orbitrap Mass spectrometer. The data was converted to MGF format with the help of MSConvert tool. The data was also searched in multi stage manner using EuGenoSuite. The parameters for the search: - 5ppm parent mass tolerance, 0.01Da fragment mass tolerance, trypsin as the enzyme with one missed cleavage allowed. Carbamidomethylation at cysteine and oxidation at methionine were used as fixed and variable modifications respectively. The FDR cut-off was set to 1%. First stage search was conducted on UniProt human database, second stage used GENCODE and third stage was conducted on the known variants for subset of identified proteins in previous stage of neXtProt search. The spectra that remained unassigned at each stage were used as input for subsequent search.

2.4. Database creation and search

All the datasets in this study were searched using a three-stage search. For the first stage, we used protein database from SwissProt database from UniProt. The database also contained isoform sequences if present from any proteins. Since all studies were conducted on human dataset, we downloaded human SwissProt database containing 42168 protein sequences. This helps in identification of unmodified and unaltered protein sequences in the datasets. For the second stage, we downloaded the human GENCODE database (version 40) containing 246,624 transcripts and translated it in three frames for covering all possible products in different frames using an in-house Perl Script. This database search helps in identification of novel peptides, alternative splicing, variants peptides, junction peptides and frameshift peptides if present in the data. The third stage database is created by downloading human protein and variant sequences from neXtProt database. In the third stage, instead of searching peptides in the extremely large variant databases, a variant database subset is created on the fly for use. The subset is created based on variants from the previously identified proteins (in first two stages) to control the database size inflation. It is assumed that chances of occurrence and discovery of variant peptides might be boasted if search is limited to variants of already discovered proteins. To create the subset database, this custom database is simply filtered based on the protein identifications from the first and the second stage search for each dataset. Third stage database search can help in identification of known variant peptides, thus differentiating between the known and novel identifications. Corollary to global PTM discovery approach, this can be considered a global variant discovery approach (gVAR).

2.5. FDR calculation / estimation

Proteogenomics searches are often performed using a large search database and single or multiple search engines. One of the most challenging issue in variant proteogenomics is the high probability of incorrect assignment which requires statistical assessment for filtering the

false hits. Target decoy (TD) approach is a common method to estimate false positives and set thresholds to control false positives based on ratio of decoys to target PSMs in search results (Aggarwal and Yadav, 2016). But, due to large size of search database, the standard FDR estimation rate of 1% typically used in shotgun proteomics does not hold true (Aggarwal *et al.*, 2022) and thus the proportion of false positives is higher than estimated. To curtail this effect, the use of multiple search engines with combined FDR strategy as suggested by Jones et al (Jones *et al.*, 2009) was implemented, in which the FDR was first calculated individually for each search engine, and then a final combined FDR score was calculated as described previously (Kumar *et al.*, 2016c). However, Sticker et al suggested that the presence of novel and variant peptides requires stronger evidence than known peptides (Sticker, Martens and Clement, 2017). Therefore, a class specific FDR was implemented for sub classes of identified peptides after combining the results from all the searches into three categories – known, novel and variant peptides. The three categories were created based on fuzzy mapping the identified peptides on reference proteome using an in-house Perl script and labelling each peptide in one of these categories. The program allows a fuzzy mapping with two amino acids as variant, and higher mismatches are labelled into novel category. Once all the identified peptides are categorised in these three categories, class specific FDR is estimated in combined target-decoy approach similar to applied standard FDR method. We have estimated the class specific FDR based on target-decoy method using following equation:

$$FDR = \frac{2 * number\ of\ decoys\ PSMs}{number\ of\ target\ PSMs + number\ of\ decoy\ PSMs}$$

Based on the estimated FDR, 1% threshold was used and peptide hits passing the threshold are selected. The variant ambiguity score calculation and evaluation is performed on these FDR passed hits in variant category.

2.6. Manual annotation of spectra one fraction of Brain AD dataset for evaluation

We performed rigorous manual annotation of spectra identified in the variant proteogenomic search conducted in section 2.3 for Brain AD dataset to evaluate the performance of VAS. We selected the first fraction of MS data from AD dataset i.e., AD_PL01 and evaluated PSMs which passed the 1% FDR threshold. The PSM visualizations for the spectra were generated through Plabel (Wang *et al.*, 2007) using batch mode for the variant peptides and assigned a manual annotation of "good", "ambiguous" and "poor" hits. This manual annotation was based on the numbers of b and y ion matches and their intensity peak, noise level of the spectra, length of peptide and modifications present. A total of 875 PSMs were analysed and categorised in these three categories.

2.7. Evaluating VAS with Percolator

We also compared results of VAS with machine learning (ML) based tool Percolator to evaluate how they are performing with same MS/MS data. We used Percolator (Version 3.6) to re-score the 1% FDR pass PSMs hits from AD_PL01 brain AD dataset, same dataset for which were manually annotated. For negative training set, for each target peptide, a shuffled variant decoy peptide was generated which score best among all shuffled amino acid positions. All the quality and variant MS features used for estimating VAS was given as ML features to percolator with default parameters.

2.8. Variant annotation

An important step after identification of variants is to map those to phenotype or disease from dbSNP or other curated SNP databases. Most of these databases are integrated into neXtProt. All the variant peptides analysed by PgxSAVy, were also annotated based on biological classification provided in neXtProt. A module BioAnnotation was developed in PgxSAVy to annotate variants using neXtProt XML file that contains information on isoforms, disease relation, corresponding wildtype amino acids and mutated amino acid. The current integrated

database for annotation contains 56,65,956 variant amino acid information with their role in disease context. These are mapped from the XML file into variant PSM results to provide biological meaning and context to the variant peptides discovered in proteogenomics searches.

3. Results

Protein sequence variations caused by single nucleotide polymorphisms (SNPs) are important for understanding the different phenotypic behavior of proteins. Mass spectrometry (MS) based proteomics provides a high throughput way to detect single amino acid variations (SAVs). But, detection of variant peptides is prone to false identification for which specific measures are required and often further validation is necessary to confirm their existence. Class specific FDR and probability methods help to filter out false positives during FDR estimation or posteriorly based on different MS features or distribution. However, these methods still have limitations to completely filter out false positives. Various rescoring tools have been developed like Percolator (Käll et al., 2007), MS^2 Rescore (Declercq et al., 2022), INFERYS (Zolg et al., 2021) which provide a rescore on the top results given by search algorithms to discriminate between correct and false hits. But these tools mostly focus on inclusion of peptides which may be falling below FDR threshold to maximize the results. Our goal is to scrutinize the identified variant peptides which have passed the FDR threshold and provide a confidence score. We have developed a method for controlling the false positives based on analyzing various features of mass spectrometry spectra in variant peptides. We have benchmarked our method on three different datasets namely, Alzheimer dataset, HEK293 cell line dataset and simulated dataset as described in method section 2.3. The workflow for the study is presented below.

3.1. Workflow of the analysis

The variant proteogenomic search was carried out using the workflow shown in the figure 1. Three MS datasets as mentioned earlier were used to validate the VAS scoring. The MS data were searched against three databases namely, ReferenceDB, TranscriptDB and VariantDB. ReferenceDB was built from 42,362 canonical proteins with their isoforms downloaded from the UniProt database. TranscriptDB was built from the GENCODE database using 246,624 transcripts and contains 1,676,122 translated three frame ORFs. VariantDB was built from neXtProt variant data containing 93,669,236 variant peptides generated from mapping of variants on 42,329 proteins. For searching, EugEnoSuite pipeline was used which has three search engines OMSSA, X!Tandem and MSGF-Plus. The search was conducted in a multi stage manner where the first search was conducted on ReferenceDB and all scans matching known proteins were assigned. The second stage was conducted with unassigned spectra left after the first search against TranscriptDB which helped to identify those peptides which may be novel in nature. The unassigned spectra left after second search may contain variant peptides which can be searched against VariantDB. A subset of VariantDB was created to capture only those variant peptides for which proteins have been identified in previous stage searches, so as to reduce the chances of false positives by controlling the size of search space. This reduces the search space and increases the chances of identifying correct variant peptides.

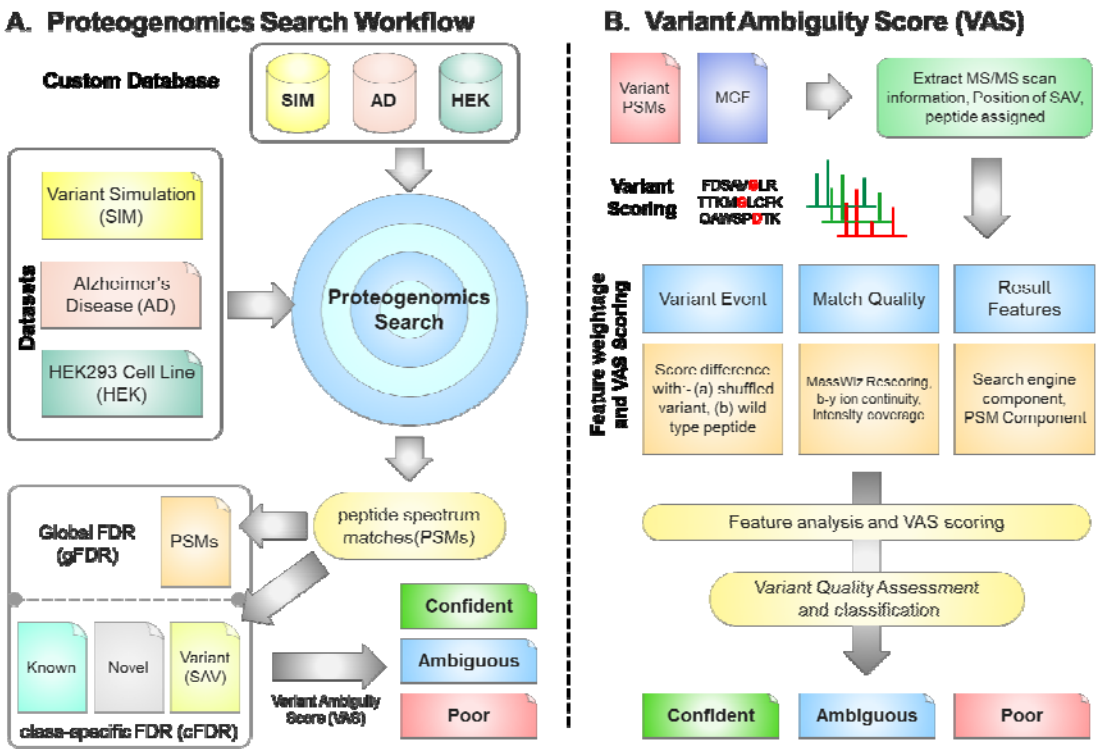


Figure 1: Workflow for identification and post search validation of variant peptides from proteogenomics searches.

For all the three searches, a global FDR was estimated using the combined FDR approach as implemented in EuGenoSuite. A separate class-specific FDR estimation was performed after combining and segregating the results of all three search engines into three different classes i.e. known, novel and variants. A classification step as discussed in section 2.5 was performed for second stage search to classify identified peptides in three classes as these results are based on translated transcripts which may contain known, novel and variant peptides. The class specific FDR is also subjected to 1% FDR. VAS was calculated for all datasets as described in section 2.1.

3.2. Evaluation of VAS on simulated data

Simulated data for 10,000 variant peptides containing one or two variants per peptide sequence was generated using MassSimulator (Table 1). Along with the variant data, 40,000 non related wild type peptides were also used for generating simulated spectra at 10% noise level. While searching the data, a database was created for 10,000 true variant peptides and 5,00,000 false variant hits. No wild type peptides present in spectral data were provided in the search database. This enabled us to identify false positives classified as identified in table 1 and strategy explained in figure 2.

Table 1 Composition of Simulated MS Input and search database

Spectra	Database	Reason
Spectra with 1 variant: 5000	5000 Variant peptides (1 variant)	True Positives
Spectra with 2 variants: 5000	5000 Variant peptides (2 variant)	True Positives
Spectra without any variant (Wildtype): 40,000	Sequence absent corresponding to these 40,000 peptides	False Positives (To enhance the competition and make WT spectra hit on variant)
	5,00,000 Non related variant peptides which are absent in simulated spectra	False Positives

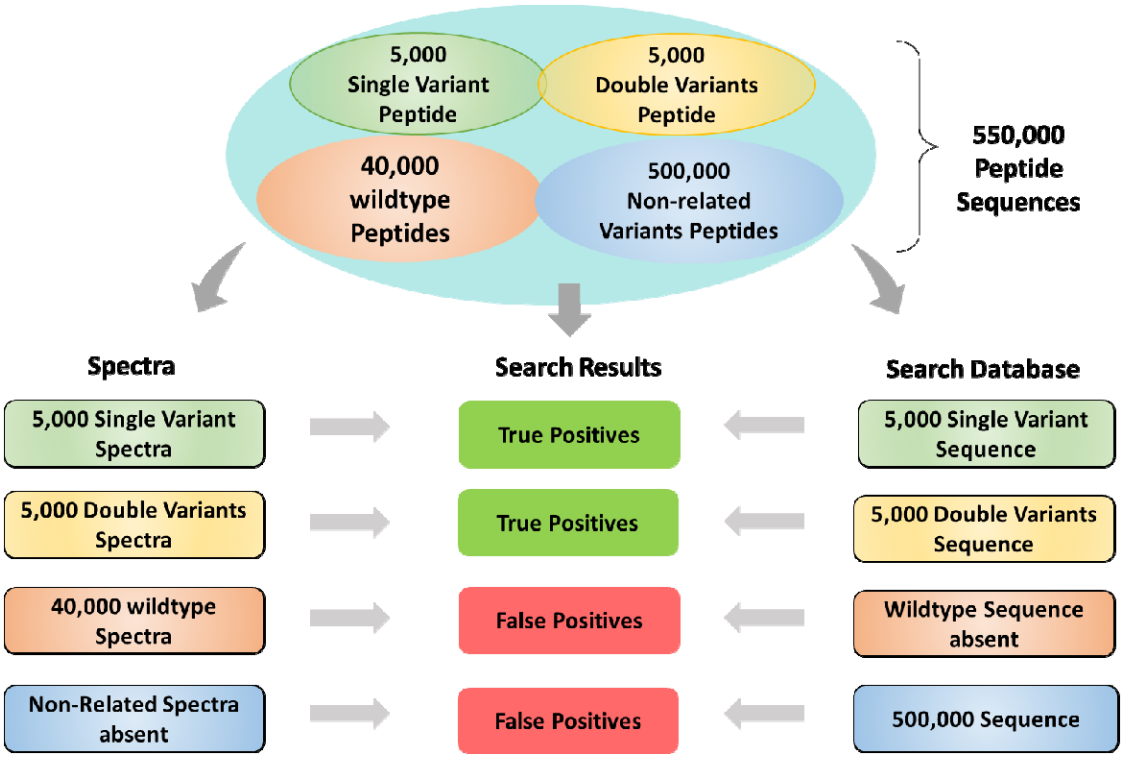


Figure 2: Workflow for generation and analysis of simulation data to evaluate the Variant Ambiguity Score (VAS)

Based on the identifications, the various MS features calculated on 1% global FDR pass data were divided in two categories- true positive and false positive. The features were assessed on their ability to separate the two groups as shown in figure 3. The search conducted on simulated MS data leads to identification of a total 49,931 non redundant spectra out of 50,000 spectra given as input. The total unique PSM identified for these spectra was 100,219 leading to an average of two PSMs per spectra. More than 24,186 PSMs were identified by two or more search engines. Also, out of 100,219 PSMs, 82983 PSMs mapped to target database and 17236 PSMs matched to decoy database; 763 PSMs mapped to contaminants. From total identified PSMs, 11,274 PSMs passed the 1% FDR threshold consisting of 11219 Target mapped PSMs and 55 Decoy PSMs. After removing the contaminants and decoy hits, we have a total 11,094 identified PSMs belonging to 11,074 unique spectra. The true

positives and false positives identified in the 11,074 PSMs were 9,593 and 1481, respectively. We analyzed these 1481 false positives and found 995 falls in the range of MS1 tolerance i.e., below 0.1 Da taken in this case and the rest 486 hits are those for which true peptide sequence was not provided in the search database. We also observed out of 1481 false positives, many of which mapped to isobaric peptides. Only 15 hits are those spectra for which the correct match sequence was provided in the search database but they failed to map them. In these 15 incorrect search hits, 7 were within MS1 tolerance range and the rest match due to isotopic search conditions. We performed MS features analysis on these identified total 11,074 PSMs to analyze the distribution and segregation of true and false positives.

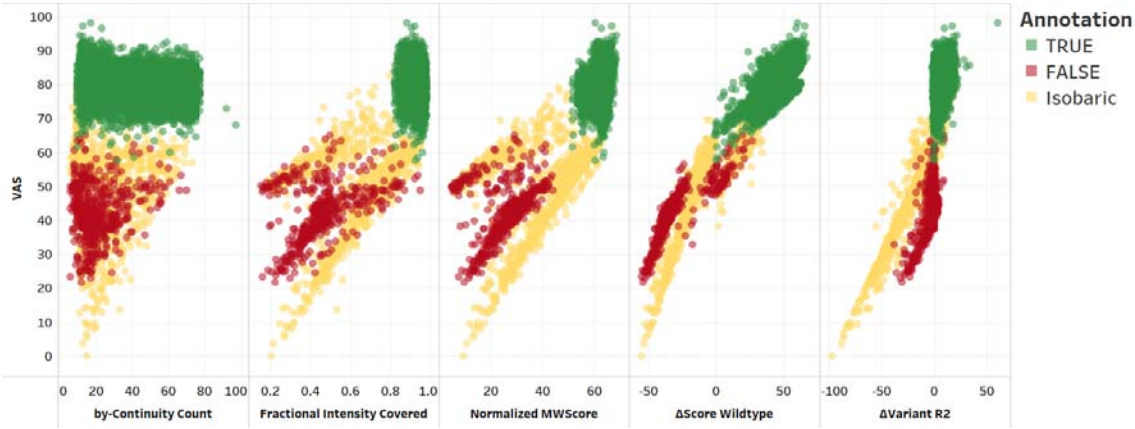


Figure 3: Distribution of features for identified PSMs from simulated data, showing segregation of true and false hits against VAS scores.

We observed that there are MS features like by continuity score, fractional intensity coverage and nMassWiz (nMW) perform very well and are able to segregate between true and false positives. While some features were expected to perform well and not show any clear correlation, the reason behind this behavior can be due to limited competition and less noise in the data.

When we analyzed the VAS scores against nMW scores we observed a lot of false positives in higher score sections. This was because of the isobaric nature of substitution or two substitutions creating a similar peptide as the wild type.

These two peptides have two amino acid changes at 10 (D->E) and 12 (I->V) positions. The mass difference between these two amino acids is 0 and being in close proximity to each other i.e., making them isobaric can lead to identification of incorrect matches. Since this simulation study was an ideal case scenario as the data present here behaves more ideal as compared to the sample-based data generated from mass spectrometry. For this purpose, we subsequently tested our hypothesis on a set of identified PSMs at 1% FDR after careful manual validation on the quality of identification.

3.3. Evaluation on the manually validated dataset

A single fraction of the Brain AD dataset was selected for manual annotation as described in section 2.8. The fraction contains 0.17 million spectra which were searched in three stage searches against ReferenceDB, TranscriptDB and VariantDB as discussed in section 2.2 and 2.3. The search conducted on the MS data leads to identification of a total 64,191; 51 and 893 PSMs in first, second and third stage search, respectively at 1% FDR estimated separately at each stage. The total unique spectra identified in all three stages combined was 64,991 representing most spectra identified as single PSMs. However, more than 60,309 PSMs were identified by two or more search engines. From a total 65,137 identified PSMs passing above the 1% FDR threshold consist of 64,817 Target mapped PSMs and 320 Decoy PSMs. After removing the contaminants and decoy hits, a total of 62,907 PSMs were identified. As our study focuses on variant peptides, we selected the 887 variant PSMs above 1% FDR in variant class FDR. If two search algorithms gave two different peptides for the same spectra, higher scored PSMs were retained. This led to filter out 12 PSMs. From this, 875 unique

PSMs were manually annotated in three categories named ‘Good’, ‘Average’ and ‘Bad’ based on visualization of PSMs in spectra visualization pLabel tool as discussed in section 2.6. The factors considered during annotation of PSMs were numbers of b and y ion matches and their intensity peak, noise level of the spectra, length of peptide and modifications present. Through manual validation, 206, 397 and 272 PSMs were categorized in the Good, Average and Bad category, respectively. This manual annotation demonstrates that even after 1% FDR estimation in variant proteogenomics using class specific FDR approach, half of hits are likely to be false. This estimation may vary in different proteogenomic pipelines but still there will be a good number of false positives present which surely require a rescoring method to filter out. Once manually annotated, VAS was calculated on these PSMs to evaluate our method and find its precision. VAS using the empirical formula as described in section 2.1, categorizes these PSM in three groups: confident, ambiguous and poor.

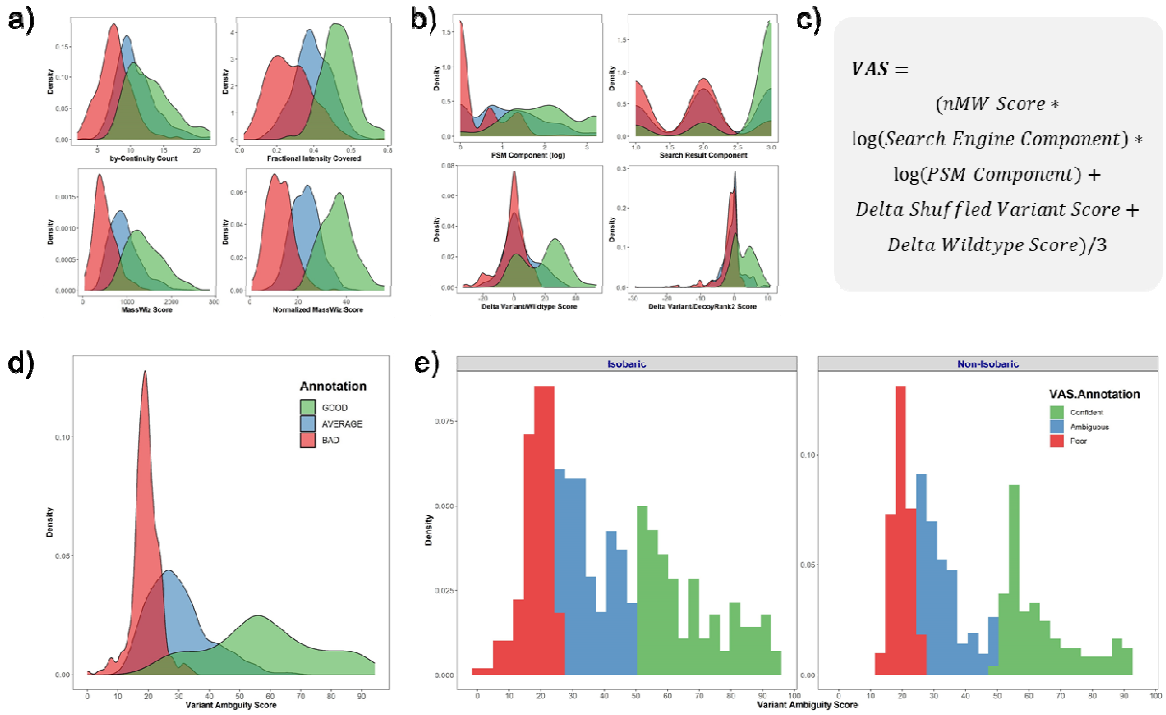


Figure 4: Distribution of features for manually annotated spectra to evaluate VAS scoring. Distribution of features for – (a) match quality, and (b) variant events. Manually annotated spectra good, average and bad PSMs are shown in green, blue and red, respectively. (c) Composition of VAS formula to combine the various features (d) Density distribution of VAS annotated spectra showing a clear segregation between good and bad hits (e) VAS Classification validation with manually annotated search results and isobaric filter.

Out of 875 PSMs, there were 161 confident PSMs, 310 semi-confident and 404 ambiguous PSMs in VAS estimation. When analyzed for the overlap of this categorization of PSMs with our manually annotated results, there is around 80% overlap between the results. As VAS uses multiple MS features to distinguish between true and false positive hits, we also observed the density distribution of manually annotated PSMs and VAS annotated PSMs for each of the features (Figure 4 a & b). These features include b-y ions continuity count, fractional intensity ratio, MassWiz Scores and Normalized MassWiz Score for match quality features. PSM counts per peptide and search engine identification counts with delta variant/wildtype score and delta variant/decoy rank2 score were included as variant event features in final scoring. We also analyzed that out of 875 identified PSMs, how many variant PSMs fall in the category of isobaric nature. We selected 11 such conditions (Supplementary Table 1) where amino acid substitution with any other variant amino acid might fall in the condition of isobaric (Choong et al., 2017). Based on the list, those variant peptides which fall in isobaric, even with good VAS scores are considered ambiguous in nature. We separated isobaric and non-isobaric variant peptides and analyzed the distribution of VAS classified variant peptides (Figure 4 e).

We also evaluated VAS results in comparison with percolator to check if percolator can classify correct variant peptides. We rescored 1% FDR passed PSMs hits of AD_PL01

dataset with percolator and 321 PSMs passed the q-value cut-off of 0.01. These 321 PSMs include manually annotated 45 good, 151 average and 125 bad PSMs (supplementary figure 1). Only one-third of PSMs passing the percolator threshold matched with VAS (Confident and Ambiguous) while rest other matched totally with VAS (Poor) (supplementary figure 1). This suggests that although percolator is an effective tool for rescoring but it requires large training data, not generally available in FDR controlled variants and also that there is not sufficient large and clearly demarcated negative training dataset in variant peptides where PSMs are passing FDR threshold.

3.4. Global vs class-specific FDR

Global FDR (gFDR) estimates the proportion of false positive identifications across all classes or categories of peptides or proteins identified in the single search. But, as the search space is very large in proteogenomics search, the 1% threshold for FDR cut off still allows many more false positives. To deal with this, class-specific FDR (csFDR) has been suggested which measures the proportion of false positive identifications within a specific class or category. This means that the FDR is calculated separately for each class or category of peptides or proteins, such as those resulting from different known and novel identifications. To test this FDR estimation, we used the first fraction of the brain AD dataset which we used for manual annotation. For gFDR estimation, we combined all search results from different stages of searching ReferenceDB, TranscriptDB and VariantDB and estimated FDR using the Combined FDR estimation method. This is different from previous FDR estimation which was done at each stage. Out of total 856,131 search results, 61,487 PSMs pass the 1% FDR cut off and then classified in three categories based on our classification tools in 60,706 known, 53 Novel and 728 Variant PSMs. For csFDR estimation, we initially created the three classes of known, novel and variant of search hits and then estimated FDR. 60,672 known

PSMs, 41 novel PSMs and 778 variant PSMs pass the 1% FDR cut off in this class specific FDR estimation (Figure 5). We observed that the csFDR behaves little stringent in filtering false positives when divided in classes initially. When we mapped these classes with the annotation, we observed that the percentage of good hits passing the threshold in csFDR are higher but also bad PSM hits are able to pass the threshold. This showed csFDR perform better than gFDR but still require VAS rescoring to remove false positives. For complete datasets, the third stage search containing only variants act as class-specific FDRsearch.

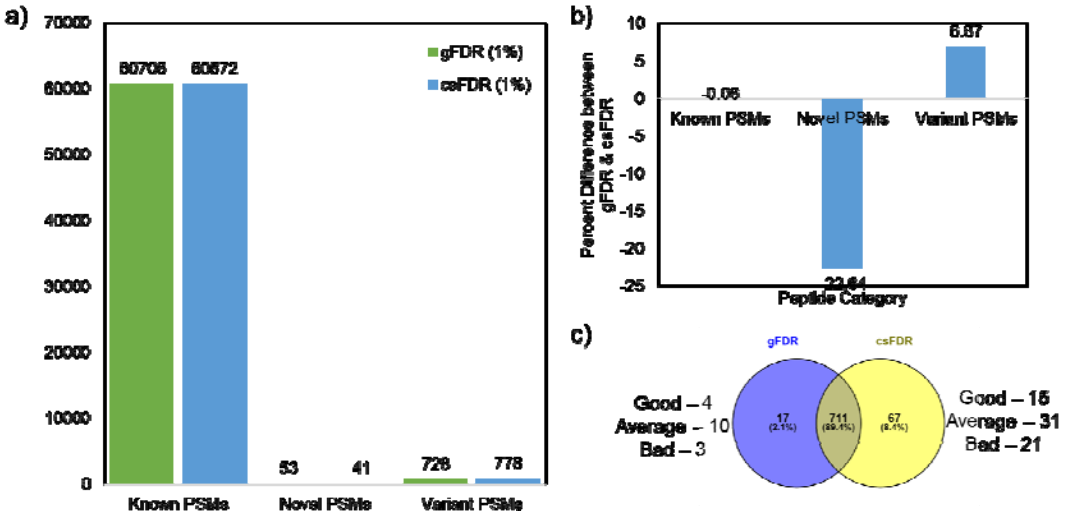


Figure 5: Comparison between gFDR and csFDR with respect to classification of variant peptides in known, novel and variant a) Number of identifications of non-redundant PSMs as Known, Novel and Variant PSMs b) Percentage difference of identification between gFDR and csFDR

3.5. Brain AD Dataset and HEK Cell line data variant peptides identification

The complete Brain AD dataset, containing 1,718,768 spectra, was searched using the EuGenoSuite algorithm in a multi-stage approach. This dataset consisted of 10 fractions, each representing a distinct sample of brain tissue in the Alzheimer's disease condition. Following

the methodology described in sections 2.2 and 2.3, three stage searches were conducted against the ReferenceDB, TranscriptDB, and VariantDB. In the analysis of the MS data, a total of 631503, 2362, and 9897 peptide-spectrum matches (PSMs) were identified in the first, second, and third stage searches, respectively, with a 1% false discovery rate (FDR) estimation for each stage independently. After removing contaminants and decoy hits, the final counts of identified PSMs were reduced to 614,695, 2,357, and 9,858 for the first, second, and third stages, respectively.

Given the specific focus of our study on variant peptides, we selected the 9,858 variant PSMs that exceeded the 1% FDR threshold within the variant class FDR for further analysis using the Variant Annotation Scoring (VAS) method. VAS, employing the empirical formula described in section 2.1, classified these PSMs into three distinct groups: confident, ambiguous, and poor. The classification of variant peptides into these classes is presented in table 2. Notably, nearly 50% of the identified variant peptides in the variant proteogenomics analysis had a high likelihood of being false hits. Approximately 12% of the variant peptides received a confident label, indicating the stringent nature of PgxSAVy and its ability to identify the most reliable variant peptides. This provides researchers with the confidence to perform downstream analyses on these variant peptides. Furthermore, figure 6 displays the matched spectra of some identified variant peptides from different classes, showcasing their diversity and characteristics.

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

VAS	**AD dataset**	**HEK Cell line**
Total	9853	4011

Confident	1171	35
Ambiguous	3726	472
Poor	4696	3504

Similarly, the HEK cell line data was subjected to analysis using the EuGenoSuite pipeline and the PgxSAVy rescoring approach to identify variant peptides. The HEK cell line dataset consisted of approximately 1.12 million spectra, which were searched against the ReferenceDB, TranscriptDB, and VariantDB databases. Following the same methodology as the Brain AD dataset, a three-stage search was performed on the MS data.

In the first stage search, a total of 487,697 peptide-spectrum matches (PSMs) were identified, while the second and third stage searches yielded 1,071 and 4,011 PSMs, respectively. These identifications were made at a 1% false discovery rate (FDR) estimated separately for each stage. The PgxSAVy rescoring approach allowed for the successful detection of variant peptides in the HEK cell line data, enabling the investigation of their role in cellular processes and disease mechanisms.

By employing this comprehensive search strategy and rigorous FDR estimation, the analysis of the HEK cell line data revealed valuable insights into the presence and abundance of variant peptides. This information contributes to a better understanding of the molecular landscape of the HEK cell line and provides a foundation for further investigations into its functional implications.

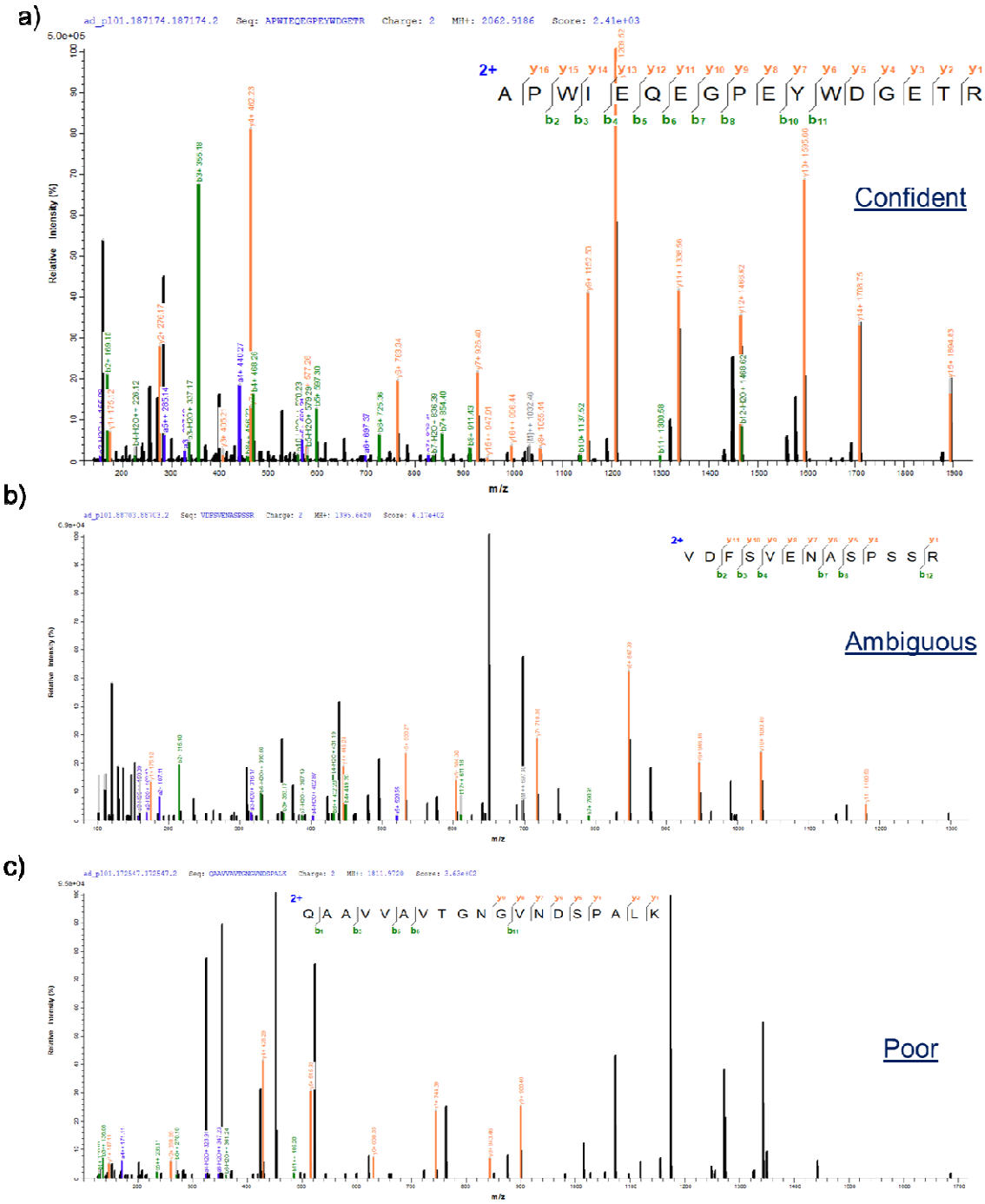


Figure 6 Spectra classified with VAS scoring in a) confident b) ambiguous c) poor

3.6. Annotation of variants peptides in PgxSAVy

While the proteogenomic identification of SAVs is tedious and computationally challenging process, the variant identification and validation using post-processing tool like PgxSAVy's VAS score is not enough. TO understand the functional implication of these variants, one must also get information about their functional biological or disease impact. Many resources for genomic variant assessment are available and nextprot is a rich source of such compiled information at one portal, a manual laborious process follows. To facilitate this tedious step, PgxSAVy contains an annotation module that can help researchers take their variant findings further.

After the segregation of variant peptides, the identified peptides in the results are mapped against their annotated counterparts in the nextProt database for providing further biological meaning to the results gained. This is provided for biologists to make informed decisions about the next set of validation experiments to understand the biological implications of the findings. As an example, figure 7 shows the frequency graph of wild type amino acids getting mutated to other amino acids in one of the brain AD datasets. This can inform the researchers on the dominant mutation type in a disease context.

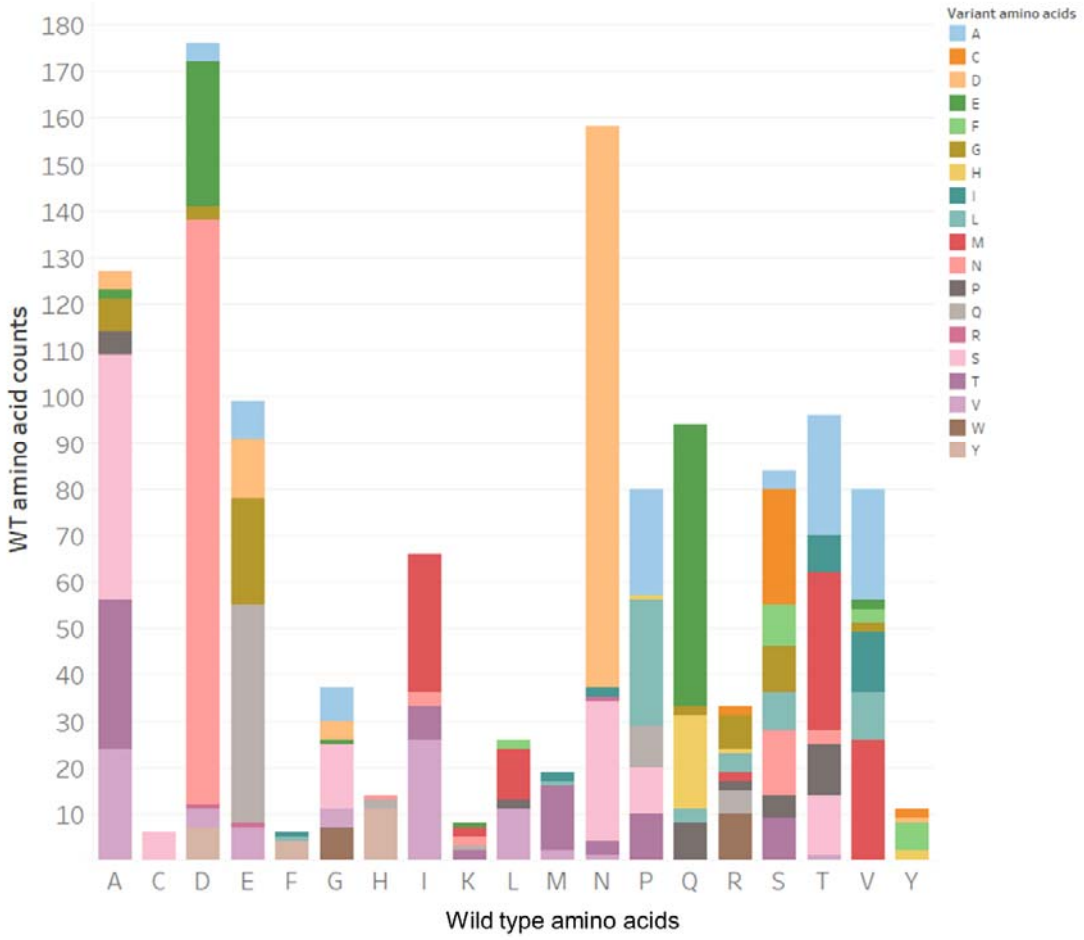


Figure 7: Frequency of mutational changes mapped by PgxSAVy

Another example shown in figure 8 depicts the type of variant amino acids distributed for each category disease information. This can help researchers evaluate the potential functional and disease impact of the variant events with respect to their frequencies.

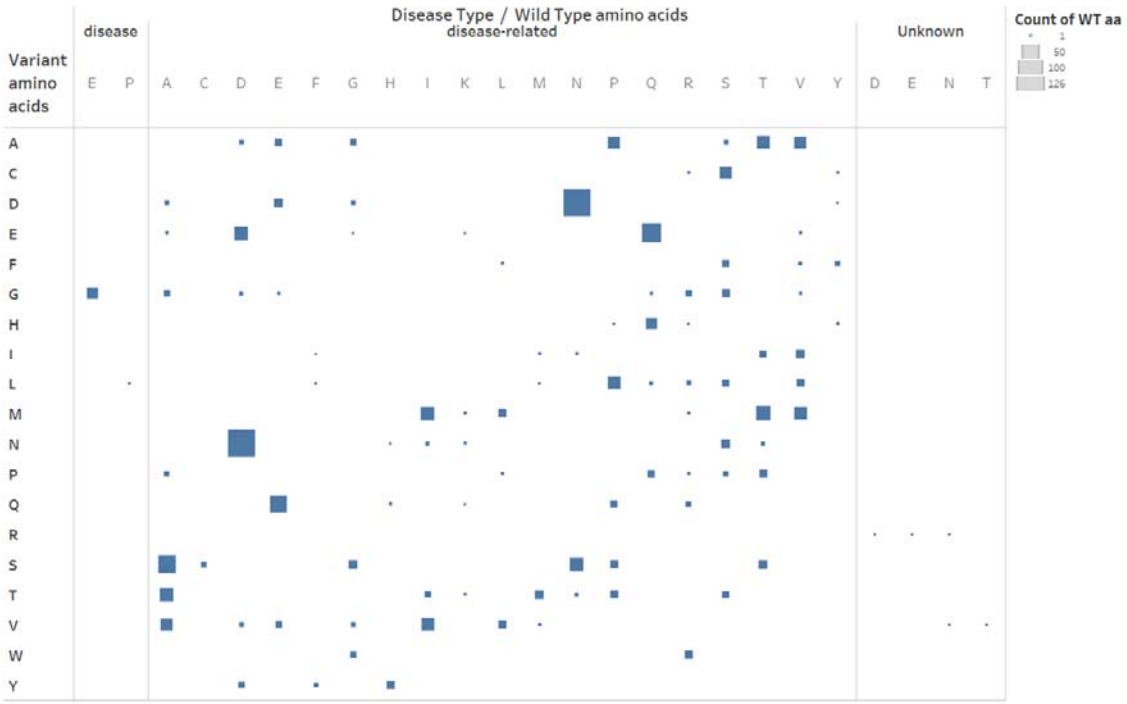


Figure 8: Disease annotation and frequency of SAAVs.

4. Discussion and Conclusion

The role of peptide and protein sequence variations is crucial in understanding the functions of proteins under various biological conditions. However, detecting variant peptides presents challenges in controlling false identifications. Existing methods, such as class-specific false discovery rate (FDR) estimation and probability-based filters, have their limitations. To address this, a new approach is needed to analyze mass spectrometry spectra features of variant peptides and effectively control false positives. With this goal, we have developed a workflow that involves a variant proteogenomic search using three databases: ReferenceDB, TranscriptDB, and VariantDB. This search is conducted in multiple stages to identify known, novel, and variant peptides. FDR estimation is performed using a combined FDR approach and class-specific FDR estimation tailored to different peptide classes. To evaluate this

method, we utilized three datasets: Alzheimer dataset, HEK293 cell line dataset, and a simulated dataset.

In the evaluation of our method, we generated simulated data containing variant peptides and wild-type peptides that are not related. By analyzing non-redundant spectra, we were able to distinguish true positives from false positives. Notably, certain MS features such as normalized b-y continuity score, average variant peak position score, and nMassWiz exhibited strong performance in differentiating true and false positives. Additionally, we discussed the challenges posed by isobaric peptides and their impact on identification accuracy. We further evaluated VAS using a manually validated dataset from a specific fraction of the Brain AD dataset. Through manual annotation of identified PSMs, we categorized them as 'Good,' 'Average,' or 'Bad' based on multiple factors. VAS scores were calculated for these PSMs, and we observed a significant overlap between manual annotation and VAS annotation, indicating the method's precision. We also examined the distribution of MS features for variant PSMs and observed distinct patterns among 'Good,' 'Average,' and 'Bad' PSMs.

Moreover, we discussed the comparison between global FDR and class-specific FDR estimation. Global FDR estimates the proportion of false positive identifications across all classes, while class-specific FDR measures false positives within specific classes. Our observations revealed that class-specific FDR was more stringent in filtering false positives, particularly in the 'Bad' category. Lastly, we applied our method to the complete Brain AD dataset and compared it with other existing tools. Through variant PSM identification and VAS scoring, we categorized the PSMs into confidence categories and discovered a high proportion of variants even after FDR are falling in the false positive category as discovered by VAS.

In summary, our study underscores the effectiveness of the VAS method in controlling false positives during the identification of variant peptides. We provide insights into its performance across different datasets and emphasize the significance of manual validation and simulated data. PgxSAVy is a handy tool that can perform post search variant quality control and signifies the importance of such tools by highlighting the need for removing such false variants that use up critical resources for chasing incorrect findings owing to proteogenomics database size issues. It can also help in biological evaluation of variants by proving facile integration with neXtProt knowledge for variant annotation and highlighting their role in disease, if any.

References:

Aggarwal, S. *et al.* (2021) 'Posttranslational modifications in systems biology', in Donev, R. and Karabencheva-Christova, T.B.T.-A. in P.C. and S.B. (eds) *Proteomics and Systems Biology*. Academic Press, pp. 93–126. doi:10.1016/bs.apcsb.2021.03.005.

Aggarwal, S. *et al.* (2022) 'False discovery rate: the Achilles' heel of proteogenomics', *Briefings in Bioinformatics*, pp. 1–15. doi:10.1093/bib/bbac163.

Aggarwal, S. and Yadav, A.K. (2016) 'False discovery rate estimation in proteomics', in *Methods in Molecular Biology*. Humana Press Inc., pp. 119–128. doi:10.1007/978-1-4939-3106-4_7.

Agosto, L.M. *et al.* (2019) 'Deep profiling and custom databases improve detection of proteoforms generated by alternative splicing', *Genome Research*, 29(12), pp. 2046–2055. doi:10.1101/gr.248435.119.

Anurag, M. *et al.* (2022) 'Proteogenomic Markers of Chemotherapy Resistance and Response in Triple-Negative Breast Cancer', *Cancer Discovery*, pp. OF1–OF20. doi:10.1158/2159-8290.CD-22-0200.

Awan, M.G. and Saeed, F. (2018) 'MaSS-Simulator: A Highly Configurable Simulator for Generating MS/MS Datasets for Benchmarking of Proteomics Algorithms', *Proteomics*, 18(20), pp. 1–4. doi:10.1002/pmic.201800206.

Bouwmeester, R. *et al.* (2021) 'DeepLC can predict retention times for peptides that carry as-yet unseen modifications', *Nature Methods*, 18(11), pp. 1363–1369. doi:10.1038/s41592-021-01301-5.

Cesnik, A.J. *et al.* (2020) 'Spritz: A Proteogenomic Database Engine', *Journal of Proteome Research*, p. acs.jproteome.0c00407. doi:10.1021/acs.jproteome.0c00407.

Chambers, M.C. *et al.* (2012) 'A cross-platform toolkit for mass spectrometry and proteomics', *Nature Biotechnology*, 30(10), pp. 918–920. doi:10.1038/nbt.2377.

Chick, J.M. *et al.* (2015) 'A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides', *Nature Biotechnology*, 33(7), pp. 743–749. doi:10.1038/nbt.3267.

Choong, W.K. and Sung, T.Y. (2022) 'Multiaspect Examinations of Possible Alternative Mappings of Identified Variant Peptides: A Case Study on the HEK293 Cell Line', *ACS Omega*, pp. 1–9. doi:10.1021/acsomega.2c00466.

Fermin, D. *et al.* (2013) 'LuciPHOr: Algorithm for phosphorylation site localization with false localization rate estimation using modified target-decoy approach', *Molecular and Cellular Proteomics*, 12(11), pp. 3409–3419. doi:10.1074/mcp.M113.028928.

Frankish, A. *et al.* (2021) 'GENCODE 2021', *Nucleic Acids Research*, 49(D1), pp. D916–D923. doi:10.1093/nar/gkaa1087.

Gonzalez-Teran, B. *et al.* (2022) 'Transcription factor protein interactomes reveal genetic determinants in heart disease', *Cell*, 185(5), pp. 794-814.e30. doi:10.1016/J.CELL.2022.01.021.

Hubler, S.L. *et al.* (2019) ‘Challenges in Peptide-Spectrum Matching□: a Robust and Reproducible Statistical Framework for Removing Low-Accuracy , High-Scoring Hits’, *bioRxiv*, (612), pp. 1–41. doi:10.1101/839290.

Ivanov, M. V. *et al.* (2014) ‘Empirical multidimensional space for scoring peptide spectrum matches in shotgun proteomics’, *Journal of Proteome Research*, 13(4), pp. 1911–1920. doi:10.1021/pr401026y.

Johnson, E.C.B. *et al.* (2022) ‘Large-scale deep multi-layer analysis of Alzheimer’s disease brain reveals strong proteomic disease-related changes not observed at the RNA level’, *Nature Neuroscience*, 25(2), pp. 213–225. doi:10.1038/s41593-021-00999-y.

Jones, A.R. *et al.* (2009) ‘Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines’, *Proteomics*, 9(5), pp. 1220–1229. doi:10.1002/pmic.200800473.

Kelkar, D.S. *et al.* (2011) ‘Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry’, *Molecular and Cellular Proteomics*, 10(12), pp. 1–13. doi:10.1074/mcp.M111.011627.

Krasnov, G.S. *et al.* (2015) ‘PPLine: An Automated Pipeline for SNP, SAP, and Splice Variant Detection in the Context of Proteogenomics’, *Journal of Proteome Research*, 14(9), pp. 3729–3737. doi:10.1021/acs.jproteome.5b00490.

Kumar, D. *et al.* (2013) ‘Proteogenomic analysis of bradyrhizobium japonicum USDA110 using genosuite, an automated multi-algorithmic pipeline’, *Molecular and Cellular Proteomics*, 12(11), pp. 3388–3397. doi:10.1074/mcp.M112.027169.

Kumar, D. *et al.* (2014) ‘Discovery of rare protein-coding genes in model methylotroph methylobacterium extorquens AM1’, *Proteomics*, 14(23–24), pp. 2790–2794. doi:10.1002/PMIC.201400153.

Kumar, D. *et al.* (2016a) ‘Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation’, *Molecular and Cellular Proteomics*, 15(1), pp. 329–339. doi:10.1074/mcp.M114.047126.

Kumar, D. *et al.* (2016b) ‘Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation’, *Molecular and Cellular Proteomics*, 15(1), pp. 329–339. doi:10.1074/mcp.M114.047126.

Kumar, D. *et al.* (2016c) ‘Integrated transcriptomic-proteomic analysis using a proteogenomic workflow refines rat genome annotation’, *Molecular and Cellular Proteomics*, 15(1), pp. 329–339. doi:10.1074/mcp.M114.047126.

Kumar, D., Yadav, A.K. and Dash, D. (2017) ‘Choosing an optimal database for protein identification from tandem mass spectrometry data’, in Keerthikumar, S. and Mathivanan, S. (eds) *Methods in Molecular Biology*. Humana Press Inc., pp. 17–29. doi:10.1007/978-1-4939-6740-7_3.

Lau, E. *et al.* (2019) ‘Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome’, *Cell Reports*, 29(11), pp. 3751-3765.e5. doi:10.1016/j.celrep.2019.11.026.

Li, Y., Wang, X., Cho, J.-H.H., *et al.* (2016) ‘JUMPg: An integrative proteogenomics pipeline identifying unannotated proteins in human brain and cancer cells’, *Journal of Proteome Research*, 15(7), pp. 2309–2320. doi:10.1021/acs.jproteome.6b00344.

Li, Y., Wang, X., Cho, J.H., *et al.* (2016) 'JUMPg: An integrative proteogenomics pipeline identifying unannotated proteins in human brain and cancer cells', *Journal of Proteome Research*, 15(7), pp. 2309–2320. doi:10.1021/acs.jproteome.6b00344.

Menschaert, G. and Fenyö, D. (2017) 'Proteogenomics from a bioinformatics angle: A growing field', *Mass Spectrometry Reviews*, 36(5), pp. 584–599. doi:10.1002/mas.21483.

Mertins, P. *et al.* (2016) 'Proteogenomics connects somatic mutations to signalling in breast cancer', *Nature*, 534(7605), pp. 55–62. doi:10.1038/nature18003.

Noble, W.S. (2015) 'Mass spectrometrists should search only for peptides they care about', *Nature Methods*, 12(7), pp. 605–608. doi:10.1038/nmeth.3450.

Omenn, G.S. *et al.* (2022) 'The 2022 Report on the Human Proteome from the HUPO Human Proteome Project', *Journal of Proteome Research* [Preprint]. doi:10.1021/acs.jproteome.2c00498.

Patrie, S.M. (2016) *Modern Proteomics – Sample Preparation, Analysis and Practical Applications*. Edited by H. Mirzaei and M. Carrasco. Cham: Springer International Publishing (Advances in Experimental Medicine and Biology). doi:10.1007/978-3-319-41448-5.

Salz, R. *et al.* (2021) 'Personalized Proteome: Comparing Proteogenomics and Open Variant Search Approaches for Single Amino Acid Variant Detection', *Journal of Proteome Research*, 20(6), pp. 3353–3364. doi:10.1021/acs.jproteome.1c00264.

Sticker, A., Martens, L. and Clement, L. (2017) 'Mass spectrometrists should search for all peptides, but assess only the ones they care about', *Nature Methods*. Nature Publishing Group, pp. 643–644. doi:10.1038/nmeth.4338.

Wang, H. *et al.* (2015) 'Systematic optimization of long gradient chromatography mass spectrometry for deep analysis of brain proteome', *Journal of Proteome Research*, 14(2), pp. 829–838. doi:10.1021/pr500882h.

Wang, L. *et al.* (2007) 'pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry', *Rapid Communications in Mass Spectrometry*, 21(18), pp. 2985–2991. doi:https://doi.org/10.1002/rcm.3173.

Wang, X., Zhang, B. and Wren, J. (2013) 'customProDB□: an R package to generate customized protein databases from RNA-Seq data for proteomics search', *Bioinformatics*, 29(24), pp. 3235–3237. doi:10.1093/bioinformatics/btt543.

Wen, B. *et al.* (2020) 'Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis', *Nature Communications*, 11(1), pp. 1–14. doi:10.1038/s41467-020-15456-w.

Woo, S., Cha, S.W., Merrihew, G., *et al.* (2014) 'Proteogenomic database construction driven from large scale RNA-seq data', *Journal of Proteome Research*, 13(1), pp. 21–28. doi:10.1021/pr400294c.

Woo, S., Cha, S.W., Na, S., *et al.* (2014) 'Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data', *Proteomics*, 14(23–24), pp. 2719–2730. doi:10.1002/pmic.201400206.

Woo, S. *et al.* (2015) 'Advanced proteogenomic analysis reveals multiple peptide mutations and complex immunoglobulin peptides in colon cancer', *Journal of Proteome Research*,

14(9), pp. 3555–3567. doi:10.1021/acs.jproteome.5b00264.

Yadav, A.K. *et al.* (2013) ‘ProteoStats - A library for estimating false discovery rates in proteomics pipelines’, *Bioinformatics*, 29(21), pp. 2799–2800. doi:10.1093/bioinformatics/btt490.

Yi, X. *et al.* (2018) ‘Quality control of single amino acid variations detected by tandem mass spectrometry’, *Journal of Proteomics*, 187(July), pp. 144–151. doi:10.1016/j.jprot.2018.07.004.

Zahn-Zabal, M. *et al.* (2020) ‘The neXtProt knowledgebase in 2020: Data, tools and usability improvements’, *Nucleic Acids Research*, 48(D1), pp. D328–D334. doi:10.1093/nar/gkz995.

Zhu, Y. *et al.* (2018) ‘Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow’, *Nature Communications*, 9(1), p. 903. doi:10.1038/s41467-018-03311-y.

Zickmann, F. and Renard, B.Y. (2015) ‘MSProGene: integrative proteogenomics beyond six-frames and single nucleotide polymorphisms’, *Bioinformatics*, 31(12), pp. i106–i115. doi:10.1093/bioinformatics/btv236.

Figures and Tables

Figure legends

Figure 1: Workflow for identification and post search validation of variant peptides from proteogenomics searches.

Figure 2: Workflow for generation and analysis of simulation data to evaluate the Variant Ambiguity Score (VAS)

Figure 3: Distribution of features for identified PSMs from simulated data, showing segregation of true and false hits against VAS scores.

Figure 4: Distribution of features for manually annotated spectra to evaluate VAS scoring. Distribution of features for – (a) match quality, and (b) variant events. Manually annotated spectra good, average and bad PSMs are shown in green, blue and red, respectively. (c) Composition of VAS formula to combine the various features (d) Density distribution of VAS annotated spectra showing a clear segregation between good and bad hits (e) VAS Classification validation with manually annotated search results and isobaric filter.

Figure 5: Comparison between gFDR and csFDR with respect to classification of variant peptides in known, novel and variant a) Number of identifications of non-redundant PSMs as Known, Novel and Variant PSMs b) Percentage difference of identification between gFDR and csFDR.

Figure 6: Spectra classified with VAS scoring in a) confident b) ambiguous c) poor.

Figure 7: Frequency of mutational changes mapped by PgxSAVy.

Figure 8: Disease annotation and frequency of SAAVs.

Supplementary Figure 1 : Percolator and VAS comparison shows that FDR results are not amenable directly for ML based segregation of variant events

Table legends

Table 1: Composition of Simulated MS Input and search database.

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

Supplementary Table 1: Potential Isobaric substitution list.

Figure 1

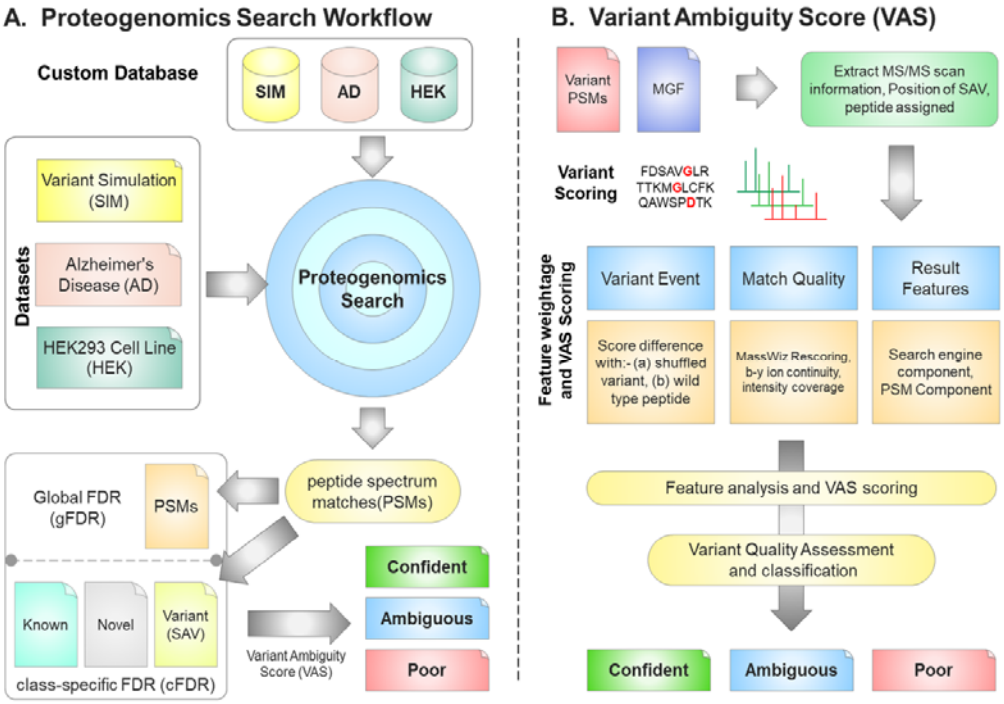

A. Proteogenomics Search Workflow
Custom Database
SIM
AD
HEK
Datasets
Variant Simulation (SIM)
Alzheimer's Disease (AD)
HEK293 Cell Line (HEK)
Proteogenomics Search
peptide spectrum matches(PSMs)
Global FDR (gFDR)
PSMs
Known
Novel
Variant (SAV)
class-specific FDR (cFDR)
Variant Ambiguity Score (VAS)
Confident
Ambiguous
Poor
B. Variant Ambiguity Score (VAS)
Variant PSMs
MGF
Extract MS/MS scan information, Position of SAV, peptide assigned
Variant Scoring
FDSAVGLR
TTKMGLCFK
QAWSPDTK
Feature weightage and VAS Scoring
Variant Event
Match Quality
Result Features
Score difference with:- (a) shuffled variant, (b) wild type peptide
MassWiz Rescoring, b-y ion continuity, intensity coverage
Search engine component, PSM Component
Feature analysis and VAS scoring
Variant Quality Assessment and classification
Confident
Ambiguous
Poor

Figure 2

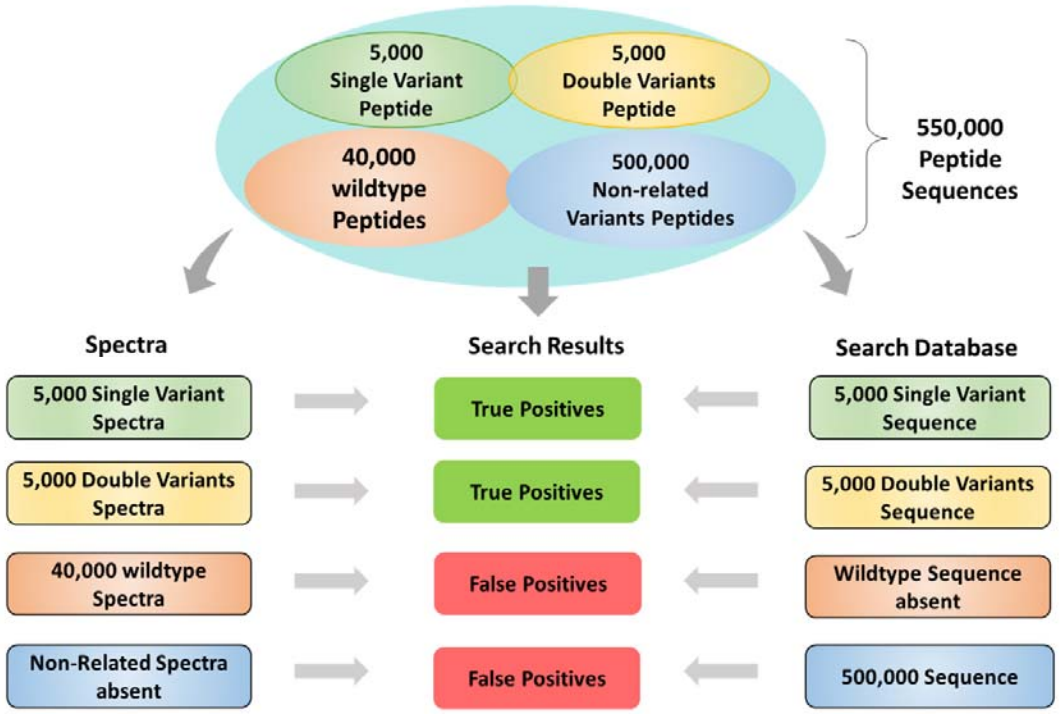

5,000
Single Variant
Peptide
5,000
Double Variants
Peptide
40,000
wildtype
Peptides
500,000
Non-related
Variants Peptides
550,000
Peptide
Sequences
Spectra
Search Results
Search Database
5,000 Single Variant
Spectra
True Positives
5,000 Single Variant
Sequence
5,000 Double Variants
Spectra
True Positives
5,000 Double Variants
Sequence
40,000 wildtype
Spectra
False Positives
Wildtype Sequence
absent
Non-Related Spectra
absent
False Positives
500,000 Sequence

Figure 3

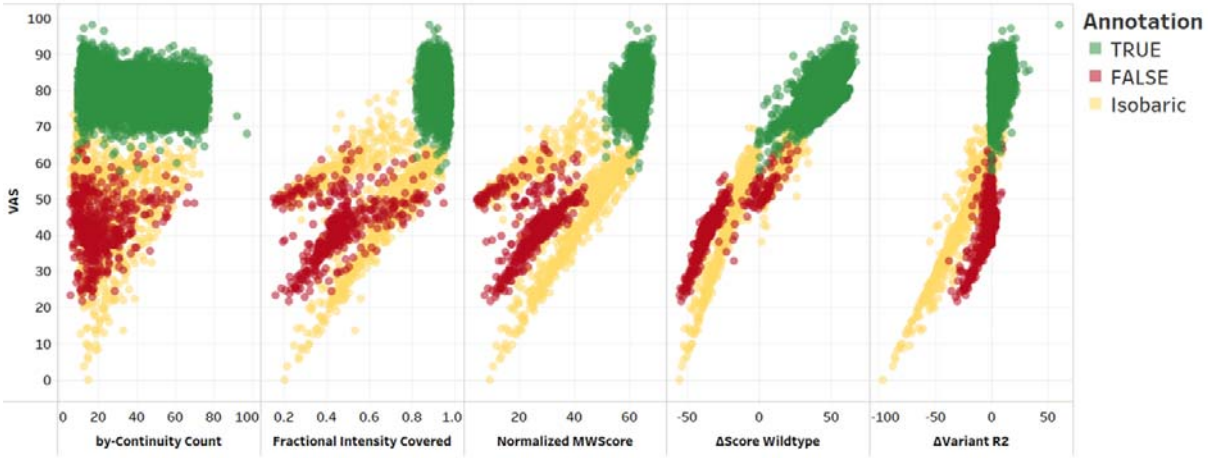

Annotation
TRUE
FALSE
Isobaric
VAS
by-Continuity Count
Fractional Intensity Covered
Normalized MWScore
ΔScore Wildtype
ΔVariant R2

Figure 4

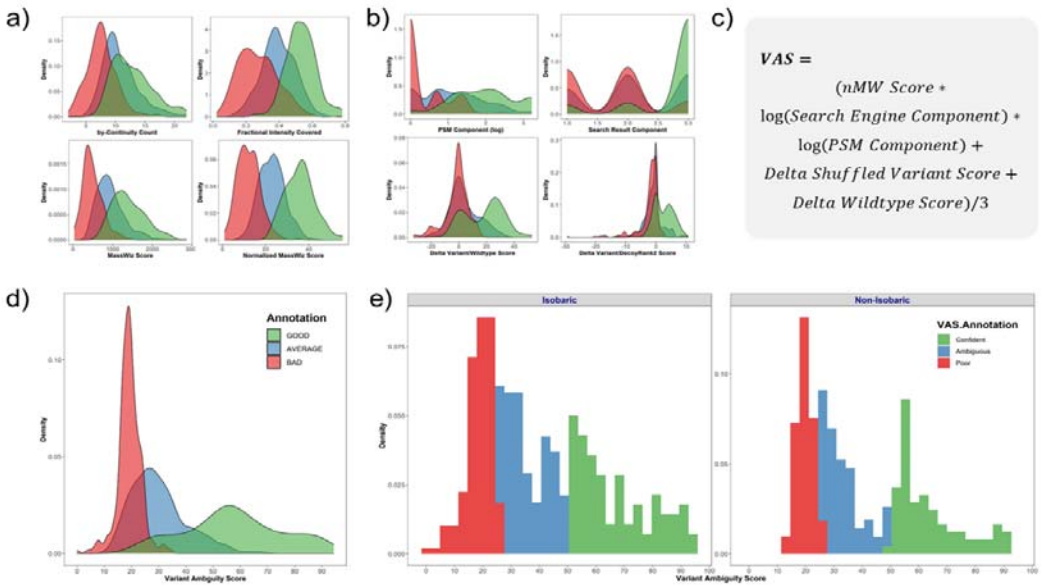

a)
b)
c)
VAS =
(nMW Score *
log(Search Engine Component) *
log(PSM Component) +
Delta Shuffled Variant Score +
Delta Wildtype Score)/3
d)
Annotation
GOOD
AVERAGE
BAD
Density
Variant Ambiguity Score
e)
Isobaric
Non-Isobaric
VAS.Annotation
Density
Variant Ambiguity Score

Figure 5

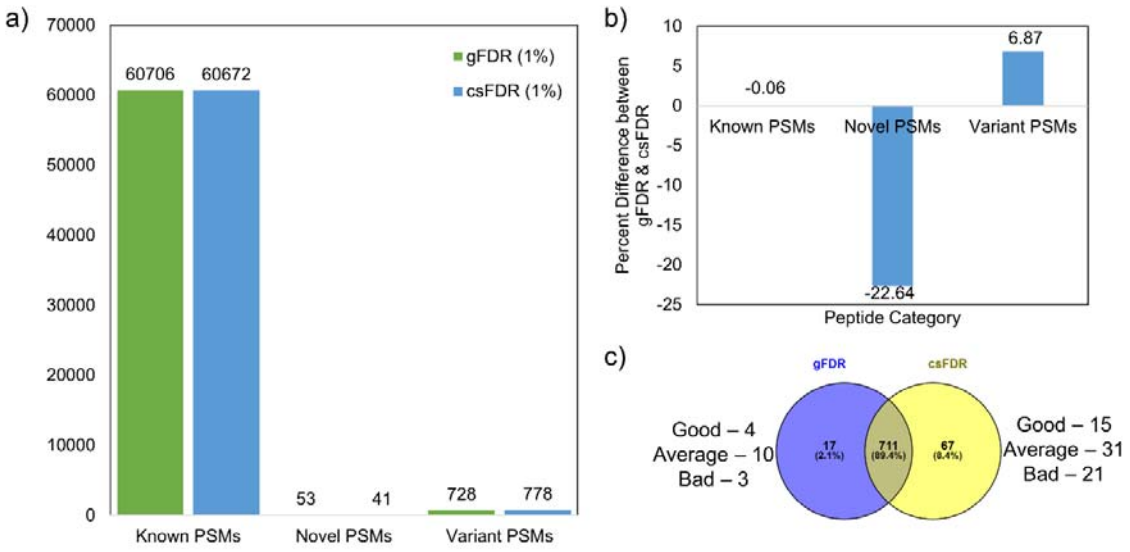

a)
70000
60000
50000
40000
30000
20000
10000
0
60706
60672
53
41
728
778
gFDR (1%)
csFDR (1%)
Known PSMs
Novel PSMs
Variant PSMs
b)
10
5
0
-5
-10
-15
-20
-25
-0.06
-22.64
6.87
Known PSMs
Novel PSMs
Variant PSMs
Percent Difference between gFDR & csFDR
Peptide Category
c)
gFDR
csFDR
Good – 4
Average – 10
Bad – 3
17 (2.1%)
711 (89.4%)
67 (8.4%)
Good – 15
Average – 31
Bad – 21

Figure 6

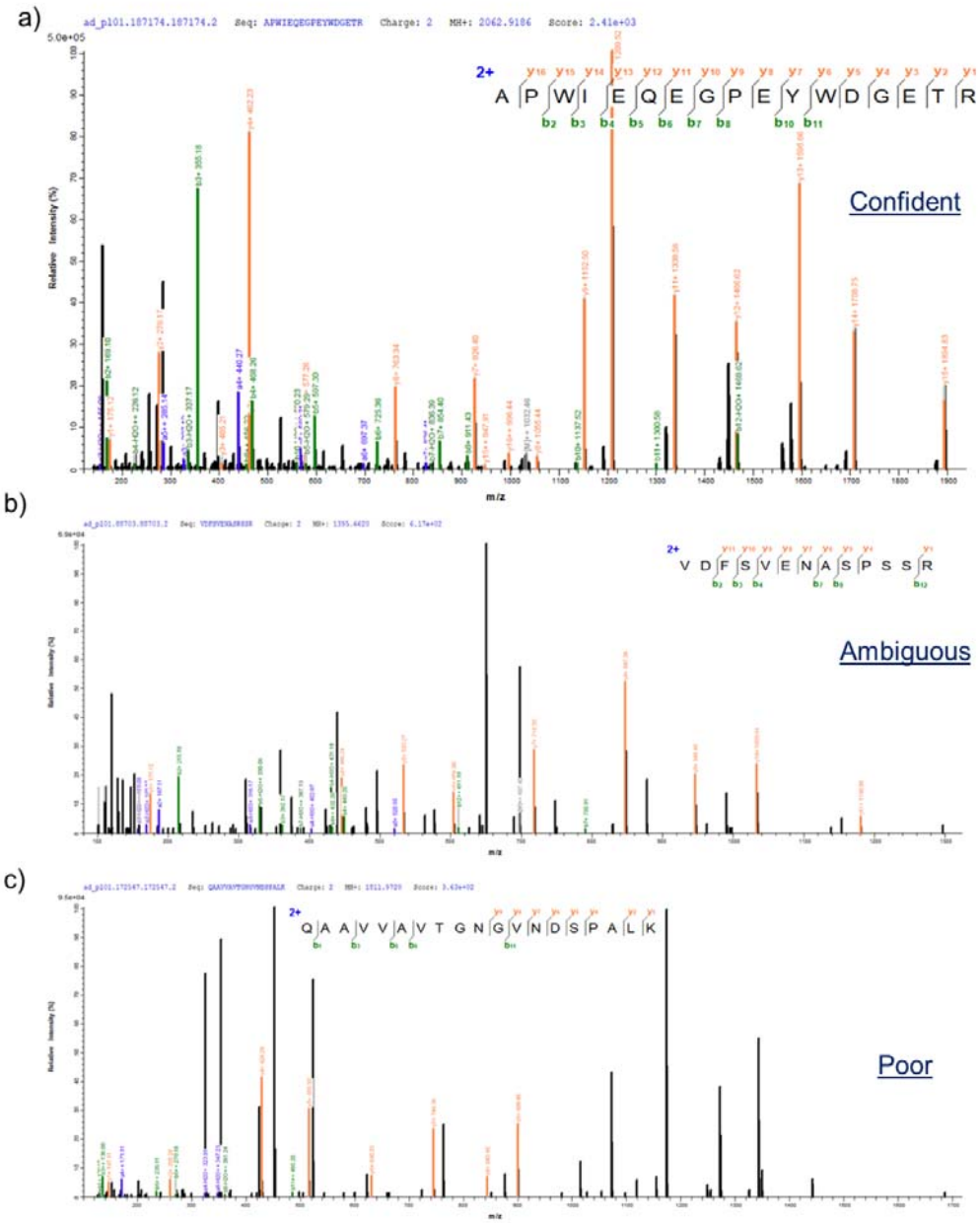

a)
2+
A P W I E Q E G P E Y W D G E T R
Confident
Relative Intensity (%)
m/z
b)
2+
V D F S V E N A S P S S R
Ambiguous
c)
2+
Q A A V V A V T G N G V N D S P A L K
Poor

Figure 7

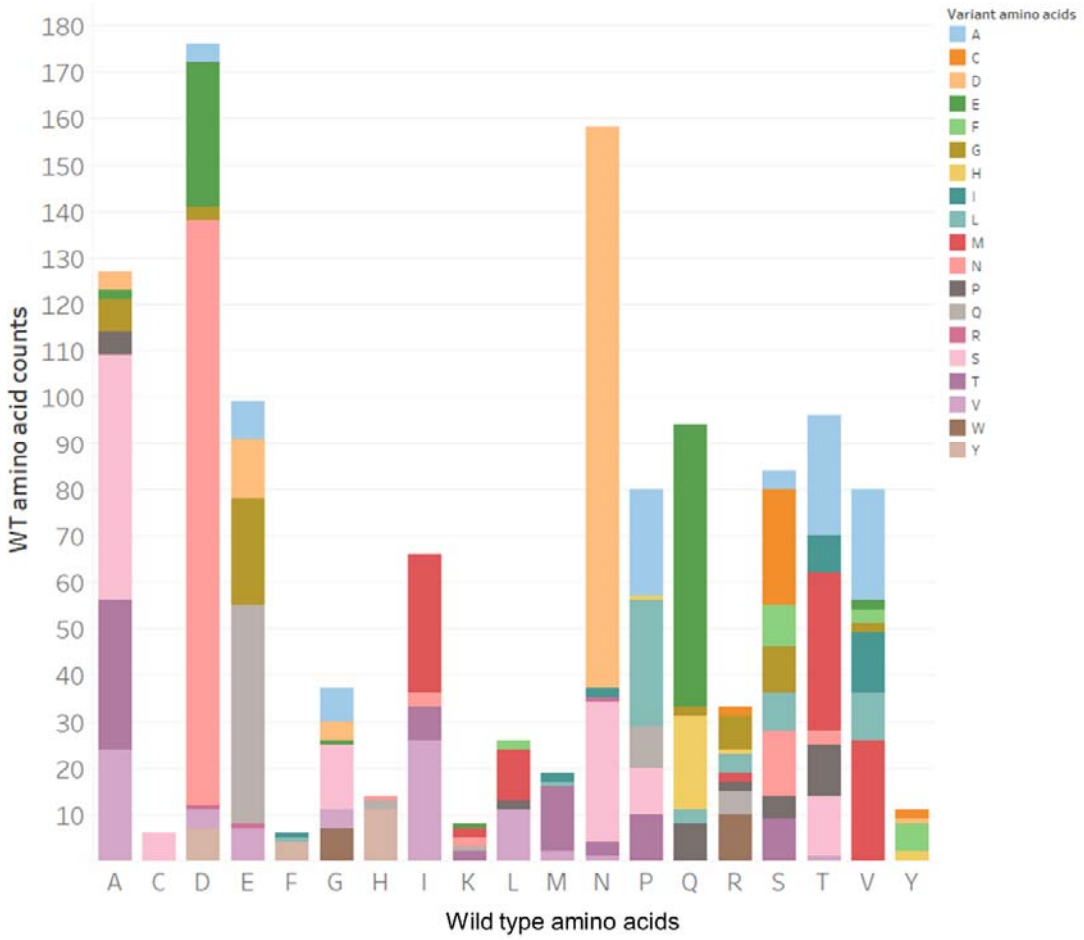

Variant amino acids
A
C
D
E
F
G
H
I
L
M
N
P
Q
R
S
T
V
W
Y
WT amino acid counts
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
A C D E F G H I K L M N P Q R S T V Y
Wild type amino acids

Figure 8

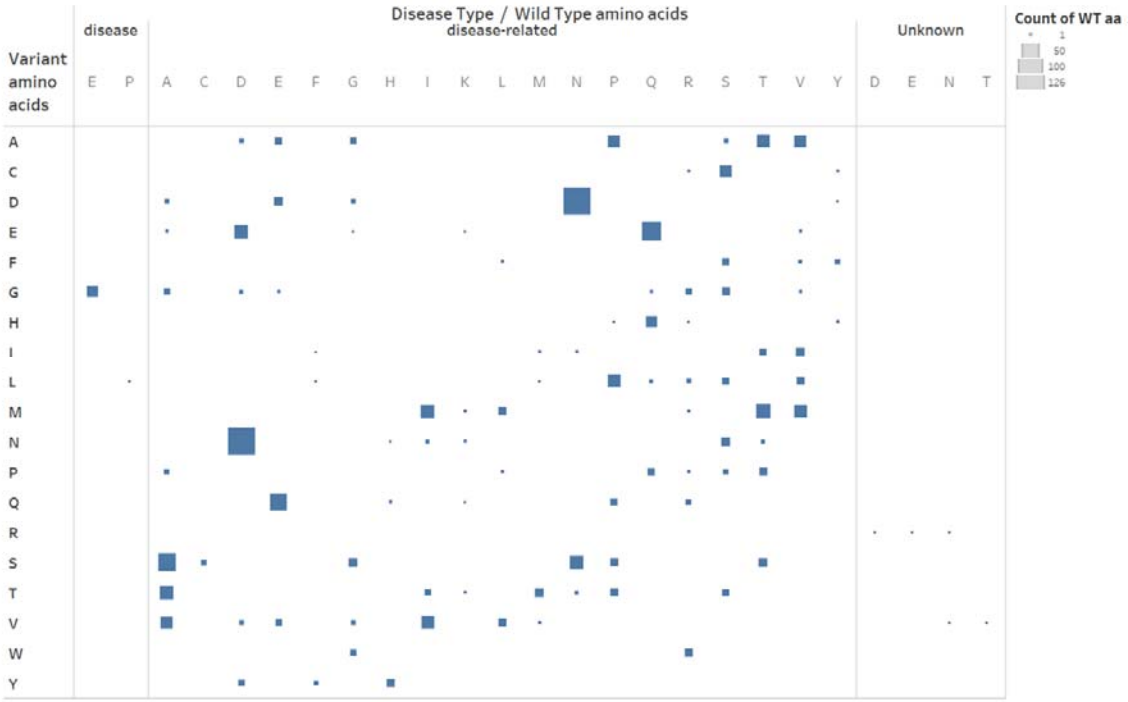
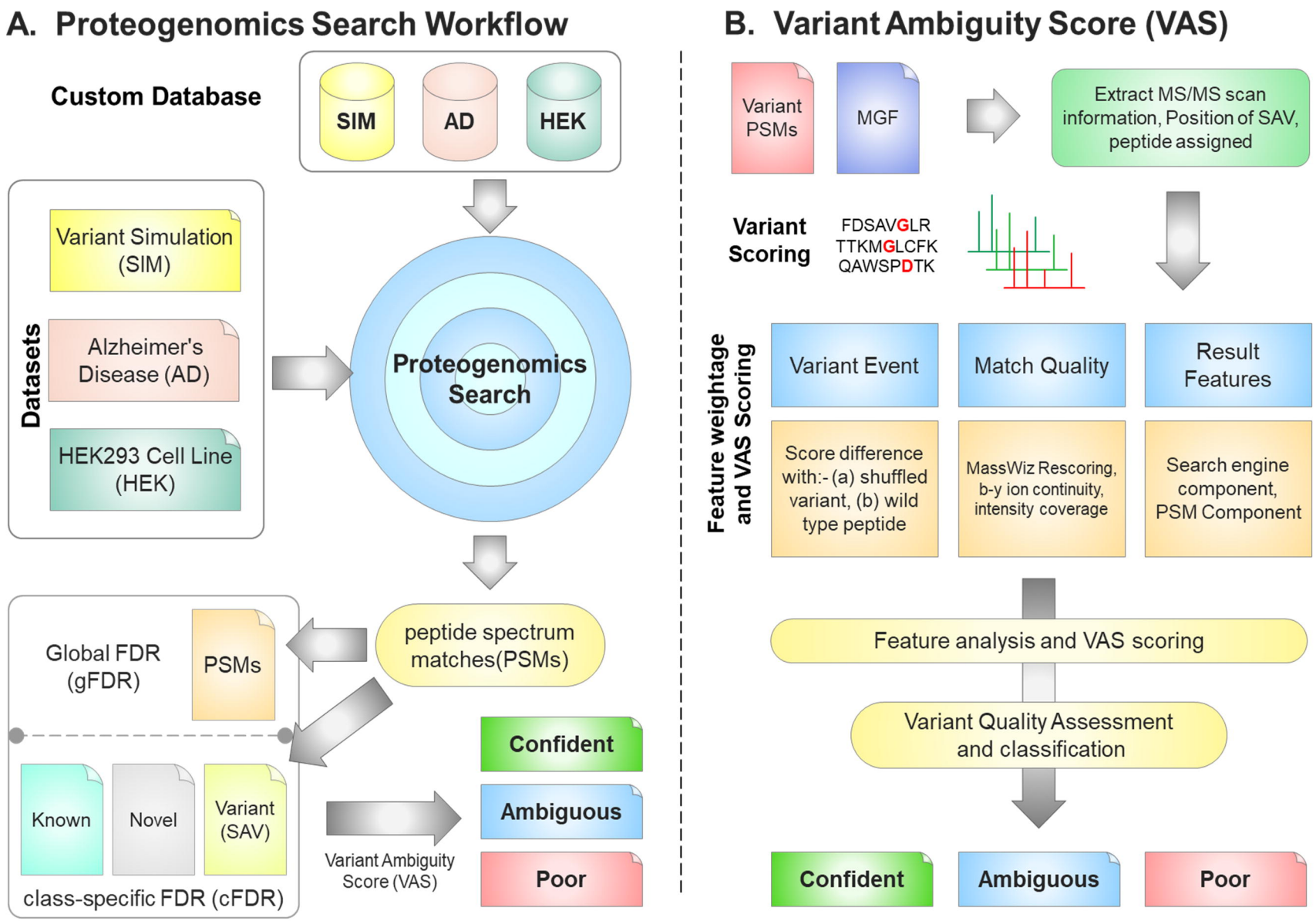

Disease Type / Wild Type amino acids
disease
disease-related
Unknown
Count of WT aa
1
50
100
126
Variant amino acids
E P A C D E F G H I K L M N P Q R S T V Y D E N T
A
C
D
E
F
G
H
I
L
M
N
P
Q
R
S
T
V
W
Y

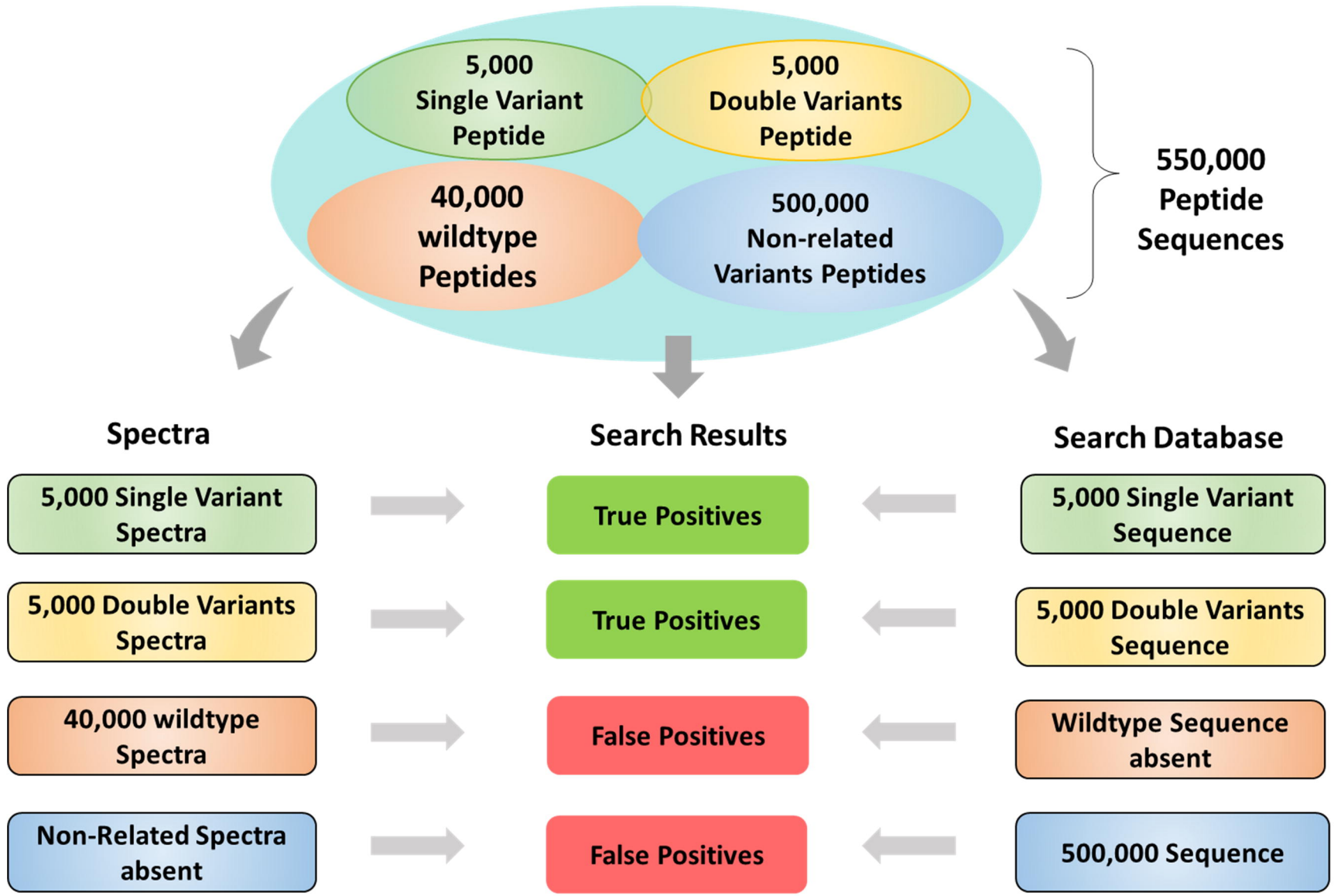
Table 1: Composition of Simulated MS Input and search database.

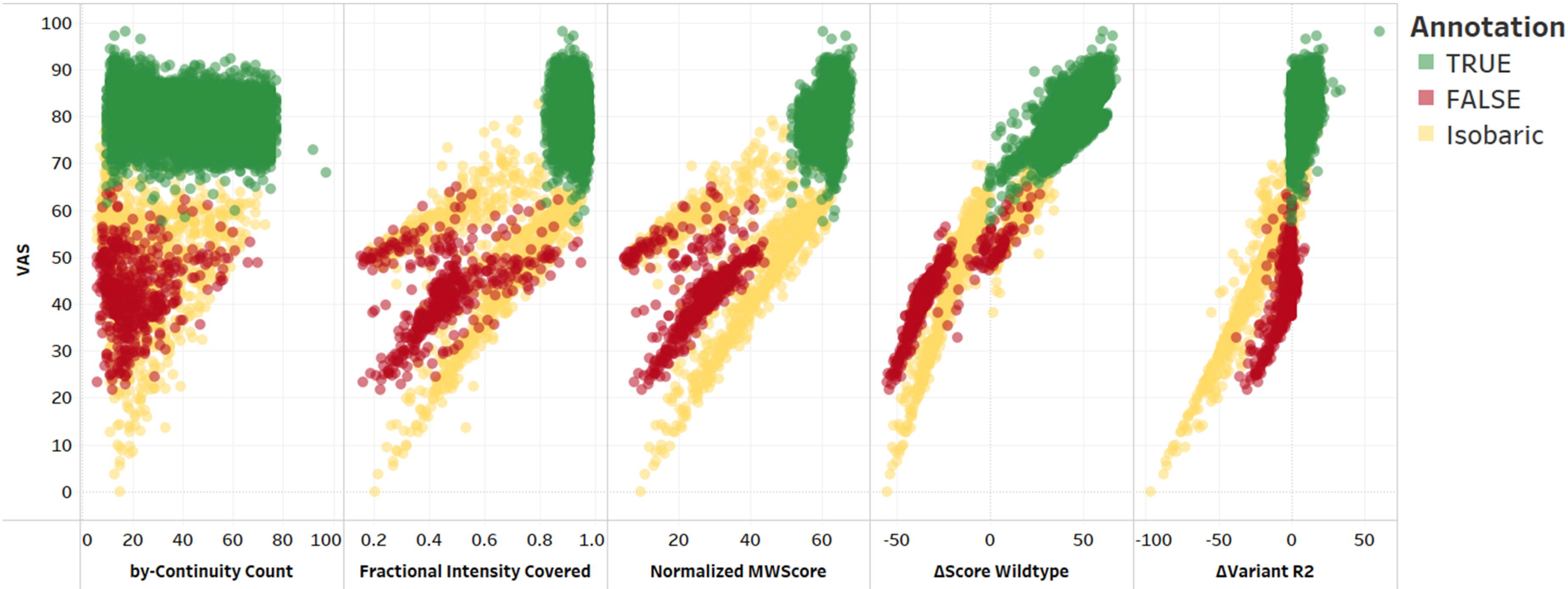
Spectra	Database	Reason
Spectra with 1 variant: 5000	5000 Variant peptides (1 variant)	True Positives
Spectra with 2 variants: 5000	5000 Variant peptides (2 variant)	True Positives
Spectra without any variant (Wildtype): 40,000	Sequence absent corresponding to these 40,000 peptides	False Positives (To enhance the competition and make WT spectra hit on variant)
	5,00,000 Not related variant peptides which are absent in simulates spectra	False Positives

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

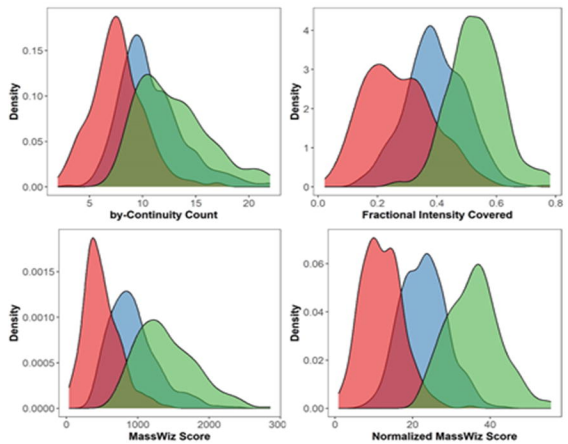
VAS	AD dataset	HEK Cell line
Total	9853	4011
Confident	1171	35
Ambiguous	3726	472
Poor	4696	3504


A. Proteogenomics Search Workflow
Custom Database
SIM
AD
HEK
Datasets
Variant Simulation (SIM)
Alzheimer's Disease (AD)
HEK293 Cell Line (HEK)
Proteogenomics Search
peptide spectrum matches(PSMs)
Global FDR (gFDR)
PSMs
Known
Novel
Variant (SAV)
class-specific FDR (cFDR)
Variant Ambiguity Score (VAS)
Confident
Ambiguous
Poor
B. Variant Ambiguity Score (VAS)
Variant PSMs
MGF
Extract MS/MS scan information, Position of SAV, peptide assigned
Variant Scoring
FDSAVGLR
TTKMGLCFK
QAWSPDTK
Feature weightage and VAS Scoring
Variant Event
Match Quality
Result Features
Score difference with:- (a) shuffled variant, (b) wild type peptide
MassWiz Rescoring, b-y ion continuity, intensity coverage
Search engine component, PSM Component
Feature analysis and VAS scoring
Variant Quality Assessment and classification
Confident
Ambiguous
Poor

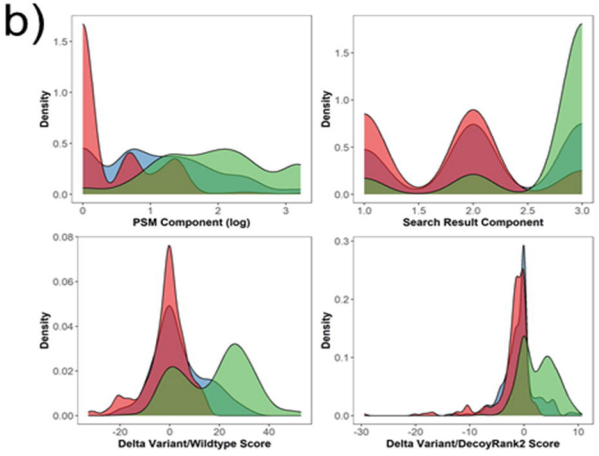

5,000
Single Variant
Peptide
5,000
Double Variants
Peptide
40,000
wildtype
Peptides
500,000
Non-related
Variants Peptides
550,000
Peptide
Sequences
Spectra
Search Results
Search Database
5,000 Single Variant Spectra
True Positives
5,000 Single Variant Sequence
5,000 Double Variants Spectra
True Positives
5,000 Double Variants Sequence
40,000 wildtype Spectra
False Positives
Wildtype Sequence absent
Non-Related Spectra absent
False Positives
500,000 Sequence


VAS
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100
by-Continuity Count
0.2
0.4
0.6
0.8
1.0
Fractional Intensity Covered
20
40
60
Normalized MWScore
-50
0
50
ΔScore Wildtype
-100
-50
0
50
ΔVariant R2
Annotation
TRUE
FALSE
Isobaric

a)



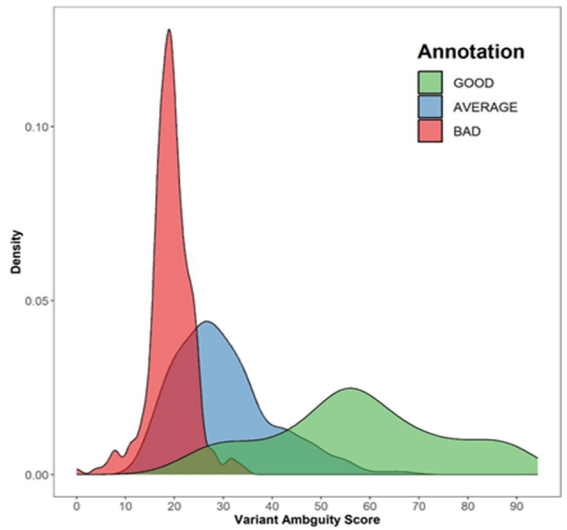
b)



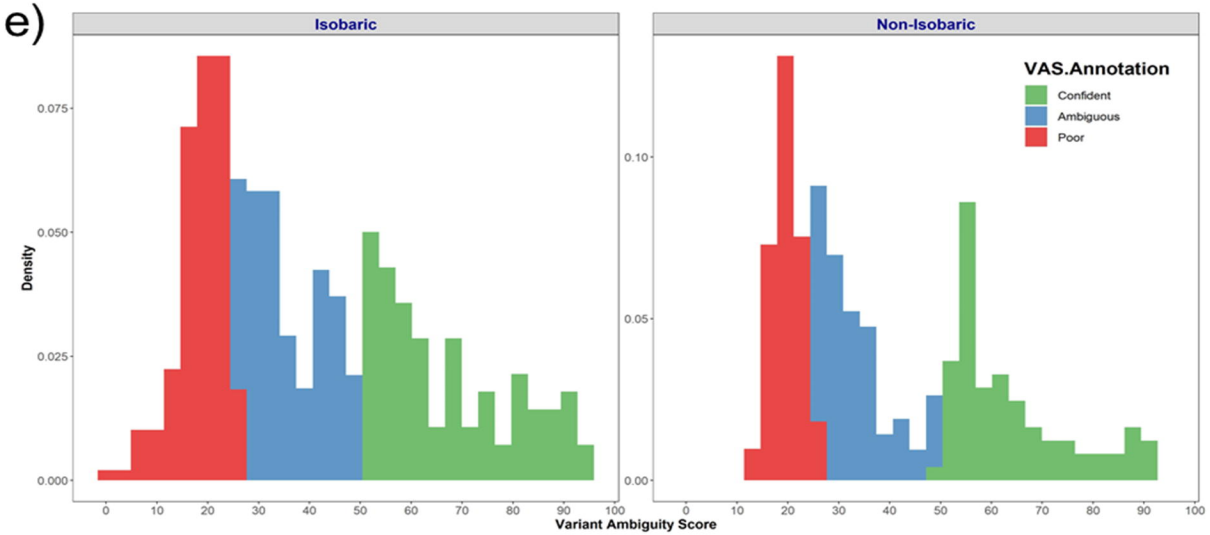
c)

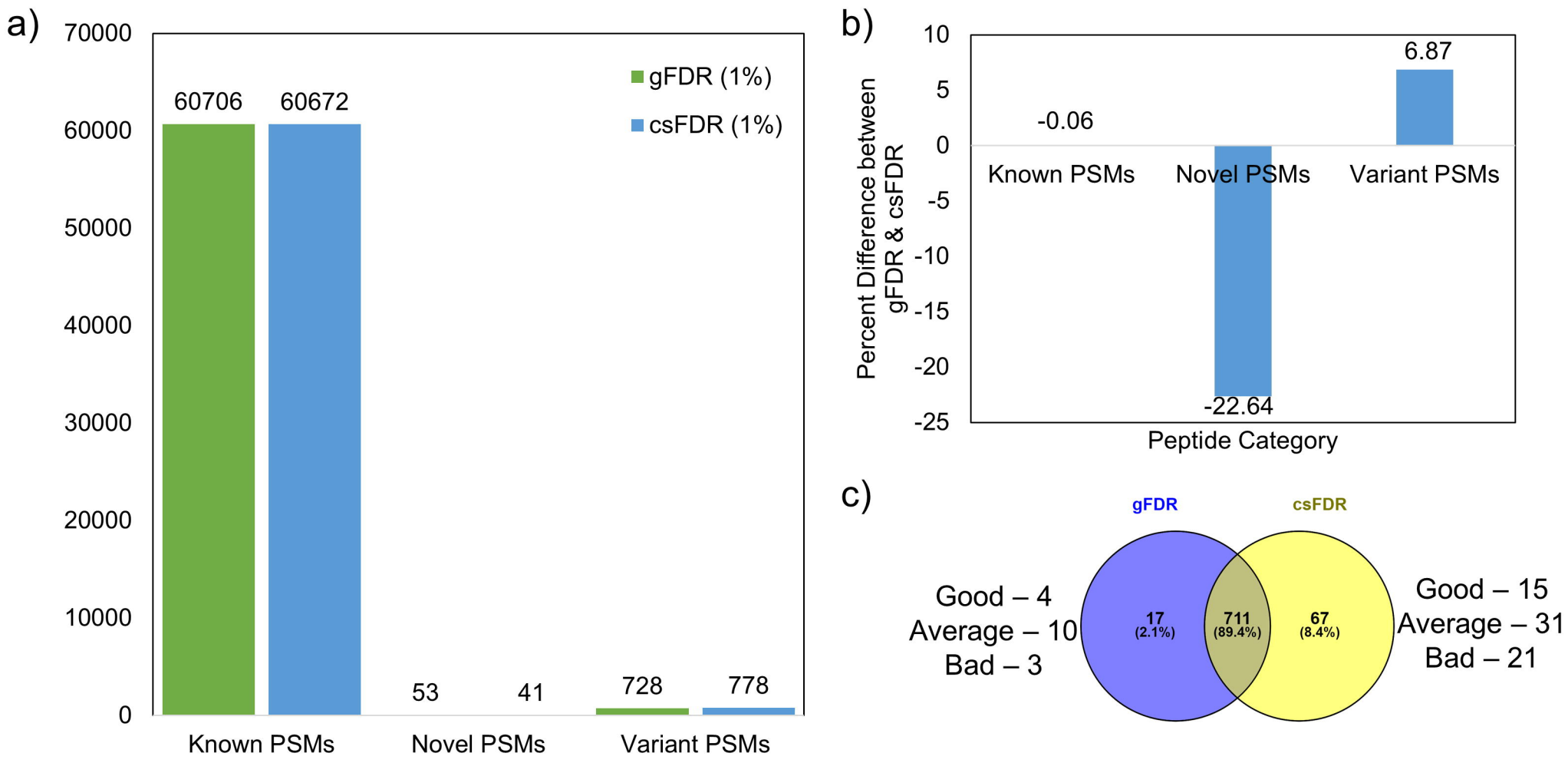
$$\boldsymbol{VAS} = (nMW\ Score * \log(Search\ Engine\ Component) * \log(PSM\ Component) + Delta\ Shuffled\ Variant\ Score + Delta\ Wildtype\ Score)/3$$

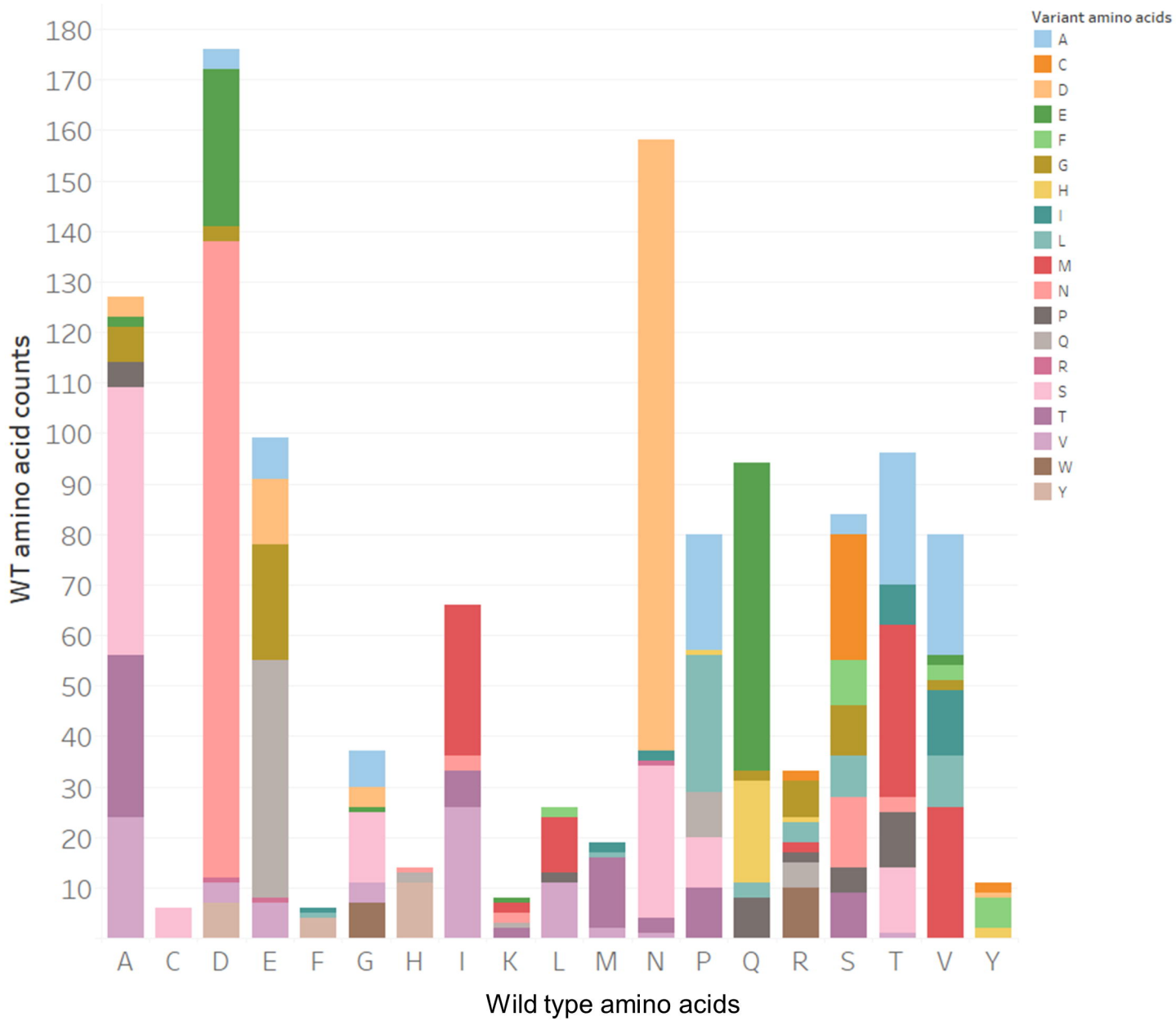
d)

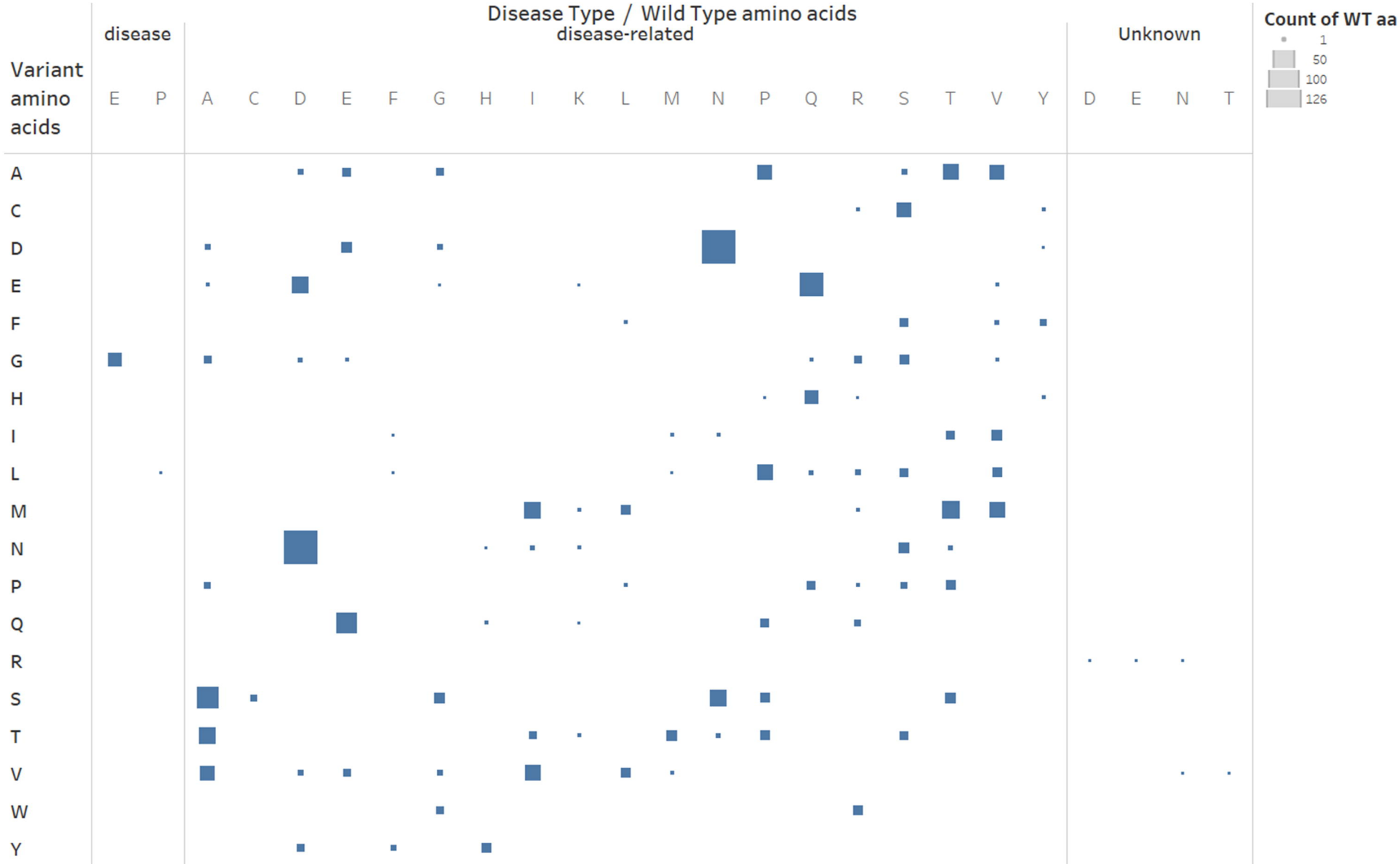


e)

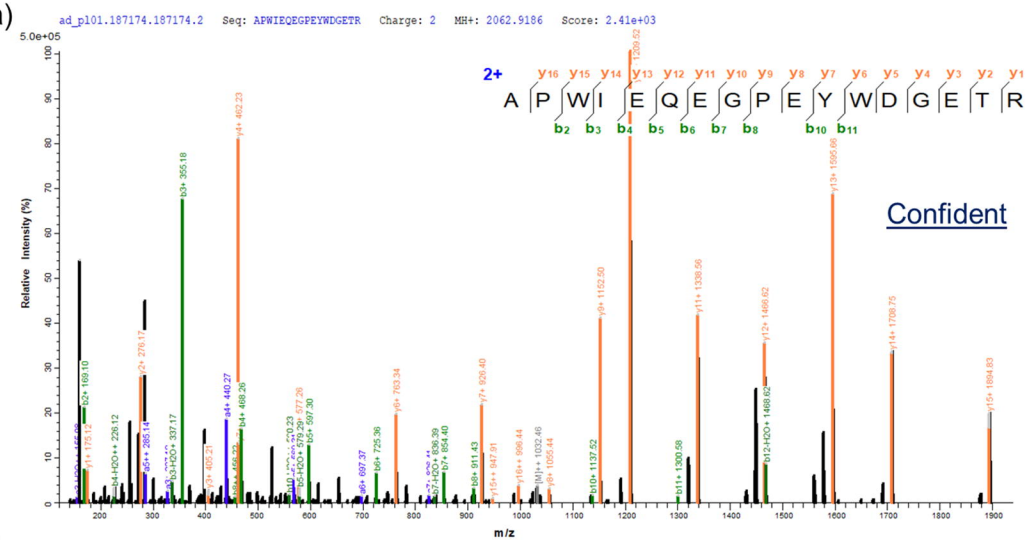



a)
70000
60000
50000
40000
30000
20000
10000
0
60706
60672
53
41
728
778
gFDR (1%)
csFDR (1%)
Known PSMs
Novel PSMs
Variant PSMs
b)
10
5
0
-5
-10
-15
-20
-25
Percent Difference between gFDR & csFDR
-0.06
-22.64
6.87
Known PSMs
Novel PSMs
Variant PSMs
Peptide Category
c)
gFDR
csFDR
Good – 4
Average – 10
Bad – 3
17
(2.1%)
711
(89.4%)
67
(8.4%)
Good – 15
Average – 31
Bad – 21

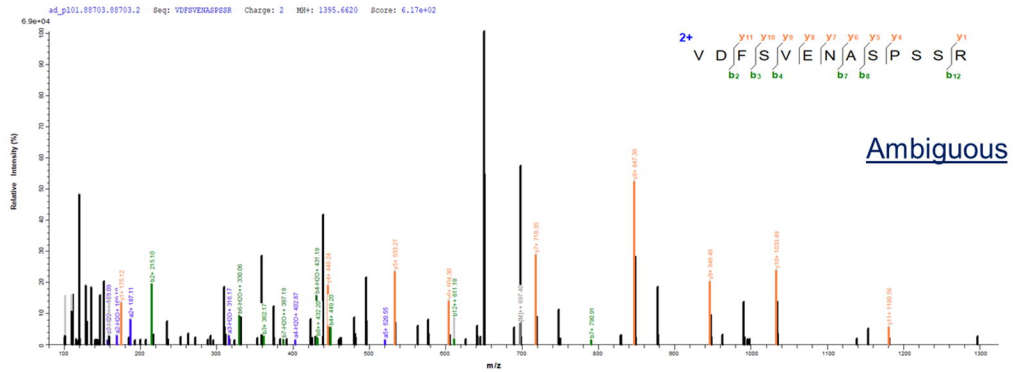

Variant amino acids
A
C
D
E
F
G
H
I
L
M
N
P
Q
R
S
T
V
W
Y
WT amino acid counts
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
A C D E F G H I K L M N P Q R S T V Y
Wild type amino acids


Disease Type / Wild Type amino acids
disease
disease-related
Unknown
Variant amino acids
E P A C D E F G H I K L M N P Q R S T V Y D E N T
A
C
D
E
F
G
H
I
L
M
N
P
Q
R
S
T
V
W
Y
Count of WT aa
1
50
100
126

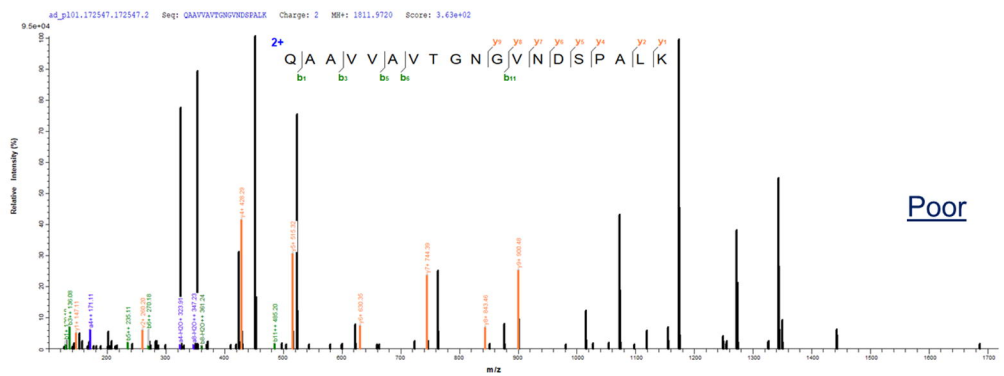
a)


ad_p101.187174.187174.2 Seq: APWIEQEGPEYWDGETR Charge: 2 MH+: 2062.9186 Score: 2.41e+03
2+
A P W I E Q E G P E Y W D G E T R
Confident
Relative Intensity (%)
m/z

b)


2+
V D F S V E N A S P S S R
Ambiguous
Relative Intensity (%)
m/z

c)


2+
Q A A V V A V T G N G V N D S P A L K
Poor
Relative Intensity (%)
m/z