Journal of proteome research

pubs.acs.org/jpr

Article

Personalized Proteome: Comparing Proteogenomics and Open Variant Search Approaches for Single Amino Acid Variant Detection

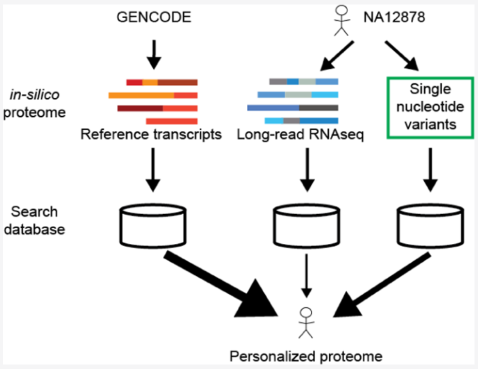
Renee Salz, Robbin Bouwmeester, Ralf Gabriels, Sven Degroeve, Lennart Martens, Pieter-Jan Volders, and Peter A.C. ’t Hoen*

Cite This: *J. Proteome Res.* 2021, 20, 3353−3364

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Discovery of variant peptides such as a single amino acid variant (SAAV) in shotgun proteomics data is essential for personalized proteomics. Both the resolution of shotgun proteomics methods and the search engines have improved dramatically, allowing for confident identification of SAAV peptides. However, it is not yet known if these methods are truly successful in accurately identifying SAAV peptides without prior genomic information in the search database. We studied this in unprecedented detail by exploiting publicly available long-read RNA sequences and shotgun proteomics data from the gold standard reference cell line NA12878. Searching spectra from this cell line with the state-of-the-art open modification search engine *ionbot* against carefully curated search databases resulted in 96.7% false-positive SAAVs and an 85% lower true positive rate than searching with peptide search databases that incorporate prior genetic information. While adding genetic variants to the search database remains indispensable for correct peptide identification, inclusion of long-read RNA sequences in the search database contributes only 0.3% new peptide identifications. These findings reveal the differences in SAAV detection that result from various approaches, providing guidance to researchers studying SAAV peptides and developers of peptide spectrum identification tools.



KEYWORDS: *long-read RNA sequence, deep proteomics, open search, direct RNA sequencing*

■ INTRODUCTION

Proteomes display significant interindividual variability[1,2] and personal proteomes may delineate disease risk and pave the way for personalized disease prevention and treatment. Personalized cancer treatment, for instance, is already instigated based on the detection of peptides containing single amino acid variants (SAAVs) that often serve as excellent biomarkers.[3−8] Detecting these SAAV peptides reliably, however, is a formidable challenge. Previously, scientists looked for protein evidence of a small number of variants in particular and resorted to targeted proteomics approaches such as selected reaction monitoring.[9−12] Alternatively, BLAST-like query tools such as Peptimapper and PepQuery[13,14] or database tools such as XMAn v2[15] and dbSAP[16] can be used to investigate single events.[17,18] Proteogenomics, the integration of genome and transcriptome information, is a more holistic and higher-throughput form of mass spectrometry (MS)-based detection of variant peptides.

A main limiting factor of SAAV peptide (called “variant peptide” in the remainder of the paper) detection with shotgun proteomics is the tandem MS (MS/MS) technology itself. Since MS/MS spectra are generally too noisy to call a peptide sequence de novo, current MS/MS analysis methods rely on a database of known peptides. This limits the ability to detect unknown peptides such as variant peptides. The most flexible way to detect variant peptides is an exhaustive search; allowing any possible amino acid substitution at any position in the peptide sequence.[19,20] However, this strategy increases the search space immensely to a point where it is no longer useful in practice. The larger search space leads to ambiguity in peptide identification and thus, a high number of false-positive hits.[21,22] Therefore, more careful curation of sequences in the search database pays off.

Databases of peptides containing variants from dbSNP have been created to facilitate the search for SAAVs,[3,16] and simply adding these variant peptides to the database showed promise early on.[3,23] Not all dbSNP variants, however, are expected to be found in every sample, and including them all may lead to false identifications.[24] In addition, rare and unique variants may be overlooked. A proteogenomics approach where only those variant peptides predicted from genome or transcriptome

Received: April 2, 2021
Published: May 17, 2021

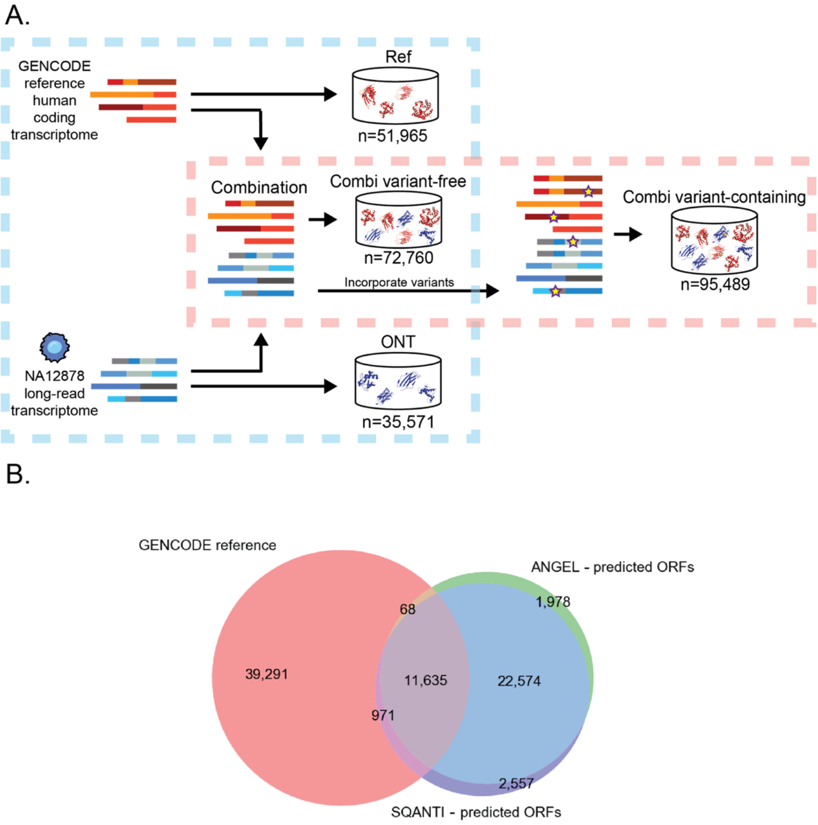


Figure 1. Creation of the search databases. (A) Three databases were made to make comparison between use of different sources of sequences. One with only translations of transcriptome sequences (ONT), one with only the reference proteome (GENCODE), and one with the union of the two. This comparison is denoted with a blue square. Variants from NA12878 were incorporated into the combination database from A and compared to the combination database without variants. This comparison is denoted with a red square. (B) Number of (predicted) ORFs in the different sources used to construct the VF search database and their overlap. The sources included the GENCODE v29 reference ORFs and the predicted ORFs from ONT RNAseq. Two ORF prediction software (ANGEL and SQANTI) were used to determine candidate ORFs, and the intersection was included in the final search database.

information are added to the peptide search databases can improve their detection. Proteogenomics pipelines have streamlined this process of incorporating personal genome information into a proteomic search database.[25−29] In addition, there is evidence that including correct sequence variant information, including often-overlooked sample-specific indels and frameshifts, improves variant peptide identification workflows.[30] Yet, false discovery rate (FDR) correction is needed to compensate for the increase of database size and complexity.[21,22] When searching for evidence of specific peptides such as variant peptides, an additional subset-specific FDR correction should be made.[31]

In addition to SAAVs, alternative splicing may also introduce sample specific peptides. Alternative splicing is commonplace as 90% of genes undergo alternative splicing.[32] Since protein reference databases do not cover all protein isoforms produced by alternative splicing, sample-specific transcriptome information is advantageous. Typically, the information on alternatively spliced sequences comes from RNA sequencing. A short-read RNA sequence is however not ideal for properly capturing the complete splicing patterns and the resulting open reading frames (ORFs). Traditionally, this is circumvented by including 3- or 6-frame translations of the sample's transcriptome. However, this approach was found to expand the database far too much for eukaryotic organisms, leaving few remaining hits after FDR correction.[33] Studies utilizing long-read RNA sequences frequently discover previously unannotated transcript structures. Thus, full-length transcripts may add essential information for correct ORF prediction and peptide identification.

An emerging alternative to proteogenomics methods for the detection of variant peptides is the "open search" method. This allows unexpected post-translational modifications and amino acid substitutions in the peptide spectrum match while maintaining an accurate FDR and a workable computation time. Using sequence tag-based approaches, the search space is narrowed with de novo sequence tags, which makes room for the addition of all possible SAAV peptides in the search space.[34−38] These methods were historically not as effective as classical proteogenomics searches in finding variant peptides since there is difficulty in discerning between post-translationally modified and SAAV peptides. However, this situation has recently improved with the inclusion of optimized probabilistic models.[39] One implementation of the tag-based method improved with such models is *ionbot* (manuscript in preparation; compomics.com/ionbot), which is a machine learning search engine that uses MS2PIP[40] and ReSCore[41] to significantly improve the accuracy of peptide match scoring.

The main objective of this study is to compare a previously established proteogenomics approach based on long-read sequencing with a recently-developed open search method for the detection of true variant peptides. In simpler terms, we compare a genome-informed search space with typical spectrum identification settings to a genome-uninformed search space with advanced identification settings. We aim to understand the power of, and potential biases associated with, using an open search method without prior information about the genome. For this, we make use of high-confidence nucleotide sequencing and (ultra)-deep proteomics data from a gold standard cell line NA12878. Using correct ORFs from the long-read transcriptome and high-confidence phased variants belonging to this cell line, we gain a unique perspective on exactly what advantages can be gained by each approach.

■ EXPERIMENTAL SECTION

NA12878 Data Sources

Variant information was obtained from Illumina platinum genomes (ftp://platgene_ro@ussd-ftp.illumina.com/2017-1.0/hg38/small_variants/NA12878/). The reference genome used was GRCh38, which can be downloaded from the precomputed 1000 genomes GRCh38 BWA database at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/GRCh38_reference_genome/ (with decoys). Transcript structures for NA12878 were sourced from the ONT consortium.[42] In the consortium, Workman et al. sequenced 9.9 million reads corresponding to 33,894 transcripts and 20,289 genes. The reference transcriptome and proteome are from GENCODE v29.

Shotgun proteomics data came from the[43] study, downloaded from Peptide Atlas (http://www.peptideatlas.org/PASS/PASS00230). This data set consists of 417 TMT6plex runs from 54 samples, with the reference tag (126.77) on NA12878 in every case.

Creation of the Search Databases

In total, four search databases were created (Table S1). (1) Database based on ONT transcriptome sequences only (referred to as "ONT"), (2) database based on GENCODE coding transcriptome only (referred to as "Ref"), (3) a database, i.e., the union of (1) and (2) and contains no NA12878 specific variants (referred to as variant-free or VF), and (4) the same sequences as the database (3) but contains NA12878 specific variants (referred to as variant-containing or VC). A simple depiction can be found in Figure 1A and Table S1, while the detailed full workflow can be found in Figure S1. Each database had MaxQuant[44] contaminant sequences appended before search.

The Ref search database was made by filtering GENCODE v29-predicted ORFs for those that were complete (no 5′ or 3′ missingness). The ONT database was created using transcript structures provided by the NA12878 consortium (https://github.com/nanopore-wgs-consortium/NA12878/blob/master/RNA.md). The coordinates in the junction file (PSL format) provided were converted to BED with BEDOPS[45] and used to fetch the corresponding stretch of the sequence from the GRCh38 genome with bedtools[46] getfasta. The exons were assembled using in-house scripts to form the full transcripts, and those that were nonidentical to transcript sequences in GENCODE ("novel") were then submitted to two ORF prediction software; ANGEL v2.4 ("dumb" ORF prediction on default settings) and SQANTI2 v2.7 (https://github.com/Magdoll/SQANTI2). The translations of transcripts predicted by both prediction programs were added to the search database. ORFs from GENCODE were used for transcript sequences in ONT identical to transcript sequences in GENCODE.

The VF database was simply the union of the Ref and ONT databases. The VC database was created by first creating full-length coding sequences (CDS) with variants included by replacing reference nucleotides according to the VCF file per CDS fragment for every CDS fragment. If only homozygous variant(s) were present in a CDS fragment, only one variant CDS fragment was generated. If a CDS contained at least one heterozygous variant, two variant CDS sequences were generated corresponding to the different alleles. Fragments were then assembled to full CDS. If a full CDS contained at least one CDS fragment with a heterozygous variant, two full CDS were generated corresponding to each allele. For those full CDS that contained at least one variant, the variant version(s) of the sequences replaced the nonvariant versions in the VF database to create the VC database.

Spectral Search and Post Processing

Each run from Wu et al. 2013[43] was first converted to the mascot generic format using msconvert[47] with MS2 peak picking enabled. Each data set was then searched against the four search databases described in the previous section, using *ionbot* version 0.5. Fixed and variable modifications were set according to the protocol in Wu et al.[43] Open modification settings were enabled for all four runs, while open variant settings (for SAAV detection) were enabled for all runs except for on the VC database. Searches allowed for up to two missed cleavages. When parsing the search results, only spectra with an observed TMT6plex reporter ion 126.77 (corresponds to cell line NA12878) were retained.

Since subsetting PSMs into groups such as variant peptides requires separate FDR correction,[31] both VC and VF underwent a separate FDR correction for the variant peptide subset. Successful FDR correction requires the modeling of potential false-positive peptide identifications using appropriate decoy peptides. In the case of variant peptides, this means a sufficient number of decoy variant peptide identifications must be present to accurately model the population of false-positive peptides. Reversed sequences thus underwent the same processing steps as the true sequences in order to create the appropriate decoys. The distributions were checked for successful modeling (Figure S2).

A variant peptide list was created to compare with *ionbot* identifications from searches of the VC and VF. The list was created with an in-house Python script that performs an in-silico trypsin digest (allowing for up to two missed cleavages) with the pyteomics v 4.2[48] package and checks per protein for peptides that differ by only one amino acid between the VF and VC database. I and L were treated as identical, and a potential variant peptide was disqualified if it appears in any other reference protein sequence.

ionbot identifications presumed to be variant peptides (and variant peptide decoys) underwent subset-specific FDR correction for both combination databases, but the exact subset of variant peptides differed between the two searches due to different assumptions. The assumption in the VF database is that variants in the genome are unknown, so all predicted variant peptides (and predicted variant decoy peptides) were pooled for FDR correction. In the VC database, only known variant peptides (and corresponding decoy peptides) are pooled for FDR correction. We expect the different approaches to the subset FDR to be comparable, as *ionbot* does not include duplicate peptides in the search database. This means that the databases being compared are of similar size at the peptide level, which is the level at which the FDR correction is performed. q value calculation and cutoff ($q < 0.01$) were performed with an in-house python script (distribution can be seen in Figure S2).

Retention time predictions were calculated with DeepLC.[49] All scripts referred to in this paper can be found in the GitHub repository (https://github.com/cmbi/NA12878-saav-detection).

RESULTS

Search Database Makeup

The main goal of this study is to evaluate the added value of transcriptomics data for SAAV identification in proteomics data. In this evaluation, SAAV identification with and without transcriptomics prior knowledge is compared for a state-of-the-art open search engine. To this end, we searched the NA12878 deep shotgun proteomics data set with four distinct search databases corresponding to two comparisons, as outlined in Figure 1A. The first comparison was between databases based on the Oxford Nanopore (ONT) long-read transcriptome, the GENCODE reference proteome (referred to as Ref), or the combination of the two (referred to as combi, Figure 1B). In this comparison, all searches were run with open modification settings that allow for one mutation in the peptide match. The second comparison was between a regular and an open variant search using databases that did and did not include NA12878 genome sequencing-derived variants, respectively. This comparison was performed for the combi databases only. The analysis with the VF combi database will be referred to as the VF method and the analysis with the variant-containing combi database will be referred to as the VC method. In this comparison, open modification search was enabled for both methods, but open variant search was only enabled in the VF method to allow for the detection of SAAVs. Open variant search is disabled in the VC method because the variants were already incorporated in the VC search database.

Adding the Long-Read Transcriptome for the Cell Line Does Not Contribute to Additional Peptide Identifications in Practice

Reliable peptide identification normally requires a comprehensive search database. We first investigated whether novel transcripts from long-read transcriptome sequencing would contribute to peptide identifications in the NA12878 shotgun proteomics data. The ONT database contained 35,248 full-length transcript sequences, 64% of which were novel. Although the combi database containing these novel predicted ORFs was 42% larger than the Ref database (Figure 1B), the number of unique peptides from these sequences made up a mere 2.3% of the search database (Figure 2, top panel). The addition of ONT-derived ORFs to the Ref ORFs thus translated to an only modest increase in the number of unique peptides in the search database (Figure 2, lower panel). A likely explanation for this is the fact that many of the novel ONT transcripts demonstrate high similarity to existing reference sequences. The sequences usually only differed in the length of the 3′ or 5′ UTR or in the use of alternative exon junctions rather than completely novel exons. The exact frequencies of these events are difficult to estimate, but when looking at the set of novel ORFs from the ONT transcriptome, 73% of them can be attributed to known GENCODE coding genes. Conversely, the GENCODE genes that had novel isoforms in the ONT set corresponded to 27% of all GENCODE coding genes. In terms of observed peptide identifications, 67% of the ORFs in ONT set had at least one peptide match (when including PSMs that also matched to peptides present in GENCODE). However, the number of unique peptide matches to the novel ONT transcripts was much smaller: only 0.3% of unique peptides identified to the combi databases mapped exclusively to novel ONT transcripts. This indicates that the transcriptome database does not contribute significantly to the proteomic search results and suggests that alternative splicing and mRNA processing events do not contribute much to the diversity of the MS-detectable fraction of the proteome.

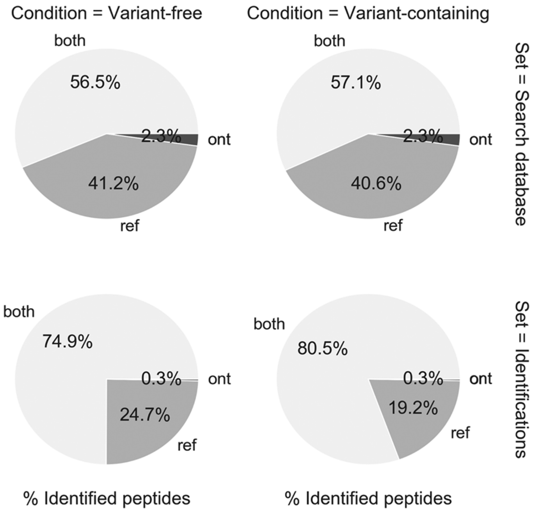
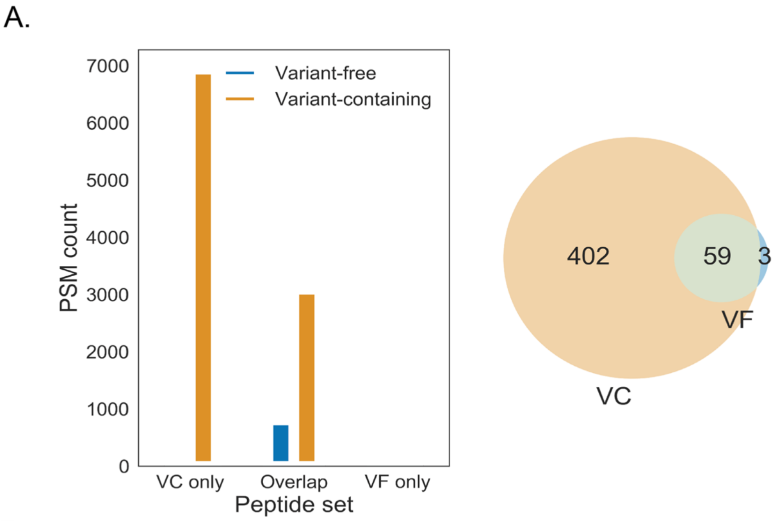


Figure 2. Detectable peptides per method. Theoretical (upper pie charts) and observed (lower pie charts) proportions of peptides when searching against VC (right) or VF (left) search databases. This shows percentages of matched peptides attributed only to GENCODE proteins, only ONT proteins, and those that match to proteins in both databases.

Aside from the contributions from the ONT-only sequences, it is also interesting to investigate protein identifications that were not found in the ONT transcriptome. While these should theoretically not be present, roughly 20% of identified peptides are exclusively matched with the ENCODE transcripts (Figure 2). As expected, this percentage is smaller than the 42% of peptides in the search database that are exclusive to GENCODE transcripts but still a significant fraction. This suggests that it is best to still use a reference transcript database, even if there is full transcriptome sequencing data available.

Variant-Containing Method Allows Detection of Many More Genome-Supported Variant Peptides

We subsequently studied the effect of the inclusion of sample-specific variants in the search database. In the VF method, the data is analyzed with an open variant search, thus letting the search engine predict single amino acid substitutions. This is in contrast to the VC method, where no variants are predicted and only genetically supported variants are present in the search database. We detected 461 variant peptides by the VC method and 62 by the VF method, with 59 overlapping between the two methods (Figure 3A). The greater majority of variant peptides that were detected by the VF method only (n = 1805) were not supported by the genome and are likely false positives (Figure 3B). In addition, one-third of the variant peptide matches that appeared to be supported by the genome actually contained an incorrect amino acid substitution. Thus, the inclusion of variant peptides derived from personal genomes in search databases is far superior to the use of a variant free database combined with an open variant search. Some examples of identified variant peptides can be found in Figure S3.



B.

	VC	VF genome-unsupported	VF genome-supported	VF genome-supported and correctly identified SAAV
PSM	10,015	41,340	807	592
Peptide	461	1,805	61	50

Figure 3. Detection of variant peptides using (combination) VF and VC databases. (A) Variant PSMs (left) and unique peptides (right) attributed to genome-supported variant peptides. (B) PSM and peptide counts found by each method.

Detectible Variant Peptides Have Attributes That Differ from Expected Variant Peptides

Out of the 34,968 peptides in the genome-supported variant peptide list, only 462 were detected by either or both the VC and VF methods (Figure 4A). They are not a random sample of all possible variant peptides. Namely, some variant peptides are easier to detect than others depending on their abundance and/or properties, and this differs even between methods. For instance, the VF method tends to find longer variant peptides (in a range of 16−27 aa) and misses the shorter variant peptides (Figure 4B). This highlights the larger amount of ambiguity in variant peptide identification proportional to the lower number of peaks in the spectra. The VC method does not suffer from this ambiguity and allows for detection of a wider range of variant peptide lengths than VF, especially shorter variant peptides (p = 0.0017 K2 samp). While there is a bias in variant peptide length, we did not find clear evidence that the position of the variant within the peptide affects detection of the variant peptide in either of the methods. In addition, the amino acid substitution itself affects detectability since the corresponding mass shift in the MS/MS spectrum needs to be separated from noise or similar mass shifts corresponding to other modifications in order to be identified. There are some predefined limitations to SAAV detection with the VF method that lead to certain amino acid substitutions getting detected less than expected (Figure 4C). Amino acids on which there are fixed modifications cannot have variants in the open variant search, meaning that substitutions at K and C are not detected. Substitutions affecting the trypsin digest, such as those involving R, can also not be detected.

Erroneous Variant Peptide Identifications Are Difficult to Discern from True Variant Peptide Identifications

The misidentifications from the open VF approach can be separated into false negatives and false positives. False-negative identification is where the VC method identifies variant peptides, but those same spectra are identified by the VF method as nonvariant peptides. False-positive misidentification is where the VF method identified variant peptides that were not supported by the genome.

There were 402 unique false-negative peptides observed (Figure 5A). These false-negative peptides were classified as variant peptides by the VC method but not by VF, although they were contained in the VF search space. Identifying causes of false negatives requires investigation of how the VC peptides were identified with the VF method. There was no particular length peptide that was misidentified more than others in general, despite the difference in detectible peptide length (Figure S4). The peptide identifications were similar between the VF and VC methods. In general, the length correlated highly between the identifications of the two methods (R^2 = 0.9071, p = 0). When comparing individual peptide identifications per method for mismatches and length difference, the largest source of error was a 1 aa length difference. Nonvariant peptides with a 1 aa length difference from the variant peptide were being identified instead of the correct variant peptide in >30% of the false negatives (Figure S4). Another possible source of false-negative errors that was investigated is SAAVs being mistaken for unexpected post-translational modifications. In the false-negative set, this did not appear to be an issue. The false-negative VF identifications had approximately the same rate of unexpected PTMs (Figure S4).

To further understand how false negatives could occur, we compared the peptide matching scores of the false-negative spectra for the VF and VC search methods (Figure 5B). Higher scores indicate higher confidence in assignment of spectra. VC scores for false-negative peptides were generally higher than the VF scores (mean score ratio VC/VF = 1.31). However, a large fraction of the false negatives received comparable scores in the

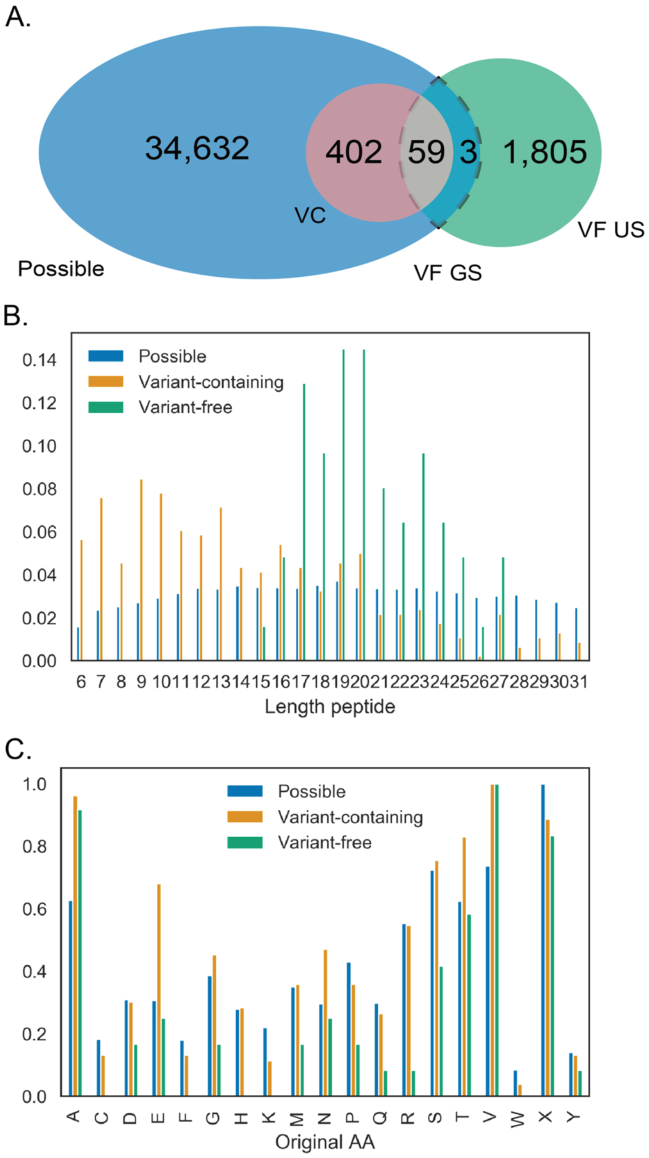


Figure 4. Properties of detected variants compared to those expected. (A) Groups of variant peptides being compared. All circles, including all overlaps, are being compared to each other. (B) Length distribution differences between detected variant peptides by the different variant detection methods. (C) Normalized (divided by max) frequency of variation per original (reference) amino acid.

VC and VF search methods. This could indicate a ranking problem: the variant peptide received a score equal to another peptide, to which the peptide spectrum was ultimately assigned. Delta retention time can often be a useful independent validator when the score disagrees between the different search methods. Despite high retention time discrepancies in this particular data set, observed retention time aligns relatively well with predicted retention time for those spectra that received higher scores in the VC.

The genome-supported variants are a tiny fraction of the high-confidence variant peptide predictions from the VF database, indicating a high false-positive rate (Figure 6A). We investigated whether there are distinguishing features between genome-supported and genome-unsupported variants. Reassuringly, scores of true positives were slightly higher than false positives [Figure 6B, $p = 1.34 \times 10^{-26}$, analysis of variance (ANOVA)].

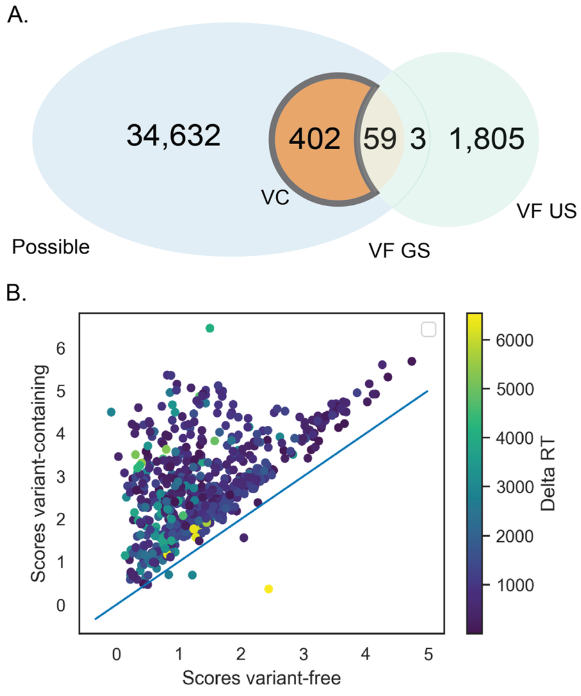


Figure 5. False-negative variant misidentifications. (A) Investigation of causes of mis-identification of peptides in the VF set. (B) Scores of those misidentified peptides in VF vs VC set. Each point corresponds to one false-negative variant peptide. Percolator PSM score is used. Color corresponds to delta retention time.

A closer inspection of genome-unsupported variants reveals potential sources of confusion for variant prediction algorithms, leading to false-positive identifications. There was a high level of concordance of peptides matched to these spectra in general. Two-thirds of spectra that corresponded to genome-unsupported variant peptide identifications by the VF had the same base peptide identifications in both the VF and VC searches. Mass shifts predicted to be SAAV in VF were commonly predicted to be "unexpected" PTMs by the VC method (Figure 6C). A common PTM mistaken as a SAAV in VF was threonine oxidation, but many PTMs contributed to this mix-up. There was no clear trend to the identification errors, underlining the difficulty of correctly classifying minor mass shifts corresponding to PTMs and SAAVs.

Evaluation of the Variant Peptides' SNPs of Origin

The detection of variant peptides is ultimately a means to understand which single-nucleotide variants (SNVs) are expressed at the protein level. By incorporating SNVs into predicted ORFs, we ended up with a theoretical set of 34,968 variant peptides originating from 9298 SNVs from all chromosomes, of which 5989 are heterozygous variants.

In the case of a heterozygous variant, both variant peptides and their reference counterparts can be identified in some ratio. A ratio different from 0.5 may be indicative of preferred expression of one of the alleles at the protein level, otherwise known as ASPE (allele-specific protein expression). The presence and magnitude of ASPE is potentially key information that can be used to understand biological mechanisms. However, technical biases of the search methodology may invalidate potential findings by distorting these ratios. For the VF method, the reference peptide was identified more frequently than the variant peptide ($p = 0.013$, one-way ANOVA). The opposite was true for the VC method.

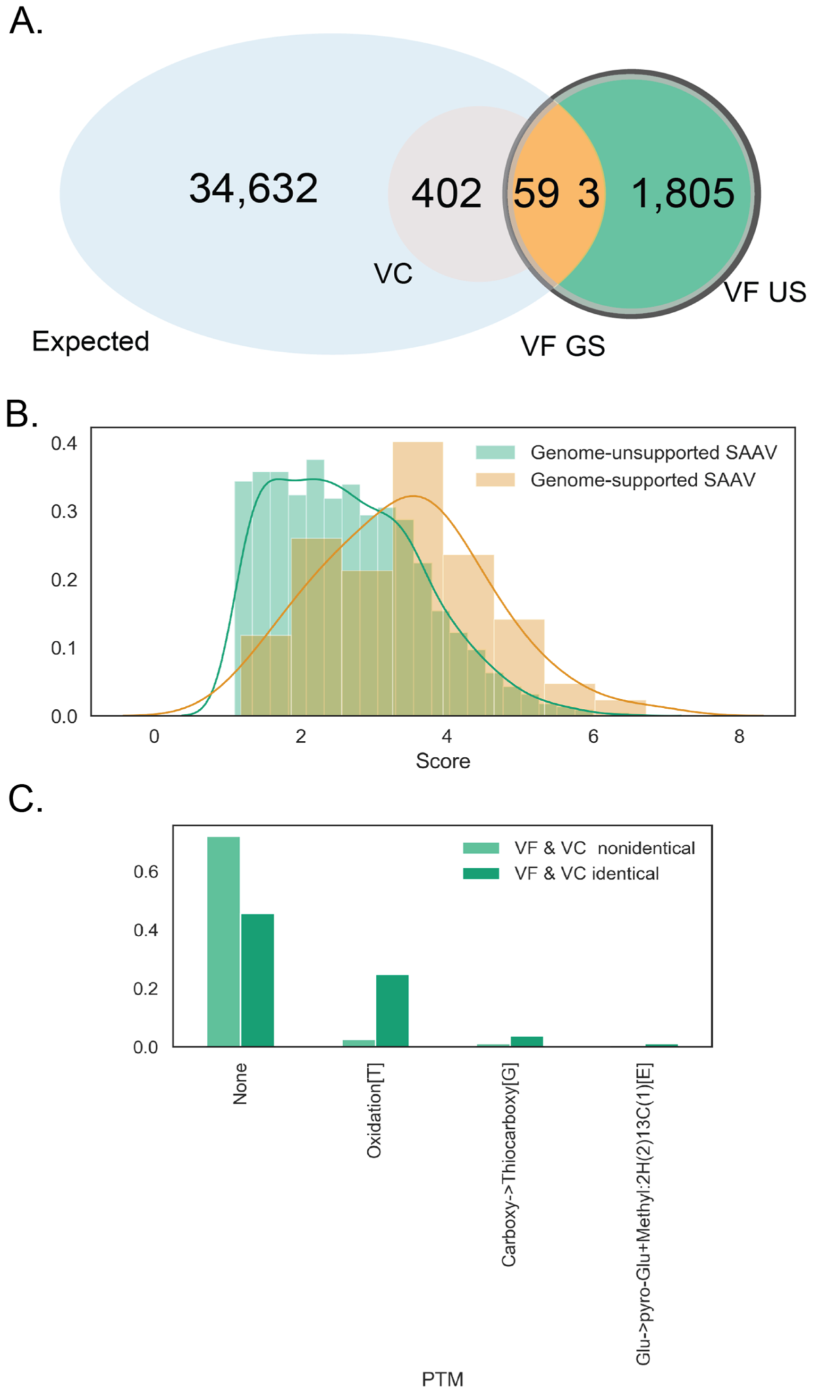


Figure 6. False-positive misidentifications. (A) False-positive misidentifications are genome-unsupported (US) variants predicted by the VF method. The Venn diagram highlights the subset of variants that are being investigated in this figure. These 2998 variants were predicted by *ionbot* to be variant peptides but were not found with the variant containing set. All but seven were variants unsupported by genome information. (B) Relative score distributions between genome supported vs unsupported variants in the VF set. (C) Unexpected modifications by the VC set corresponding to all "false-positive" predicted variant PSMs in the VF set.

Homozygous variants can be used as a type of control to understand the bias in search methods since we know that only one of the two alleles can be expressed. In case of homozygous variants, the variant peptide is expected to be present in all cases—with no reference counterpart. This was observed for the VC but not for the VF method (Figure 7A). Thus, without prior information about zygosity, the VF method tends to be conservative in identifying SAAV peptides, resulting in a higher likelihood of the reference peptide than its variant counterpart.

It is evident that some variant peptides were observed much more often than their reference counterparts or vice versa. The VC heterozygous variant peptide identifications should not suffer from the technical reference bias and allow for detection of allele-specific expression at the protein level. The VC-detected heterozygous variants were divided in two groups; one group with more counts for the reference peptide (reference-biased, $N = 78$) and one group with more counts for the alternative peptide (alternative-biased, $N = 123$). The two groups

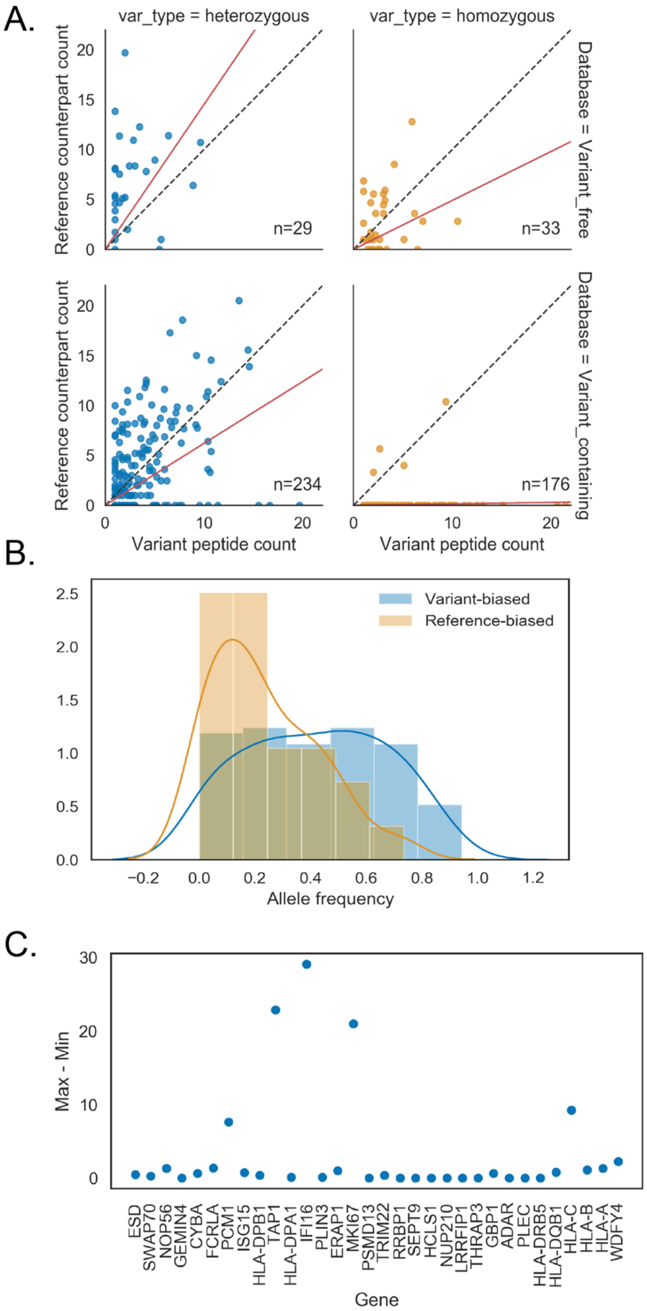


Figure 7. Underlying SNPs detected at the protein level. (A) Variant peptide abundance vs reference counterpart split by zygosity and search database, square root-transformed. (B) Separating heterozygous variants in the variant-containing database by whether more variant peptide was found (variant-biased) or more of the reference counterpart was found (reference-biased) revealed differences in allele frequency distributions. (C) Ratio variability of genes with two or more variant peptides. Ratio is defined by the variant counterpart abundance divided by variant peptide abundance. Y axis shows max − min per gene.

demonstrated a clear and significant difference in the population allele frequency ($p = 6.45 \times 10^{-8}$. Figure 7B). Those with lower allele frequencies displayed a stronger reference bias. This could be explained by the fact that rare variants in coding regions have a higher likelihood of causing undesirable effects on the resulting protein. Any deleterious effects resulting from the variant on protein stability would be visible as depletion of the alternative allele.

One significant subgroup of heterozygous variants was particularly biased toward the alternative allele. A total of 44 out of 183 variant peptides supported by more than two PSMs did not have any detected reference counterparts. One-third of these variants had a substitution involving arginine or lysine (tryptic cleavage sites). One gene, HLA-DBQ1, had two alternative alleles instead of one reference and one alternative. In general, the score distribution for these highly biased group was lower than the score distribution for all VC-detected variant peptides. The allele frequencies of this group were not different from those of the overall alternative-biased group ($p = 0.5$, ANOVA). There was also no correlation between the RNA expression of these genes to the variant peptide expression ($R^2 = 0.01$). Also, a comparison the list of genes displaying ASE on the RNA level from[42] to the heterozygous genes with variant peptides detected at the protein level yielded negligible overlap (two genes).

A total of 33 genes were detected through two or more unique variant peptides. For variant peptides within a gene, the reference-peptide-to-variant-peptide ratio should be consistent, unless there are different protein isoforms as a consequence of alternative splicing. This was the case for the majority of genes with multiple variant peptides belonging to the same gene (Figure 7C). Five of these genes were represented by multiple variant peptides with inconsistent ratios. HLA-C, IFI16, and MKI67 had peptides matching to nonidentical (sets of) isoforms within the gene. PCM1 had peptides matched to 24 isoforms, that is, four times the average number of isoforms matched by a variant peptide in the VC search. Thus, inconsistent variant-to-reference-peptide ratios within a gene can generally by attributed to differing abundances of protein isoforms.

■ DISCUSSION

Here, we have carried out an investigation of the effects of proteogenomic additions to a proteomics search database. To this end, we compared a typical proteomics approach to a purely proteomics method utilizing state-of-the-art open search. We observed that the addition of transcriptomic sequences to the search database did not have significant effects on the overall peptide identification rate. There was a roughly equal number PSMs from the three databases, despite the long-read transcriptome search database being 40% smaller than that of the union of it and the reference. At the same time, the matches to reference-only sequences in the combination database imply that >20% of peptide identifications are missed. This suggests a large portion of false identifications when using a database comprised only ONT sequences.

The fact that around a quarter of peptide identifications cannot be attributed to the transcriptomics data is rather surprising. There are a couple possible explanations. Using transcriptomics data from different cells than the proteomics data (different labs and different year) will unavoidably cause some discrepancies.[50] This could also be attributed to protein stability in the cell as proteins are detectable for some time after RNA has already been degraded.[51] Also notable is the fact that including the transcriptome sequences did not seem to add significantly to the peptide detections; the proportion of novel peptides found was lower than the proportion of novel transcripts found. As this cell line/organism is so well studied, it is likely that the vast majority of present proteins have already been characterized. For other cell types and organisms with more novel transcripts, adding (full length) transcriptomes may lead to more peptide identifications.

Two different search methods were used to identify nonreference peptides derived from SNVs: a proteogenomics approach, in which all variants known from the genome sequence were added to the search database, and an "open variant search", where only reference peptides were included in the search database and one amino acid differences were allowed by the search engine. The proteogenomics approach was clearly superior as it detected 7 times more variant peptides, whereas the open variant search suffered from many false-positive identifications that were not supported by the genome sequence and from large numbers of false negatives. Nevertheless, the proteogenomics search method also detected only a minor fraction of the variant peptides predicted to be present in the genome. It has been estimated before that maximum ~70% of variants in protein coding regions are theoretically detectible in an ideal shotgun proteomics experiment considering peptide lengths 7−40 aa.[52] The number of variants found with a proteogenomics method in practice is much lower, depending on method details. Some studies either use a statistically dubious "multitier" method[53,54] or skip FDR subsetting altogether[55] and report the number of variants detected to be in the region of 10%. We detect only 1% of the theoretically present variant peptides, despite the ~4M spectra present in this data set, making it one of the deepest proteomics datasets currently available. This is partly due to the careful control of FDRs in our study. Also, other conservative efforts to detect variant peptides using FDR subsetting or targeted proteomics validation detect <1% of all theoretically present variant peptides.[23,54,56]

While open search lags behind the proteogenomics approach for the moment, it has promise. Algorithms are being continuously improved to better differentiate signal from noise, which will reduce the false positives and false negatives in variant peptide detection.[57] There are several upcoming methodologies to further refine the open search to increase accuracy, either adding to existing peptide identification tools or standalone with promising results such as Open-pfind,[58] TagGraph,[39] MSFragger,[59] and Crystal-C.[60] There are considerable challenges still to face in their detection, particularly in noise/signal differentiation. This is especially complicated as variants often co-occur with other PTMs such as phosphorylation.[30,54] Current detection methods including *ionbot* cannot handle the complexity of two modifications on one site. However, deep neural networks show great promise with difficult peptide identifications.[61] Using methods of machine learning along with orthogonal information such as peptide retention time should result in significant improvements in open search.[62] This in combination with rapidly improving data-independent acquisition removes detection limitations of low-abundance or otherwise difficult to detect peptides,[63] which is currently a considerable hurdle in SAAV peptide detection.[55] Including open search is clearly useful and bound to get more accurate. This study used *ionbot* as the sole predictor of unexpected modifications/SAAVs, and comparison between identification tools was difficult as no other identification software tested reported the precise reporter ions per matched spectra (to be able to separate TMT tags corresponding to different cell lines). A study to compare methods given these updates is certainly warranted and ensemble methods may eventually be used to even more accurately predict these unexpected modifications/SAAVs.

One important implication of correctly detecting SAAVs is the ability to observe allele-specific expression at the protein level. A targeted proteomics approach has recently been described to study ASPE with high confidence.[64] It found no correlation between RNA and protein level ASE for the few variants studied, highlighting the utility of having higher throughput methods to study this phenomenon. One simple way to measure ASPE when using a proteogenomics approach is by comparing the spectral counts for the SAAV and its reference counterpart, since a reference counterpart usually has equal detectability by MS/MS.[52] Here, we found low correlation between the abundance of the variant and reference counterparts, regardless of VF or VC method. This is potentially indicative for a high level of ASPE. In contrast,[54] it demonstrated a high correlation between variant and reference peptides. This may be attributed to the low stringency associated with using the multitier search strategy for SAAV detection. We found no correlation between ASE and ASPE in this study, which is consistent with the findings of Shi et al.[64]

CONCLUSIONS

Our study provides guidance for the detection of variant peptides that shape the personal proteome. While personal genomes currently seem indispensable for the characterization of personal proteomes, new computational and analytical tools and new file formats to accommodate personal proteome information will allow us to get the fullest picture possible of the individual proteome, even without personal genome information.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00264.

Detailed workflow schematic, distribution of target and decoy variant peptides, annotated variant peptide spectra in mirror plots, investigation of false-negative ("mislabeled") identifications by *ionbot*, side-by-side comparison of the contents of the search database, and absolute numbers of PSMs and peptides detected per method (PDF)

Variant PSMs from the VF search database, variant PSMs from the variant-containing search database, reference counterpart PSMs from the VF search database, and reference counterpart PSMs from the variant-containing search database (ZIP)

AUTHOR INFORMATION

Corresponding Author

Peter A.C. 't Hoen − *Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen 6525 GA, The Netherlands*; Email: Peter-Bram.tHoen@radboudumc

Authors

Renee Salz − *Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen 6525 GA, The Netherlands*; orcid.org/0000-0003-1035-7866

Robbin Bouwmeester − *VIB-UGent Center for Medical Biotechnology VIB, 9052 Ghent, Belgium; Department of*

Biomolecular Medicine, Ghent University, 9052 Ghent, Belgium; orcid.org/0000-0001-6807-7029

Ralf Gabriels − *VIB-UGent Center for Medical Biotechnology VIB, 9052 Ghent, Belgium; Department of Biomolecular Medicine, Ghent University, 9052 Ghent, Belgium;* orcid.org/0000-0002-1679-1711

Sven Degroeve − *VIB-UGent Center for Medical Biotechnology VIB, 9052 Ghent, Belgium; Department of Biomolecular Medicine, Ghent University, 9052 Ghent, Belgium*

Lennart Martens − *VIB-UGent Center for Medical Biotechnology VIB, 9052 Ghent, Belgium; Department of Biomolecular Medicine, Ghent University, 9052 Ghent, Belgium*

Pieter-Jan Volders − *VIB-UGent Center for Medical Biotechnology VIB, 9052 Ghent, Belgium; Department of Biomolecular Medicine, Ghent University, 9052 Ghent, Belgium;* orcid.org/0000-0002-2685-2637

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.1c00264

Author Contributions

R.S. wrote the analysis scripts, performed the statistical analysis, and drafted the manuscript. L.M. and P.A.C.'t.H. conceived the study. P.-J.V. and P.A.C.'t.H. participated in its design and coordination and helped to draft the manuscript. S.D., R.G., and R.B. assisted with the running of the spectra identification tools. All authors read and approved the final manuscript. P.V. and P.A.C.'t.H. are shared last authors.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is funded by the Radboud Universitair Medisch Centrum.

■ REFERENCES

(1) Nagaraj, N.; Mann, M. Quantitative Analysis of the Intra-and Inter-Individual Variability of the Normal Urinary Proteome. *J. Proteome Res.* **2011**, *10*, 637−645.

(2) Kushner, I. K.; Clair, G.; Purvine, S. O.; Lee, J.-Y.; Adkins, J. N.; Payne, S. H. Individual Variability of Protein Expression in Human Tissues. *J. Proteome Res.* **2018**, *17*, 3914−3922.

(3) Li, J.; Su, Z.; Ma, Z. Q.; Slebos, R. J. C.; Halvey, P.; Tabb, D. L.; Liebler, D. C.; Pao, W.; Zhang, B. A Bioinformatics Workflow for Variant Peptide Detection in Shotgun Proteomics. *Mol. Cell. Proteomics* **2011**, *10*, M110.006536.

(4) Mertins, P.; Mani, D. R.; Mani, D. R.; Ruggles, K. V.; Gillette, M. A.; Clauser, K. R.; Wang, P.; Wang, X.; Qiao, J. W.; Cao, S.; Petralia, F.; Kawaler, E.; Mundt, F.; Krug, K.; Tu, Z.; Lei, J. T.; Gatza, M. L.; Wilkerson, M.; Perou, C. M.; Yellapantula, V.; Huang, K.-l.; Lin, C.; McLellan, M. D.; Yan, P.; Davies, S. R.; Townsend, R. R.; Skates, S. J.; Wang, J.; Zhang, B.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Ding, L.; Paulovich, A. G.; Fenyö, D.; Ellis, M. J.; Carr, S. A. Proteogenomics Connects Somatic Mutations to Signalling in Breast Cancer. *Nature* **2016**, *534*, 55−62.

(5) Subbannayya, Y.; Pinto, S. M.; Gowda, H.; Prasad, T. S. K. Proteogenomics for Understanding Oncology: Recent Advances and Future Prospects. *Expet Rev. Proteonomics* **2016**, *13*, 297−308.

(6) Zhang, B.; Wang, J.; Wang, J.; Wang, X.; Zhu, J.; Liu, Q.; Shi, Z.; Chambers, M. C.; Zimmerman, L. J.; Shaddox, K. F.; Kim, S.; Davies, S. R.; Wang, S.; Wang, P.; Kinsinger, C. R.; Rivers, R. C.; Rodriguez, H.; Townsend, R. R.; Ellis, M. J. C.; Carr, S. A.; Tabb, D. L.; Coffey, R. J.; Slebos, R. J. C.; Liebler, D. C.; Klauser, K. R.; Kuhn, E.; Mani, D. R.; Mertins, P.; Ketchum, K. A.; Paulovich, A. G.; Whiteaker, J. R.; Edwards, N. J.; McGarvey, P. B.; Madhavan, S.; Chan, D.; Pandey, A.; Shih, I. M.; Zhang, H.; Zhang, Z.; Zhu, H.; Whiteley, G. A.; Skates, S. J.; White, F. M.; Levine, D. A.; Boja, E. S.; Hiltke, T.; Mesri, M.; Shaw, K. M.; Stein, S. E.; Fenyo, D.; Liu, T.; McDermott, J. E.; Payne, S. H.; Rodland, K. D.; Smith, R. D.; Rudnick, P.; Snyder, M.; Zhao, Y.; Chen, X.; Ransohoff, D. F.; Hoofnagle, A. N.; Sanders, M. E.; Wang, Y.; Ding, L. Proteogenomic Characterization of Human Colon and Rectal Cancer. *Nature* **2014**, *513*, 382−387.

(7) Zhang, H.; Liu, T.; Zhang, Z.; Payne, S. H.; Zhang, B.; McDermott, J. E.; Zhou, J.-Y.; Petyuk, V. A.; Chen, L.; Ray, D.; Sun, S.; Yang, F.; Chen, L.; Wang, J.; Shah, P.; Cha, S. W.; Aiyetan, P.; Woo, S.; Tian, Y.; Gritsenko, M. A.; Clauss, T. R.; Choi, C.; Monroe, M. E.; Thomas, S.; Nie, S.; Wu, C.; Moore, R. J.; Yu, K.-H.; Tabb, D. L.; Fenyö, D.; Bafna, V.; Wang, Y.; Rodriguez, H.; Boja, E. S.; Hiltke, T.; Rivers, R. C.; Sokoll, L.; Zhu, H.; Shih, I.-M.; Cope, L.; Pandey, A.; Zhang, B.; Snyder, M. P.; Levine, D. A.; Smith, R. D.; Chan, D. W.; Rodland, K. D.; Carr, S. A.; Gillette, M. A.; Klauser, K. R.; Kuhn, E.; Mani, D. R.; Mertins, P.; Ketchum, K. A.; Thangudu, R.; Cai, S.; Oberti, M.; Paulovich, A. G.; Whiteaker, J. R.; Edwards, N. J.; McGarvey, P. B.; Madhavan, S.; Wang, P.; Chan, D. W.; Pandey, A.; Shih, I.-M.; Zhang, H.; Zhang, Z.; Zhu, H.; Cope, L.; Whiteley, G. A.; Skates, S. J.; White, F. M.; Levine, D. A.; Boja, E. S.; Kinsinger, C. R.; Hiltke, T.; Mesri, M.; Rivers, R. C.; Rodriguez, H.; Shaw, K. M.; Stein, S. E.; Fenyo, D.; Liu, T.; McDermott, J. E.; Payne, S. H.; Rodland, K. D.; Smith, R. D.; Rudnick, P.; Snyder, M.; Zhao, Y.; Chen, X.; Ransohoff, D. F.; Hoofnagle, A. N.; Liebler, D. C.; Sanders, M. E.; Shi, Z.; Slebos, R. J. C.; Tabb, D. L.; Zhang, B.; Zimmerman, L. J.; Wang, Y.; Davies, S. R.; Ding, L.; Ellis, M. J. C.; Townsend, R. R. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **2016**, *166*, 755−765.

(8) Tan, Z.; Zhu, J.; Stemmer, P. M.; Sun, L.; Yang, Z.; Schultz, K.; Gaffrey, M. J.; Cesnik, A. J.; Yi, X.; Hao, X.; Shortreed, M. R.; Shi, T.; Lubman, D. M. Comprehensive Detection of Single Amino Acid Variants and Evaluation of Their Deleterious Potential in a PANC-1 Cell Line. *J. Proteome Res.* **2020**, *19*, 1635−1646.

(9) Ruppen-Cañás, I.; López-Casas, P. P.; García, F.; Ximénez-Embún, P.; Muñoz, M.; Morelli, M. P.; Real, F. X.; Serna, A.; Hidalgo, M.; Ashman, K. An Improved Quantitative Mass Spectrometry Analysis of Tumor Specific Mutant Proteins at High Sensitivity. *Proteomics* **2012**, *12*, 1319−1327.

(10) Su, Z.-D.; Sun, L.; Yu, D.-X.; Li, R.-X.; Li, H.-X.; Yu, Z.-J.; Sheng, Q.-H.; Lin, X.; Zeng, R.; Wu, J.-R. Quantitative Detection of Single Amino Acid Polymorphisms by Targeted Proteomics. *J. Mol. Cell Biol.* **2011**, *3*, 309−315.

(11) Dimitrakopoulos, L.; Prassas, I.; Berns, E. M. J. J.; Foekens, J. A.; Diamandis, E. P.; Charames, G. S. Variant Peptide Detection Utilizing Mass Spectrometry: Laying the Foundations for Proteogenomic Identification and Validation. *Clin. Chem. Lab. Med.* **2017**, *55*, 1291−1304.

(12) Wang, D.; Eraslan, B.; Wieland, T.; Hallström, B.; Hopf, T.; Zolg, D. P.; Zecha, J.; Asplund, A.; Li, L.; Meng, C.; Frejno, M.; Schmidt, T.; Schnatbaum, K.; Wilhelm, M.; Ponten, F.; Uhlen, M.; Gagneur, J.; Hahne, H.; Kuster, B. A Deep Proteome and Transcriptome Abundance Atlas of 29 Healthy Human Tissues. *Mol. Syst. Biol.* **2019**, *15*, No. e8503.

(13) Guillot, L.; Delage, L.; Viari, A.; Vandenbrouck, Y.; Com, E.; Ritter, A.; Lavigne, R.; Marie, D.; Peterlongo, P.; Potin, P.; Pineau, C. Peptimapper: Proteogenomics Workflow for the Expert Annotation of Eukaryotic Genomes. *BMC Genom.* **2019**, *20*, 1.

(14) Wen, B.; Wang, X.; Zhang, B. PepQuery Enables Fast, Accurate, and Convenient Proteomic Validation of Novel Genomic Alterations. *Genome Res.* **2019**, *29*, 485−493.

(15) Flores, M. A.; Lazar, I. M. XMAn v2—a Database of Homo Sapiens Mutated Peptides. *Bioinformatics* **2020**, *36*, 1311.

(16) Cao, R.; Shi, Y.; Chen, S.; Ma, Y.; Chen, J.; Yang, J.; Chen, G.; Shi, T. DbSAP: Single Amino-Acid Polymorphism Database for Protein Variation Detection. *Nucleic Acids Res.* **2017**, *45*, D827−D832.

(17) Lichti, C. F.; Mostovenko, E.; Wadsworth, P. A.; Lynch, G. C.; Pettitt, B. M.; Sulman, E. P.; Wang, Q.; Lang, F. F.; Rezeli, M.; Marko-

Varga, G.; Végvári, Á.; Nilsson, C. L. Systematic Identification of Single Amino Acid Variants in Glioma Stem-Cell-Derived Chromosome 19 Proteins. *J. Proteome Res.* **2015**, *14*, 778−786.

(18) Nie, S.; Yin, H.; Tan, Z.; Anderson, M. A.; Ruffin, M. T.; Simeone, D. M.; Lubman, D. M. Quantitative Analysis of Single Amino Acid Variant Peptides Associated with Pancreatic Cancer in Serum by an Isobaric Labeling Quantitative Method. *J. Proteome Res.* **2014**, *13*, 6058−6066.

(19) Gatlin, C. L.; Eng, J. K.; Cross, S. T.; Detter, J. C.; Yates, J. R. Automated Identification of Amino Acid Sequence Variations in Proteins by HPLC/Microspray Tandem Mass Spectrometry. *Anal. Chem.* **2000**, *72*, 757.

(20) Roth, M. J.; Forbes, A. J.; Boyne, M. T.; Kim, Y.-B.; Robinson, D. E.; Kelleher, N. L. Precise and Parallel Characterization of Coding Polymorphisms, Alternative Splicing, and Modifications in Human Proteins by Mass Spectrometry. *Mol. Cell. Proteomics* **2005**, *4*, 1002−1008.

(21) Noble, W. S. Mass Spectrometrists Should Search Only for Peptides They Care About. *Nat. Methods* **2015**, *12*, 605−608.

(22) Nesvizhskii, A. I. A Survey of Computational Methods and Error Rate Estimation Procedures for Peptide and Protein Identification in Shotgun Proteomics. *J. Proteonomics* **2010**, *73*, 2092−2123.

(23) Song, C.; Wang, F.; Cheng, K.; Wei, X.; Bian, Y.; Wang, K.; Tan, Y.; Wang, H.; Ye, M.; Zou, H. Large-Scale Quantification of Single Amino-Acid Variations by a Variation-Associated Database Search Strategy. *J. Proteome Res.* **2014**, *13*, 241−248.

(24) Sheynkman, G. M.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Discovery and Mass Spectrometric Analysis of Novel Splice-Junction Peptides Using RNA-Seq*. *Mol. Cell. Proteomics* **2013**, *12*, 2341−2353.

(25) Wang, X.; Slebos, R. J. C.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B. Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data. *J. Proteome Res.* **2012**, *11*, 1009−1017.

(26) Wen, B.; Xu, S.; Zhou, R.; Zhang, B.; Wang, X.; Liu, X.; Xu, X.; Liu, S. P. G. A. An R/Bioconductor Package for Identification of Novel Peptides Using a Customized Database Derived from RNA-Seq. *BMC Bioinf.* **2016**, *17*, 244.

(27) Li, Y.; Wang, X.; Cho, J.-H.; Shaw, T. I.; Wu, Z.; Bai, B.; Wang, H.; Zhou, S.; Beach, T. G.; Wu, G.; Zhang, J.; Peng, J. JUMPg: An Integrative Proteogenomics Pipeline Identifying Unannotated Proteins in Human Brain and Cancer Cells. *J. Proteome Res.* **2016**, *15*, 2309−2320.

(28) Wang, X.; Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **2013**, *29*, 3235−3237.

(29) Zickmann, F.; Renard, B. Y. MSProGene: Integrative Proteogenomics beyond Six-Frames and Single Nucleotide Polymorphisms. *Bioinformatics* **2015**, *31*, i106−i115.

(30) Cesnik, A. J.; Miller, R. M.; Ibrahim, K.; Lu, L.; Millikin, R. J.; Shortreed, M. R.; Frey, B. L.; Smith, L. M. Spritz: A Proteogenomic Database Engine. **2020**, bioRxiv:10.1101/2020.06.08.140681.

(31) Sticker, A.; Martens, L.; Clement, L. Mass Spectrometrists Should Search for All Peptides, but Assess Only the Ones They Care About. *Nat. Methods* **2017**, *14*, 643−644.

(32) Wang, E. T.; Sandberg, R.; Luo, S.; Khrebtukova, I.; Zhang, L.; Mayr, C.; Kingsmore, S. F.; Schroth, G. P.; Burge, C. B. Alternative Isoform Regulation in Human Tissue Transcriptomes. *Nature* **2008**, *456*, 470−476.

(33) Blakeley, P.; Overton, I. M.; Hubbard, S. J. Addressing Statistical Biases in Nucleotide-Derived Protein Databases for Proteogenomic Search Strategies. *J. Proteome Res.* **2012**, *11*, 5221−5234.

(34) Tanner, S.; Shu, H.; Frank, A.; Wang, L.-C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. I. P. T. Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra. *Proc. Seventh Annu. Int. Conf. Comput. Mol. Biol.* **2003**, *422*, 4626−4639.

(35) Tabb, D. L.; Ma, Z.-Q.; Martin, D. B.; Ham, A.-J. L.; Chambers, M. C. DirecTag: Accurate Sequence Tags from Peptide MS/MS through Statistical Scoring. *J. Proteome Res.* **2008**, *7*, 3838−3846.

(36) Dasari, S.; Chambers, M. C.; Slebos, R. J.; Zimmerman, L. J.; Ham, A.-J. L.; Tabb, D. L. TagRecon: High-Throughput Mutation Identification through Sequence Tagging NIH Public Access. *J. Proteome Res.* **2010**, *9*, 1716−1726.

(37) Abraham, P.; Adams, R. M.; Tuskan, G. A.; Hettich, R. L. Moving Away from the Reference Genome: Evaluating a Peptide Sequencing Tagging Approach for Single Amino Acid Polymorphism Identifications in the Genus Populus. *J. Proteome Res.* **2013**, *12*, 3642−3651.

(38) Han, Y.; Ma, B.; Zhang, K. Spider: Software for Protein Identification from Sequence Tags with de Novo Sequencing Error. *J. Bioinf. Comput. Biol.* **2005**, *03*, 697−716.

(39) Devabhaktuni, A.; Lin, S.; Zhang, L.; Swaminathan, K.; Gonzalez, C. G.; Olsson, N.; Pearlman, S. M.; Rawson, K.; Elias, J. E. TagGraph Reveals Vast Protein Modification Landscapes from Large Tandem Mass Spectrometry Datasets. *Nat. Biotechnol.* **2019**, *37*, 469−479.

(40) Gabriels, R.; Martens, L.; Degroeve, S. Updated MS^{2}PIP Web Server Delivers Fast and Accurate MS^2 Peak Intensity Prediction for Multiple Fragmentation Methods, Instruments and Labeling Techniques. *Nucleic Acids Res.* **2019**, *47*, W295−W299.

(41) Silva, C.; Bouwmeester, R.; Martens, L.; Degroeve, S. Accurate Peptide Fragmentation Predictions Allow Data Driven Approaches to Replace and Improve upon Proteomics Search Engine Scoring Functions. *Bioinformatics* **2019**, *35*, 5243−5248.

(42) Workman, R. E.; Tang, A. D.; Tang, P. S.; Jain, M.; Tyson, J. R.; Razaghi, R.; Zuzarte, P. C.; Gilpatrick, T.; Payne, A.; Quick, J.; Sadowski, N.; Holmes, N.; de Jesus, J. G.; Soulette, C. M.; Snutch, T. P.; Loman, N.; Paten, B.; Loose, M.; Simpson, J. T.; Olsen, H. E.; Brooks, A. N.; Akeson, M.; Timp, W.; Timp, W. Nanopore Native RNA Sequencing of a Human Poly(A) Transcriptome. *Nat. Methods* **2019**, *16*, 1297−1305.

(43) Wu, L.; Candille, S. I.; Choi, Y.; Xie, D.; Jiang, L.; Li-Pook-Than, J.; Tang, H.; Snyder, M. Variation and Genetic Control of Protein Abundance in Humans. *Nature* **2013**, *499*, 79.

(44) Cox, J.; Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.

(45) Neph, S.; Kuehn, M. S.; Reynolds, A. P.; Haugen, E.; Thurman, R. E.; Johnson, A. K.; Rynes, E.; Maurano, M. T.; Vierstra, J.; Thomas, S.; Sandstrom, R.; Humbert, R.; Stamatoyannopoulos, J. A. BEDOPS: High-Performance Genomic Feature Operations. *Bioinformatics* **2012**, *28*, 1919−1920.

(46) Quinlan, A. R.; Hall, I. M. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* **2010**, *26*, 841−842.

(47) Adusumilli, R.; Mallick, P. Data Conversion with ProteoWizard MsConvert. *Methods in Molecular Biology*; Humana Press Inc., 2017; Vol. *1550*, pp 339−368.

(48) Goloborodko, A. A.; Levitsky, L. I.; Ivanov, M. V.; Gorshkov, M. V. Pyteomics - A Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics. *J. Am. Soc. Mass Spectrom.* **2013**, *24*, 301−304.

(49) Bouwmeester, R.; Gabriels, R.; Hulstaert, N.; Martens, L.; Degroeve, S. DeepLC Can Predict Retention Times for Peptides That Carry As-yet Unseen Modifications. **2020**, bioRxiv:10.1101/2020.03.28.013003.

(50) Hirsch, C.; Schildknecht, S. In Vitro Research Reproducibility: Keeping up High Standards. *Front. Pharmacol.* **2019**, *10*, 1484.

(51) Shao, W.; Guo, T.; Toussaint, N. C.; Xue, P.; Wagner, U.; Li, L.; Charmpi, K.; Zhu, Y.; Wu, J.; Buljan, M.; Sun, R.; Rutishauser, D.; Hermanns, T.; Fankhauser, C. D.; Poyet, C.; Ljubicic, J.; Rupp, N.; Rüschoff, J. H.; Zhong, Q.; Beyer, A.; Ji, J.; Collins, B. C.; Liu, Y.; Rätsch, G.; Wild, P. J.; Aebersold, R. Comparative Analysis of MRNA and Protein Degradation in Prostate Tissues Indicates High Stability of Proteins. *Nat. Commun.* **2019**, *10*, 2524.

(52) Mamie Lih, T.-S.; Choong, W.-K.; Chen, Y.-J.; Sung, T.-Y. Evaluating the Possibility of Detecting Variants in Shotgun Proteomics via LeTE-Fusion Analysis Pipeline. *J. Proteome Res.* **2018**, *17*, 2937−2952.

(53) Hwang, H.; Park, G. W.; Park, J. Y.; Lee, H. K.; Lee, J. Y.; Jeong, J. E.; Park, S.-K. R.; Yates, J. R.; Kwon, K.-H.; Park, Y. M.; Lee, H.-J.; Paik, Y.-K.; Kim, J. Y.; Yoo, J. S. Next Generation Proteomic Pipeline for Chromosome-Based Proteomic Research Using NeXtProt and GENCODE Databases. *J. Proteome Res.* **2017**, *16*, 4425−4434.

(54) Ma, S.; Menon, R.; Poulos, R. C.; Wong, J. W. H. Proteogenomic Analysis Prioritises Functional Single Nucleotide Variants in Cancer Samples. *Oncotarget* **2017**, *8*, 95841−95852.

(55) Ruggles, K. V.; Tang, Z.; Wang, X.; Grover, H.; Askenazi, M.; Teubl, J.; Cao, S.; McLellan, M. D.; Clauser, K. R.; Tabb, D. L.; Mertins, P.; Slebos, R.; Erdmann-Gilmore, P.; Li, S.; Gunawardena, H. P.; Xie, L.; Liu, T.; Zhou, J.-Y.; Sun, S.; Hoadley, K. A.; Perou, C. M.; Chen, X.; Davies, S. R.; Maher, C. A.; Kinsinger, C. R.; Rodland, K. D.; Zhang, H.; Zhang, Z.; Ding, L.; Townsend, R. R.; Rodriguez, H.; Chan, D.; Smith, R. D.; Liebler, D. C.; Carr, S. A.; Payne, S.; Ellis, M. J.; Fenyő, D. An Analysis of the Sensitivity of Proteogenomic Mapping of Somatic Mutations and Novel Splicing Events in Cancer. *Mol. Cell. Proteomics* **2016**, *15*, 1060−1071.

(56) Dimitrakopoulos, L.; Prassas, I.; Sieuwerts, A. M.; Diamandis, E. P.; Martens, J. W. M.; Charames, G. S. Proteome-Wide Onco-Proteogenomic Somatic Variant Identification in ER-Positive Breast Cancer. *Clin. Biochem.* **2019**, *66*, 63−75.

(57) Bittremieux, W.; Meysman, P.; Noble, W. S.; Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *J. Proteome Res.* **2018**, *17*, 3463−3474.

(58) Chi, H.; Liu, C.; Yang, H.; Zeng, W.-F.; Wu, L.; Zhou, W.-J.; Wang, R.-M.; Niu, X.-N.; Ding, Y.-H.; Zhang, Y.; Wang, Z.-W.; Chen, Z.-L.; Sun, R.-X.; Liu, T.; Tan, G.-M.; Dong, M.-Q.; Xu, P.; Zhang, P.-H.; He, S.-M. Comprehensive Identification of Peptides in Tandem Mass Spectra Using an Efficient Open Search Engine. *Nat. Biotechnol.* **2018**, *36*, 1059−1061.

(59) Yu, F.; Teo, G. C.; Kong, A. T.; Haynes, S. E.; Avtonomov, D. M.; Geiszler, D. J.; Nesvizhskii, A. I. Identification of Modified Peptides Using Localization-Aware Open Search. *Nat. Commun.* **2020**, *11*, 4065.

(60) Chang, H.-Y.; Kong, A. T.; Da Veiga Leprevost, F.; Avtonomov, D. M.; Haynes, S. E.; Nesvizhskii, A. I.; Nesvizhskii, A. I. Crystal-C: A Computational Tool for Refinement of Open Search Results. *J. Proteome Res.* **2020**, *19*, 2511−2515.

(61) Bouwmeester, R.; Gabriels, R.; Van Den Bossche, T.; Martens, L.; Degroeve, S. The Age of Data-Driven Proteomics: How Machine Learning Enables Novel Workflows. *Proteomics* **2020**, *20*, 1900351.

(62) Tran, N. H.; Zhang, X.; Xin, L.; Shan, B.; Li, M. De Novo Peptide Sequencing by Deep Learning. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, 8247−8252.

(63) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M. Deep Learning Enables de Novo Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry. *Nat. Methods* **2019**, *16*, 63−66.

(64) Shi, J.; Wang, X.; Zhu, H.; Jiang, H.; Wang, D.; Nesvizhskii, A.; Zhu, H.-J. Determining Allele-Specific Protein Expression (ASPE) Using a Novel Quantitative Concatamer Based Proteomics Method. *J. Proteome Res.* **2018**, *17*, 3606−3612.