OXFORD
(GIGA)^n SCIENCE

GigaScience, 2023, **12**, 1–13
DOI: 10.1093/gigascience/giad093
Research

Finding haplotypic signatures in proteins

Jakub Vašíček [1,2,†], Dafni Skiadopoulou [1,2,†], Ksenia G. Kuznetsova [1,2], Bo Wen [3], Stefan Johansson [1,4], Pål R. Njølstad [1,5], Stefan Bruckner [6,‡], Lukas Käll [7,‡] and Marc Vaudel [1,2,8,*,‡]

[1] Mohn Center for Diabetes Precision Medicine, Department of Clinical Science, University of Bergen, Bergen 5021, Norway
[2] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen 5008, Norway
[3] Department of Genome Sciences, University of Washington, Seattle, WA 98195, United States
[4] Department of Medical Genetics, Haukeland University Hospital, Bergen 5021, Norway
[5] Children and Youth Clinic, Haukeland University Hospital, Bergen 5021, Norway
[6] Chair of Visual Analytics, Institute for Visual and Analytic Computing, University of Rostock, Rostock 18051, Germany
[7] Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH–Royal Institute of Technology, Solna 17121, Sweden
[8] Department of Genetics and Bioinformatics, Health Data and Digitalization, Norwegian Institute of Public Health, Oslo 0473, Norway
*Correspondence address. Marc Vaudel, Universitetet i Bergen, Klinisk institutt 2, Postboks 7804, NO-5020 BERGEN, NORWAY. E-mail: Marc.Vaudel@uib.no
†These authors contributed to the work equally.
‡These authors jointly supervised the work.

Abstract

Background: The nonrandom distribution of alleles of common genomic variants produces haplotypes, which are fundamental in medical and population genetic studies. Consequently, protein-coding genes with different co-occurring sets of alleles can encode different amino acid sequences: protein haplotypes. These protein haplotypes are present in biological samples and detectable by mass spectrometry, but they are not accounted for in proteomic searches. Consequently, the impact of haplotypic variation on the results of proteomic searches and the discoverability of peptides specific to haplotypes remain unknown.

Findings: Here, we study how common genetic haplotypes influence the proteomic search space and investigate the possibility to match peptides containing multiple amino acid substitutions to a publicly available data set of mass spectra. We found that for 12.42% of the discoverable amino acid substitutions encoded by common haplotypes, 2 or more substitutions may co-occur in the same peptide after tryptic digestion of the protein haplotypes. We identified 352 spectra that matched to such multivariant peptides, and out of the 4,582 amino acid substitutions identified, 6.37% were covered by multivariant peptides. However, the evaluation of the reliability of these matches remains challenging, suggesting that refined error rate estimation procedures are needed for such complex proteomic searches.

Conclusions: As these procedures become available and the ability to analyze protein haplotypes increases, we anticipate that proteomics will provide new information on the consequences of common variation, across tissues and time.

Keywords: proteogenomics, haplotype, protein, bioinformatics, post-translational modification

Background

Linkage disequilibrium (LD) describes the nonrandom correlation between alleles at different positions in the genome in a population. LD arises when alleles at nearby sites co-occur on the same haplotype more often than expected by chance. When haplotypes are located in protein-coding portions of the genome and include nonsynonymous changes, they can alter protein sequences, forming so-called protein haplotypes, as defined by Spooner et al. [1]. Based on the co-occurrence of alleles in the 1000 Genomes Project [2] and their *in silico* translation, Spooner et al. [1] created a list of possible protein haplotype sequences. Notably, they stress that for 1 in 7 genes, the most frequent protein haplotype differs from the reference sequence in Ensembl [3]. In precision medicine, probing the proteotype—the actual state of the proteome—adds valuable information concerning the relationship between the genotype and the phenotype [4]. Therefore, it is important that genetic information, including LD, is taken into account in proteomics searches.

Proteins in biological samples can be identified by liquid chromatography coupled to mass spectrometry (LC-MS), usually after digestion into peptides [5]. Then, the measured spectra are matched to a database of expected protein sequences using a search engine [6]. The identified peptides are used to infer the presence of proteins [7] along with potential posttranslational modifications (PTMs) [8]. When the peptides cover the relevant parts of the protein sequences, it is also possible to discover the product of alternative splicing or genetic variation [9]. In precision medicine, proteomic searches need to be adapted to individual patient profiles by extending the search space to include noncanonical sequences [10].

This challenge is addressed by proteogenomics—the scientific field integrating genomics and proteomics into a joint approach [9, 11]. Recent work, mainly in the domain of cancer research, has shown that accounting for genetic variation in proteomic analyses provides the means to discover noncanonical proteins. Umer et al. [12] have developed a tool to generate databases of variant proteins derived from single-nucleotide polymorphisms (SNPs), insertions and deletions, and the 3-frame translation of pseudogenes and noncanonical transcripts, appended with a database of canonical proteins [12]. Levitsky et al. [13] use measures of

Received: April 13, 2023. **Revised:** September 24, 2023. **Accepted:** October 8, 2023

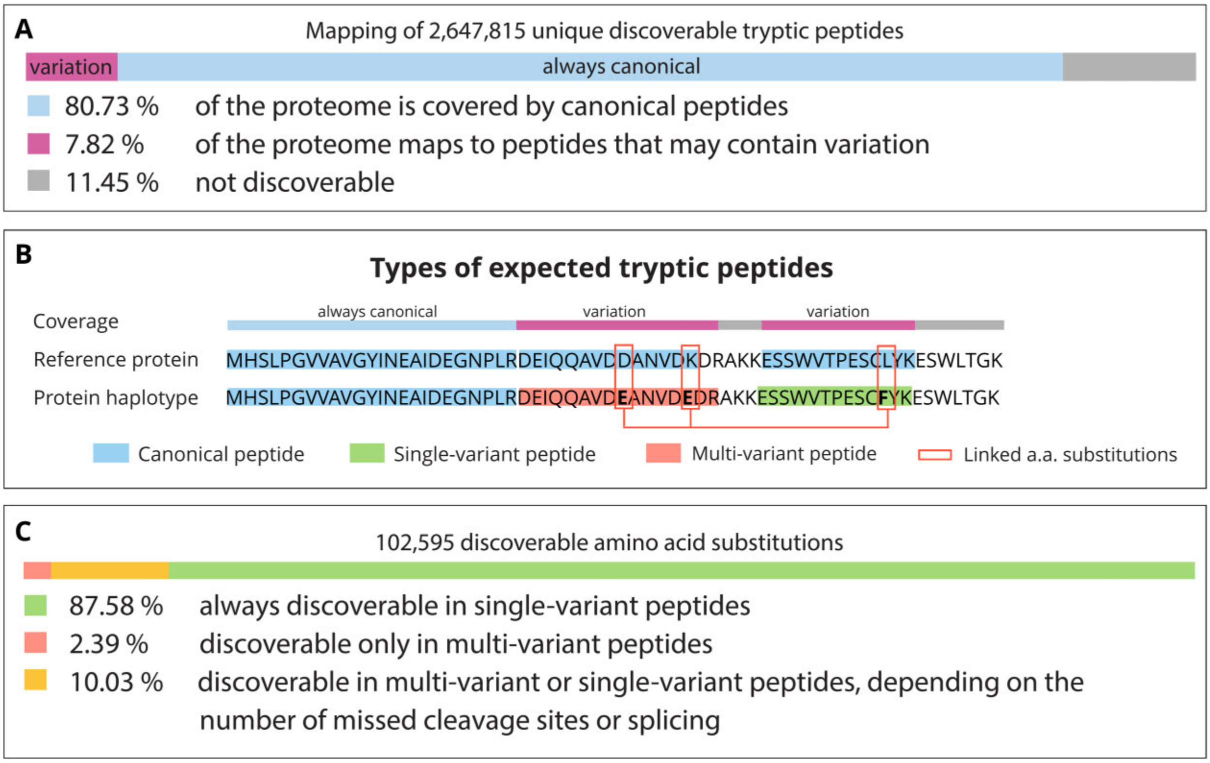
A
Mapping of 2,647,815 unique discoverable tryptic peptides
variation
always canonical
80.73 % of the proteome is covered by canonical peptides
7.82 % of the proteome maps to peptides that may contain variation
11.45 % not discoverable
B
Types of expected tryptic peptides
Coverage
always canonical
variation
variation
Reference protein
MHSLPGVVAVGYINEAIDEGNPLRDEIQQAVDDANVDKDRAKKESSWVTPESCLYKESWLTGK
Protein haplotype
MHSLPGVVAVGYINEAIDEGNPLRDEIQQAVDEANVDEDRAKKESSWVTPESCFYKESWLTGK
Canonical peptide
Single-variant peptide
Multi-variant peptide
Linked a.a. substitutions
C
102,595 discoverable amino acid substitutions
87.58 % always discoverable in single-variant peptides
2.39 % discoverable only in multi-variant peptides
10.03 % discoverable in multi-variant or single-variant peptides, depending on the number of missed cleavage sites or splicing

Figure 1: (A) Proteome coverage expressed in terms of the percentage of amino acids; that is, if 7 out of 100 residues belong to at least 1 discoverable peptide containing the product of a substitution, we say that 7% of the proteome maps to peptides containing variation. See main text for details and Materials and Methods for the handling of shared peptides. (B) Example of a reference sequence aligned to another haplotype. The classes of peptides following the cleavage pattern of trypsin are highlighted by a colored background. Three amino acid substitutions encoded by this haplotype are marked by red rectangles. The "coverage" layer indicates the alignment applied to obtain numbers shown in section A. (C) Distribution of variation in discoverable peptides. Amino acid variants are stratified based on the category of peptide in which the substitution caused by the respective variant can be identified.

proteome coverage, including variant peptides, to verify the presence of single amino acid variants. Choong et al. [14] proposed an algorithm to generate the optimal number of protein sequences containing combinations of amino acid substitutions possibly occurring in the same tryptic peptide. In their approach, the database includes not only the combinations of alleles encoded by haplotypes but all combinations possible per peptide. Lobas et al. [15, 16] showed that peptides containing variation were 2.5 to 3 times less likely to be identified than canonical peptides. Wang et al. [17] have analyzed data for 29 paired healthy human tissues from the Human Proteome Atlas project to detect amino acid variants at the protein level. However, the majority of amino acid variants predicted from exome sequencing could not be detected [17], suggesting that proteogenomics remains highly challenging and methods for discovering noncanonical proteins need further development.

Here, we used the protein haplotypes generated by Spooner et al. [1] to evaluate the ability of mass spectrometry–based proteomics to identify peptides encoded by combinations of variants in LD. We show that in some protein haplotypes, multiple amino acid substitutions affect the same peptide after digestion. Those protein haplotypes can only be identified if the combinations of amino acid variants are included in the search space, and several of these protein haplotypes are predicted to be more common than the reference sequence. Then, we mined the publicly available data from Wang et al. [17] for peptides including a combination of amino acid variants, demonstrating how such peptides can be identified according to the standards of the field but also how the quality control of the results remains challenging.

Results

The consequence of haplotypes on the proteomics search space

We digested *in silico* the protein sequences translated from haplotypes obtained from Spooner et al. [1] using the canonical cleavage pattern of trypsin, allowing for up to 2 missed cleavages. Note that indels were not considered, and we focused only on common variants with a minor allele frequency >1% in any population of the 1000 Genomes Project [2]; see Materials and Methods for details. After excluding contaminants, this yielded 2,647,815 unique tryptic peptide sequences of length between 8 and 40 amino acids (Fig. 1A). The coverage of protein sequences from Ensembl can be partitioned as follows: 80.73% can only be covered by canonical peptides, 7.82% map to peptides that may contain 1 or multiple amino acid substitutions, and the remaining yields sequences that are either too short or too long to be identified. Most peptides discoverable in proteomic studies therefore map to canonical sequences, making it challenging for nontargeted approaches to assess the allelic status of a common genetic variant using proteomics, in agreement with [15, 16].

We classify the obtained peptide sequences in 3 types (Fig. 1B): (i) canonical, no haplotype is known to yield an amino acid substitution in the sequence of this peptide; (ii) single-variant, a haplotype encodes an amino acid substitution in the sequence of this peptide; and (iii) multivariant, a haplotype encodes a set of 2 or more amino acid substitutions in the sequence of this peptide. In total, common haplotypes encode 102,595 amino acid substitutions, with 87.58% of them found only in single-variant

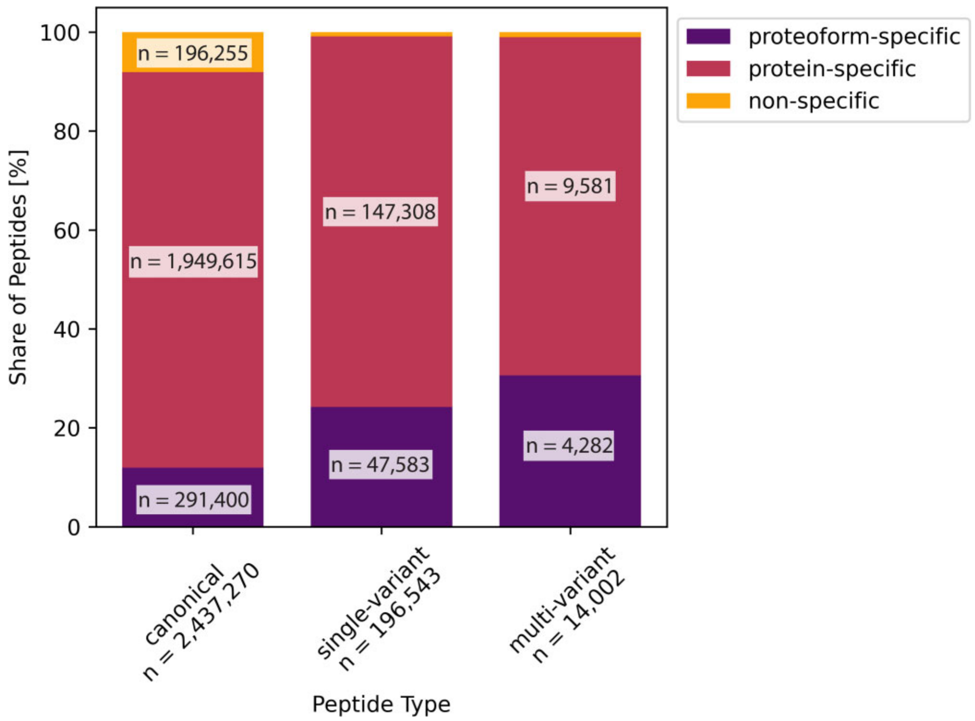

Figure 2: Classification of peptides based on their ability to distinguish between protein sequences (bar color) and to identify amino acid substitutions (position on x-axis). The height of the bars represents the distribution of categories (nonspecific, protein-specific, proteoform-specific) among the peptide types (canonical, single-variant, multivariant).

peptides, 2.39% in multivariant peptides, and 10.03% in either single- or multivariant peptides depending on the number of missed cleavages (Fig. 1C). Note that substitutions in different isoforms of the same protein are reported separately by Spooner et al. [1], creating multiple consequences for the same genetic variant. The total number of amino acid substitutions is consequently higher than the number of genetic variants. Interestingly, based on the frequencies among all participants in the 1000 Genomes project, 22.3% and 32.4% of the amino acid substitutions discoverable in single-variant and multivariant peptides, respectively, occur in protein haplotypes that are predicted to be more frequent than the Ensembl reference sequence. If these alleles are not accounted for, proteomics analyses will, therefore, not be able to identify these parts of the genome for the majority of individuals.

Peptides can be classified based on their ability to distinguish between protein sequences. We propose the following categories: (i) nonspecific peptides map to the products of different genes; (ii) protein-specific peptides map to multiple sequences, which are all products of the same gene; and (iii) proteoform-specific peptides map uniquely to a single form of a protein (i.e., single splice variant and haplotype), referred to as proteoform [18]. In this classification, based on the identification of a proteoform-specific peptide, one can uniquely identify products of a given gene. A protein-specific peptide allows for discriminating certain groups of proteoforms but does not yield a single candidate sequence (e.g., it determines which amino acid substitution is present but maps to multiple splicing variants). Nonspecific peptide sequences map to multiple genes, where the sequence of 1 gene matches the sequence of another, making it challenging to infer which protein is covered. We found 198,046 distinct nonspecific peptide sequences, covering up to 17.53% of the proteome. The prevalence of canonical, single-variant, and multivariant peptides among the above introduced types is displayed in Fig. 2, with exact numbers provided in Table 1. As expected intuitively, peptides containing the product of 1 or multiple variants present a higher ability to distinguish between protein products of different genes and between proteoforms of the same gene.

Table 1: Classification of peptides types and the number of *in silico* digested peptides in each of the categories

Peptide type (variation)	Peptide type (specificity)	Number of possible peptides
Canonical	Proteoform-specific	291,400
Canonical	Protein-specific	1,949,615
Canonical	Nonspecific	196,255
Single-variant	Proteoform-specific	47,583
Single-variant	Protein-specific	147,308
Single-variant	Nonspecific	1,652
Multivariant	Proteoform-specific	4,282
Multivariant	Protein-specific	9,581
Multivariant	Nonspecific	139

Matching multivariant peptides to mass spectra

To investigate the prevalence of spectra matching multivariant peptides encoded by common haplotypes and the quality of the obtained matches, we searched the deep proteomics data of healthy tonsil tissue made available by Wang et al. [17] against the sequences of common protein haplotypes using X!Tandem [19] as a search engine without refinement procedure and Percolator [20] with standard features for the evaluation of the confidence in all peptide-to-spectrum matches (PSMs). The resulting PSMs were thresholded at a 1% PSM-level false discovery rate (FDR). Note that because our study focuses on evaluating the quality of the spectrum matches, a PSM-level FDR was therefore preferred to peptide-level statistics. After thresholding, 1,318,152 target PSMs remained (13,467 decoy PSMs would have passed the threshold), representing 176,193 unique peptide sequences (8,047 decoy peptide sequences would have passed the threshold), covering the alternative amino acid of 4,582 substitutions. The distribution of alternative alleles among single- and multivariant peptides (Fig. 3A) mirrored the values obtained from the *in silico* digestion of protein haplotypes (Fig. 1C). On average, the products of 2,249.67 substitutions were found per sample (2,360, 2,165, and 2,224 in samples 1, 2, and 3, respectively). The matched peptide sequences cover 21.56% of the proteome predicted to map exclusively to canonical peptides and 16.89% of the proteome possibly mapping to peptides with substitutions (Fig. 3B). Note, however, that 19,678 peptide sequences (identified in 231,181 PSMs) map to the products of multiple genes that cannot be distinguished, hence affecting the coverage estimates.

Out of the 1,318,152 spectra matched to peptides, 0.57% were matched to single-variant peptides and 0.03% were matched to multivariant peptides. The share of spectra matched to variant peptides is thus lower than the expected error rate, and currently, no method allows the evaluation of error rates in these subgroups of matches specifically. We thus investigated whether these classes of peptides showed signs of an overrepresentation of false-positive matches. No substantial difference was noticeable in the density of the posterior error probabilities (PEPs) and *q*-values for all 3 classes of PSMs (Fig. 3C, D), indicating that a more stringent FDR threshold would not alter the prevalence of variant peptides. We also compared the observed peptide retention time and fragmentation with predictions from DeepLC [21] and MS2PIP [22], respectively. Overall, the density of the distance to prediction in both retention time and fragmentation was very similar for all 3 classes of peptides (Fig. 3E, F), displaying no obvious shift in the distribution, which would have been indicative of a strong overrepresentation of false positives. Yet the distributions of variant and multivariant peptides showed stronger tails toward high distance to prediction compared to nonvariant peptides, indicative of the presence of false-positive matches. In comparison, the distance to prediction for decoys showed high retention time difference and low spectrum similarity.

Quality metrics on all matches are available as supplementary material. Three examples, sampled from the multivariant matches passing the FDR threshold at low, medium, and high PEP, representing high, medium, and low confidence, respectively, are displayed in Fig. 4 along with the predicted spectra. As expected, the share of peaks matching predicted fragment ions decreases as the PEP increases: (A) the highly confident match presents an excellent coverage of the spectrum with fragment ion masses, with an extensive mapping of the peptide y-ion series; the retention time distance to prediction of 320.8 seconds represents only a fraction of the gradient (approximately 2.5 hours); and the spectrum similarity to prediction, 0.79, shows good but not perfect agreement, which is in the lower range of the distribution of similarity scores for the canonical matches. (B) The medium confidence match presents a good coverage of the spectrum lacking prediction for many peaks, and the agreement scores with retention time and fragmentation predictions are excellent. (C) The low confidence match presents a poor coverage of the spectrum with poor agreement with retention time and fragmentation predictions. In addition to passing commonly accepted statistical thresholds, the matches in Fig. 4A and B would pass expert quality control. On the other hand, the match in 4C is most probably a false positive. Together, while these 3 sampled PSMs represent only a limited set of examples, they are very representative of the difficulty to confidently assess the presence of individual peptides from large proteomic experiments. This task is, however, important given that chimeric spectrum matches [23–26] and partial matches are known to be difficult to account for in error rate estimation [27, 28].

As highlighted by Spooner et al. [1], depending on the population studied, specific haplotypes often have higher frequencies than the canonical haplotype by Ensembl. For example, there are 5 haplotypes of the IQ motif containing the GTPase activating protein 2 (IQGAP2, ENSP00000274364) gene that have higher predicted frequencies than the canonical haplotype in the European population (with combined frequency of 84.9% according to Spooner et al. [1]). These haplotypes encode a tryptic peptide containing 2 amino acid substitutions when compared to the canonical sequence in Ensembl: VLWLDEIQQAVD**E**ANVD**E**DR (amino acid substitutions in bold). At position 527 of the protein sequence, aspartic acid is changed to glutamic acid (527D>E, rs2431352), and at position 532, lysine is changed to glutamic acid (532K>E, rs2909888), preventing cleavage by trypsin. In our results, 2 peptides overlapped with this sequence, 1 featuring a missed cleavage, supported by 13 and 10 spectra, respectively. Fig. 4D and E display 2 examples of highly confident matches, and Supplementary Table S1 lists PEP, *q*-value, and agreement with predictors for all PSMs. Altogether, the PEPs and agreement with predictors for these PSMs support the identification of this sequence and thus the presence of these haplotypes in the data reported by Wang et al. [17], consistent with the frequencies of these haplotypes in the European population. The sequence encoded by these haplotypes cannot be detected using canonical databases.

For diploid chromosomes, in the absence of deletion or copy number alteration, each individual carries 2 versions of a given gene—1 paternally and 1 maternally inherited—which can represent different haplotypes. We thus expect to find evidence for heterozygosity in some of the identified variants. We have come across such cases in 26, 21, and 19 genes in samples 1, 2, and 3, respectively. For example, the protein CR1 (complement component [3b/4b] receptor 1, ENSP00000356016) is commonly affected by multiple SNPs. First, at the position 2060, threonine is commonly changed to serine (2060T>S, rs4844609). Haplotypes including serine at position 2060 are expected in the European population with the combined frequency of 98%. Second, at the position 2065, isoleucine can be changed to valine (2065I>V, rs6691117); valine is expected in the European population with a frequency of 22.57%. However, a valine at position 2065 is only expected when preceded by a serine at position 2060. In one of the samples, we identified spectra matching confidently to both a multivariant peptide encoded by both alternative alleles (SFF**S**LTEI**V**R, substitutions in bold) and to a single-variant peptide encoded by the alternative allele of the first SNP (SFF**S**LTEIIR, substitution in bold). Mirrored spectra and associated quality metrics are shown in Fig. 5. In this case, including haplotypes in the protein database enables the identification of not only the alternative but also the reference allele of a variant.

While including the sequences from different haplotypes offers the ability to detect new protein haplotypes, it also increases the likelihood of similar peptides to map to different proteins. For example, the protein POTE ankyrin domain family member I (POTEI, ENSP00000392718) contains in its most frequent haplotype 8 amino acid substitutions, 2 of which fall into the same tryptic peptide. In the actin-like domain of POTEI at position 918, tyrosine changes to phenylalanine (918Y>F, rs147268452), and at position 929, methionine changes into threonine (929M>T, rs201878083), thus encoding the peptide LCFVALDFEQEMATAASSSSLEK

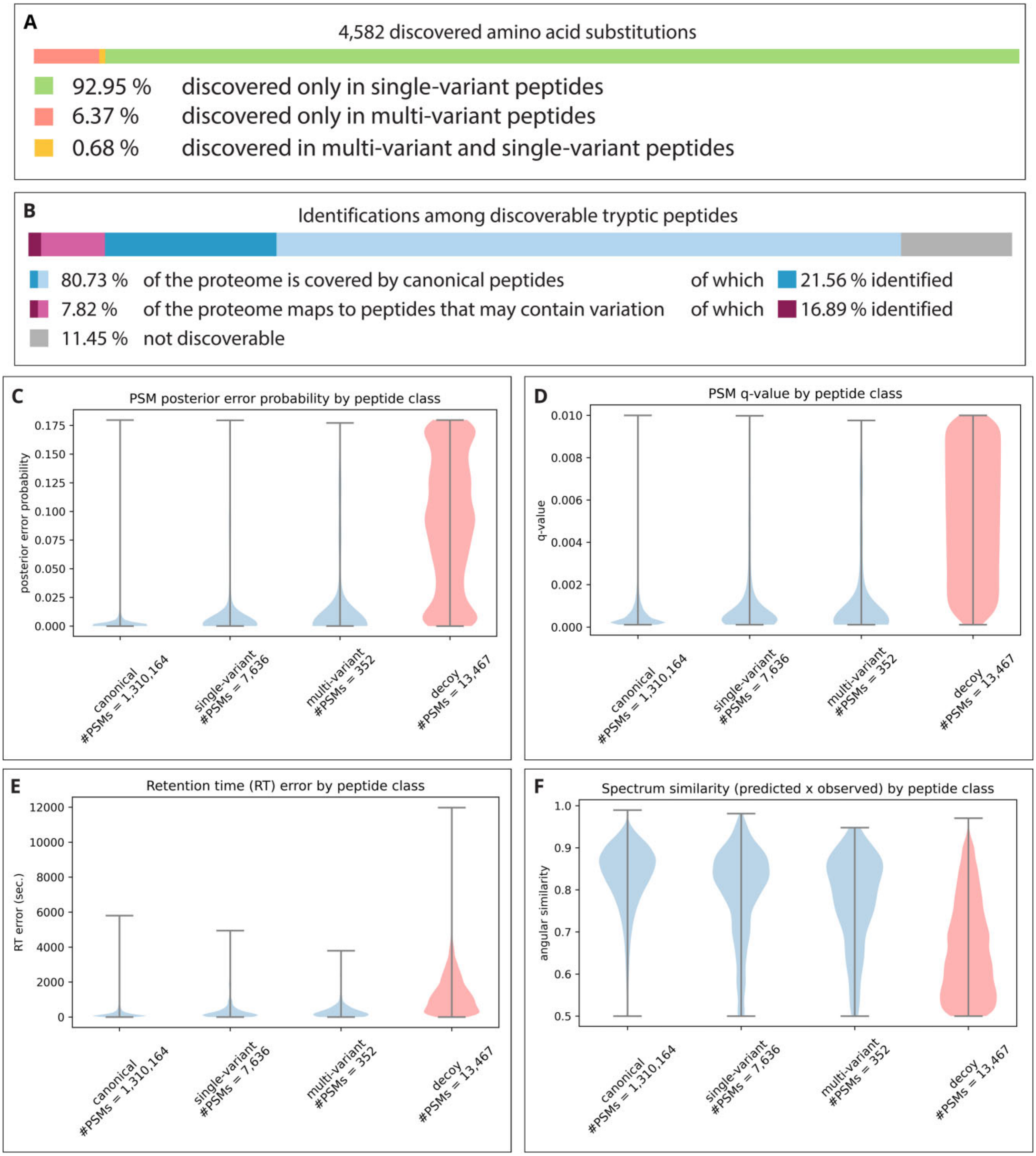

Figure 3: A: Coverage of the proteome by identified peptides, stratified by the possibility to contain variation. Lighter shades indicate the coverage by predicted peptides, darker shades represent the actual coverage by identified peptides. B: Distribution of variation in identified peptides. Amino acid variants are stratified based on the category of peptide in which the substitution caused by the respective variant can be identified. C-F: Distribution of four confidence measures among PSMs for peptide categories: posterior error probability (PEP), q-value, difference between observed and predicted retention time, and angular similarity between the observed and predicted spectrum. Decoy PSMs for this comparison were thresholded to 1% PSM-level FDR.

(Table 2). The frequency of this haplotype among participants of the European population in the 1000 Genomes project is 46%, while in this population, the aggregated frequency of all haplotypes not containing any of these substitutions is 1.98%. However, the sequence of the corresponding region of POTEI is highly similar to the sequence of actin beta (ACTB), actin gamma 1 (ACTG1), and actin alpha 1 (ACTA1), differing in 1, 1, and 2 residues, respectively. Such highly similar sequences represent peptides differing in their composition by only a few atoms, a mass difference that can be indistinguishable from a chemical or posttranslational modification (e.g., a chemical modification of methionine can be mistaken for a substitution of methionine to threonine [29]).

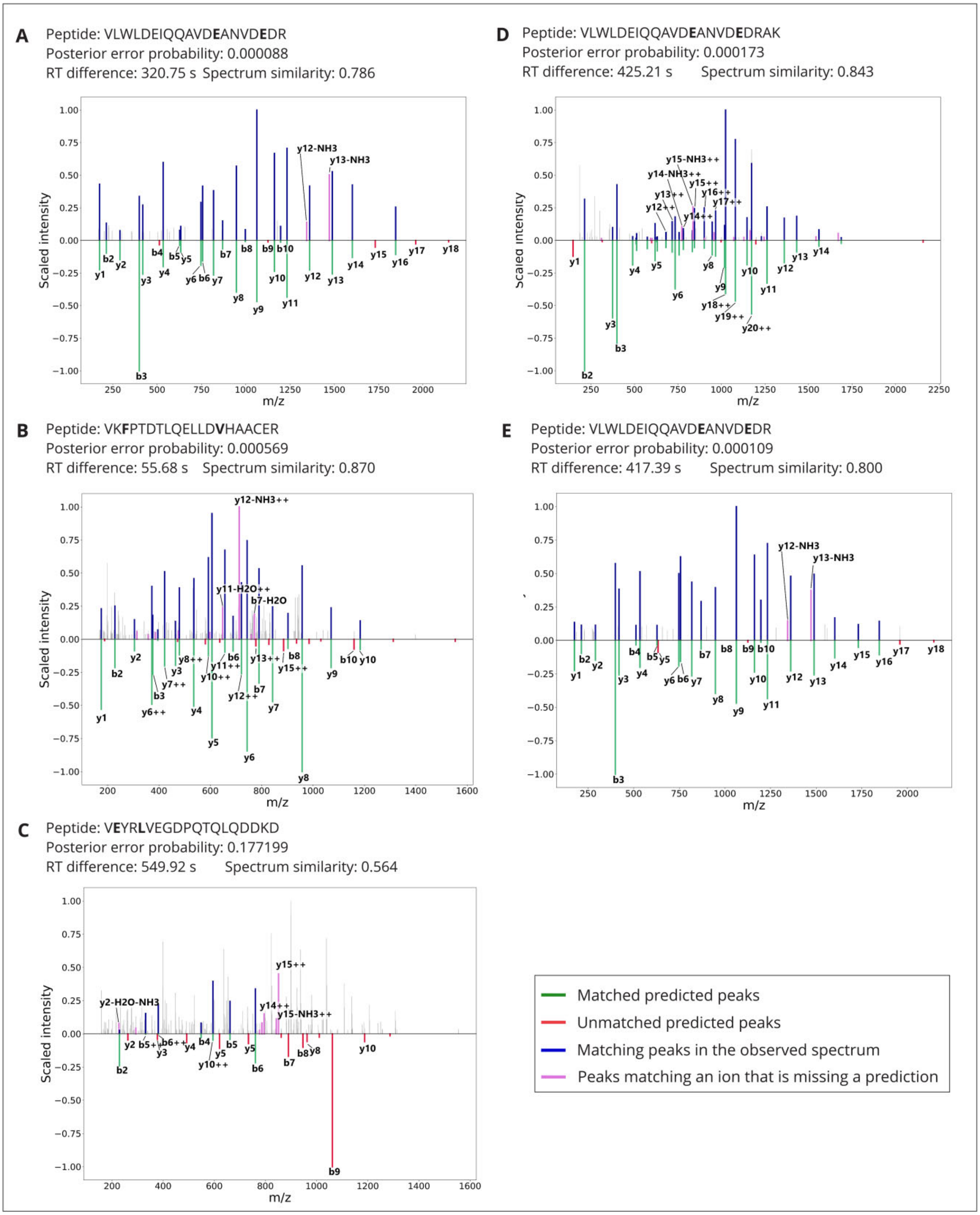

Figure 4: Quality control metrics and spectra of 5 multivariant PSMs. Amino acid substitutions are marked in bold. PSM A is among the 10% top-scoring matches to multivariant peptides by posterior error probability, B scores as the median value, and C is the lowest-scoring match to a multivariant peptide. PSMs D and E are within the 5 top-scoring matches for the most common haplotype of IQGAP2. The posterior error probability as obtained from Percolator is listed along with retention time difference to prediction as obtained from DeepLC and spectrum similarity with prediction as obtained from MS2PIP. The intensity of the measured spectrum is plotted (top; blue, pink, and gray) with the scaled predicted intensity mirrored (bottom; green and red). Peaks in the measured spectrum matching predictions are highlighted in blue, measured peaks matching an ion with a missing intensity prediction are highlighted in pink, and other measured peaks are plotted in gray. Note that in this representation, peaks matching a fragment ion with a predicted intensity of zero will not be annotated.

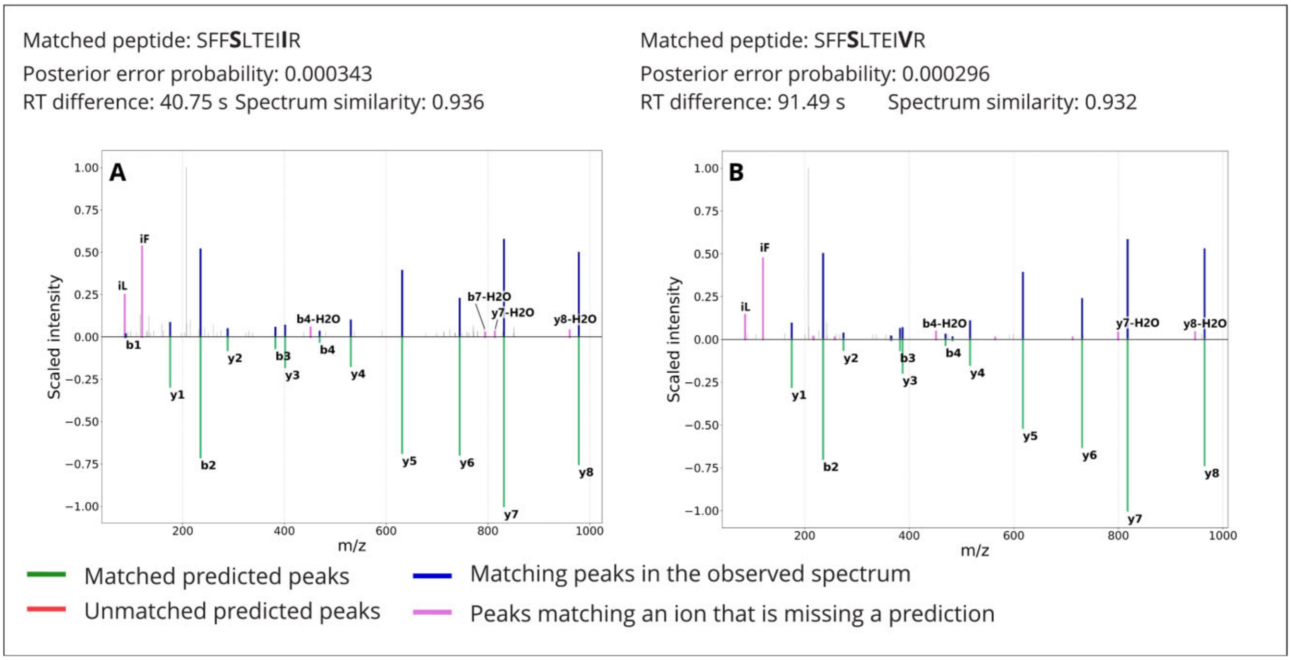

Figure 5: Quality control metrics and spectra for PSMs matching both the reference (A) and the alternative (B) allele in the same sample. The posterior error probability as obtained from Percolator is listed along with retention time difference to prediction as obtained from DeepLC and spectrum similarity with prediction as obtained from MS2PIP. The intensity of the measured spectrum is plotted (top; blue, pink, and gray) with the scaled predicted intensity mirrored (bottom; green and red). Peaks in the measured spectrum matching predictions are highlighted in blue, measured peaks matching an ion with a missing intensity prediction are highlighted in pink, and other measured peaks are plotted in gray. Note that in this representation, peaks matching a fragment ion with a predicted intensity of zero will not be annotated.

Table 2: PSMs mapping to the 5 highly similar protein sequences: actin gamma 1 (ACTG1)/actin beta (ACTB), actin alpha 1 (ACTA1), and 3 haplotypes of POTEI. REF marks the canonical sequence. We specify the number of confident PSMs matching the sequence of interest and number of samples with any spectra matching to these peptides.

Protein haplotype	Peptide sequence	No. of confident PSMs
ACTG1: REF ACTB: REF	LC**Y**VALDFE**Q**EMA**T**AASSSSLEK	3,298
ACTA1: REF	LC**Y**VALDFE**N**EMA**T**AASSSSLEK	385
POTEI: REF	LC**Y**VALDFE**Q**EMA**M**AASSSSLEK	103
POTEI: 918Y>F,929M>T	LC**F**VALDFE**Q**EMA**T**AASSSSLEK	19
POTEI: 918Y>F	LC**F**VALDFE**Q**EMA**M**AASSSSLEK	18

Therefore, telling these 2 proteins apart can be extremely challenging when accounting for variants. Conversely, if only canonical sequences are included in the database, the spectra obtained from POTEI will be arbitrarily assigned to actin. Numbers of spectra matching the corresponding regions of these proteins are listed in Table 2. Matching spectra to each of the peptide sequences in Table 2 have been identified in all 3 samples.

The need to distinguish very similar sequences makes the use of haplotype-specific databases particularly sensitive to the spectrum identification strategy. As an example, we conducted the search again after enabling the refinement procedure of X!Tandem. This procedure is a multistep approach that selects a limited set of proteins for a secondary search with different search parameters, including more modifications and relaxing thresholds (e.g., in terms of missed cleavages). While this procedure presents the advantage to quickly scan for new peptides, it makes the evaluation of matches challenging [30] and increases the likelihood to encounter cases where a modification can be mistaken for an amino acid substitution and vice versa. Fig. 6 shows such an example of 2 matches to the same spectrum, obtained using the refinement procedure: 1 peptide contains the product of the alternative allele of 2 variants (Fig. 6A) while the other has the product of the reference allele for 1 of the variants with a modification on the N-terminus compensating the mass difference (Fig. 6B). Both matches show a good matching of the higher-mass peaks and good agreements with the predictors but a high prevalence of unmatched peaks. Based on their scores, both matches would pass a 1% FDR threshold, but the similarity between the sequences makes it challenging to assess whether 1 or the other haplotype is a better match. This example shows the difficulty to distinguish variant peptides when the amino acid substitution has a mass difference equal or very similar to a modification. Overall, we observed inflated identification rates for multivariant peptides using the refinement procedure (1,060 PSMs with refinement vs. 342 PSMs without). For example, without the refinement procedure, 19 spectra matched the multivariant sequence of POTEI among the PSMs passing a 1% FDR threshold (Table 1); with the refinement procedure, the results contained 113 matching spectra. From the 94 additional matches, we suspect that many correspond to other sequences that were artifactually matched to this sequence, possibly through the addition of modifications.

Error rates derived from the target-decoy strategy rely on the modeling of the null distribution of scores using random matches. Distinguishing a variant peptide from a modified one, however, requires telling apart 2 matches that are very similar and both better than random. In such cases, it is expected that modeling the null distribution using random matches provides underestimated

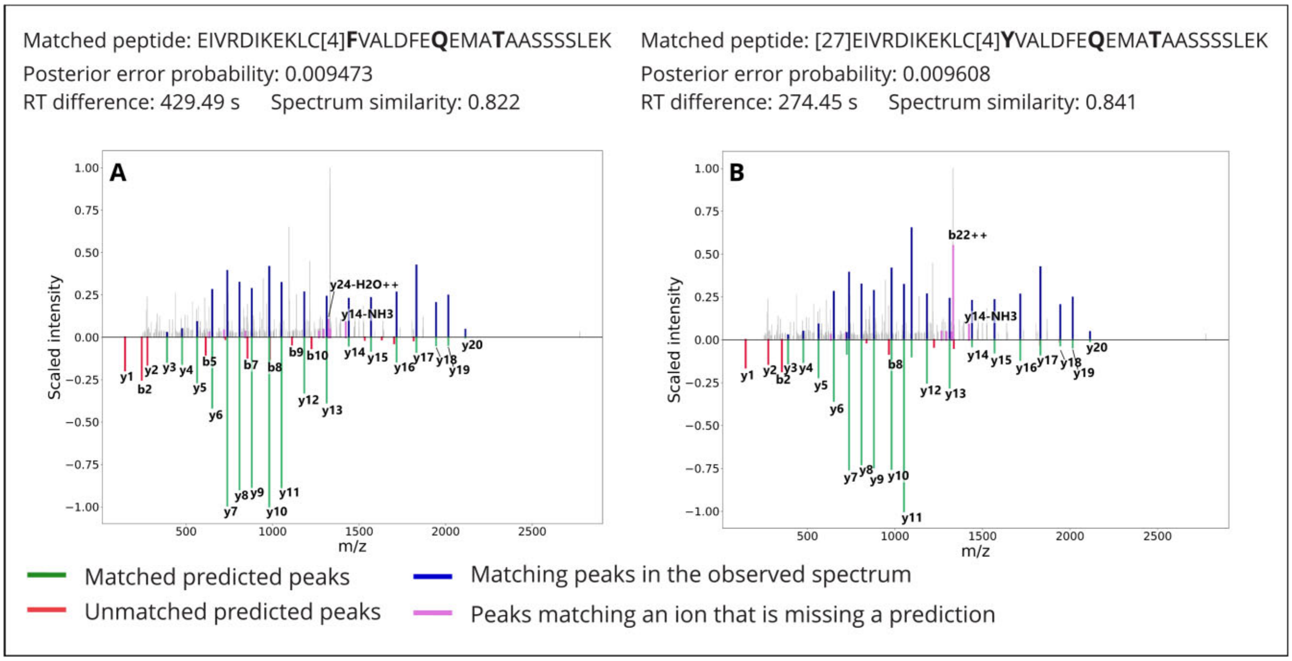

Figure 6: Comparison between predicted spectra for 2 different peptides matched to the same observed spectrum. The posterior error probability as obtained from Percolator is listed along with retention time difference to prediction as obtained from DeepLC and spectrum similarity with prediction as obtained from MS2PIP. The intensity of the measured spectrum is plotted (top; blue, pink, and gray) with the scaled predicted intensity mirrored (bottom; green and red). Peaks in the measured spectrum matching predictions are highlighted in blue, measured peaks matching an ion with a missing intensity prediction are highlighted in pink, and other measured peaks are plotted in gray. Numbers in the peptide sequence are identifiers of posttranslational modifications in UniMod [31].

error rates, and additional quality control measures can be applied to assess the quality of the matches [32]. We submitted the variant matches passing the target-decoy 1% FDR threshold in the X!Tandem search without the refinement procedure to PepQuery, a targeted peptide search engine providing additional validation for variant peptides identified using proteomics [33]. PepQuery found that a substantial share of the matches were low scoring or could also match another peptide (10% and 11% of the matches, respectively), and the prevalence of these matches decreased with the PEP (Fig. 7A). Conversely, 47% of the matches were labeled as confident, and the prevalence of confident matches increased with the PEP. The remaining matches were labeled as possibly matching a modification not considered in the original search—a rare posttranslational modification or an artifact introduced during sample preparation. Interestingly, the prevalence of such ambiguous matches was stable around 30% across PEP bins. These results highlight the difficulty posed by modifications in the confident identification of variant peptides. In the case of highly similar expected spectra between a variant and a modified peptide, analysts need to rely on prior knowledge on the likelihood of finding a given allele or modification in the sample studied or on the presence of diagnostic ions (Fig. 8). In the example of Fig. 8B, the detection of y29++ would advocate in favor of the variant peptide rather than the modified peptide, but this peak is of low intensity and therefore represents only thin evidence.

Moreover, we searched all spectra again using the search engine Tide [34], using the same parameters. Out of the 7,988 confident variant matches given by X!Tandem, 3,604 (45.12%) were confirmed by Tide. For 4,314 (54%) variant PSMs reported by X!Tandem, the spectra were not confidently matched to any peptide by Tide. The remaining 70 spectra were confidently matched to another peptide by Tide—in 51 cases to a canonical peptide, in 12 cases to a decoy peptide sequence, in 4 cases to a variant peptide encoded by a different haplotype but coming from the same gene, and in 3 cases to a contaminant.

Conclusion and Discussion

In this study, we propose to take advantage of the correlation between alleles through linkage disequilibrium to allow for the identification of peptides containing multiple linked amino acid substitutions, hence avoiding the computation of all possible combinations of alleles [14]. Co-occurring alleles in the protein-coding regions of a gene yield specific protein sequences—protein haplotypes. Building upon previous work in proteogenomics, we created a search space of protein haplotypes. We observe that 7.82% of the whole proteome maps to peptides that can contain an amino acid substitution, and up to 12.42% of all discoverable substitutions are located in peptides where multiple substitutions co-occur (multivariant peptides). These cases suggest that linkage disequilibrium between alleles resulting in amino acid substitutions should be included in a proteomics search space when identifying common variation. Subsequently, we performed a reanalysis of 3 samples of healthy tonsil tissue provided by Wang et al. [17]. We identified peptides encoded by haplotypes containing 4,582 unique amino acid substitutions compared to the reference sequences of Ensembl, 6.37% of which were found only in multivariant peptides.

Although searching haplotype-specific sequences allows for the discovery of novel peptides that match to protein haplotypes, numerous challenges still remain. Of the predicted haplotypes, 78.23% contain only substitutions, and the remaining haplotypes contain other types of polymorphisms (insertions, deletions, or polymorphisms introducing or removing a stop codon). These cannot be detected using the sequences obtained from Haplosaurus. Moreover, with the introduction of haplotypes, the search space

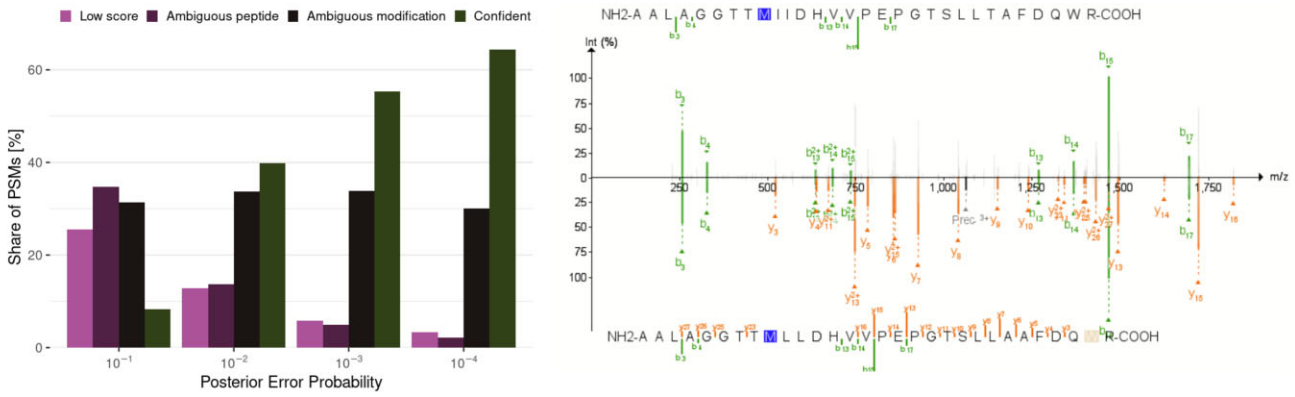

Figure 7: Analysis of variant peptides passing a target-decoy 1% FDR threshold using PepQuery. (A) Histogram of the PSM type according to PepQuery. Low score: the match was not further investigated by PepQuery due to a low score; Ambiguous peptide: the spectrum could be matched to a reference peptide at a similar score; Ambiguous modification: the spectrum could be matched to a reference peptide at a similar score when accounting for a modification that was not included in the original search; Confident: the match passed all PepQuery validation filters. (B) Mirrored annotated spectra obtained using PDV [35] of a variant PSM with better match when accounting for a modification not included in the search, here a dioxydation of tryptophan.

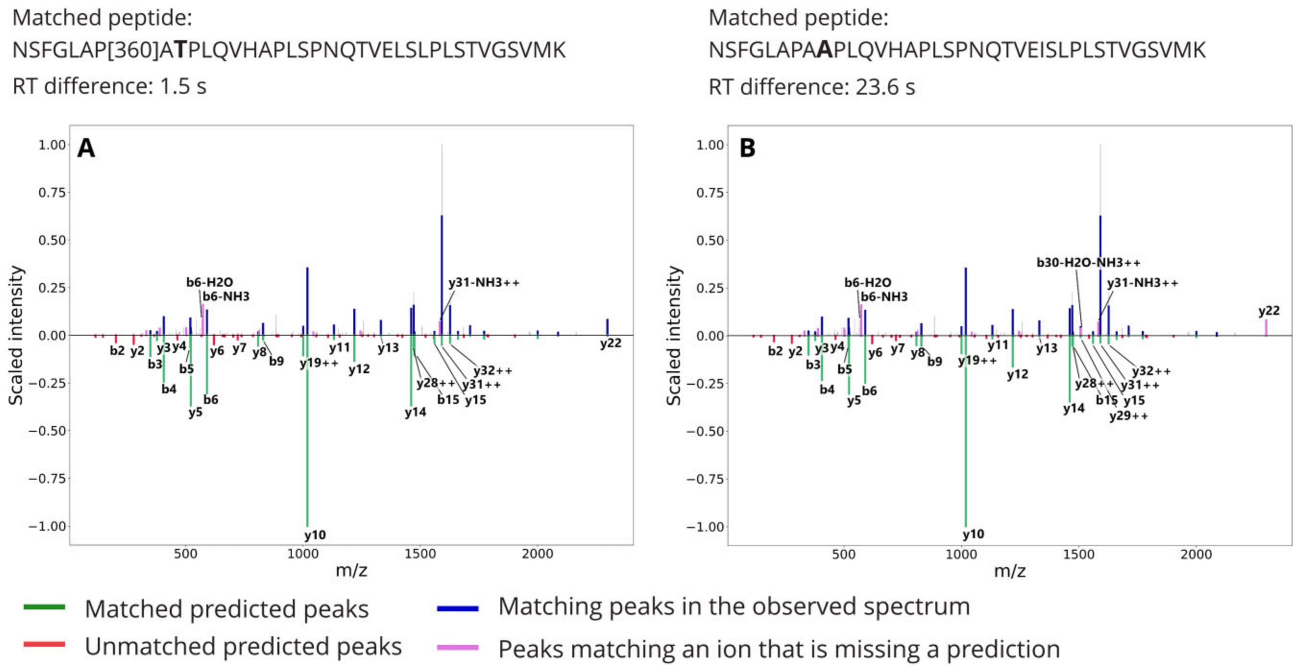

Figure 8: Comparison between predicted spectra as obtained from MS2PIP for 2 different peptides matched to the same observed spectrum. (A) Peptide candidate suggested by PepQuery is a canonical sequence with a modification. (B) Peptide candidate suggested in our search is a variant peptide. We list the retention time difference to prediction as obtained from DeepLC. The intensity of the measured spectrum is plotted (top; blue, pink, and gray) with the scaled predicted intensity mirrored (bottom; green and red). Peaks in the measured spectrum matching predictions are highlighted in blue, measured peaks matching an ion with a missing intensity prediction are highlighted in pink, and other measured peaks are plotted in gray. Numbers in the peptide sequence are identifiers of posttranslational modifications in UniMod [31].

consists of a large number of proteoforms with a high degree of similarity, making it challenging to infer which proteoform has been identified. Amino acid substitutions of a mass difference equal to a posttranslational or chemical modification are particularly challenging, as their distinction relies on the detection of few specific ions. This implies that searching without the correct haplotype or modification will generate incorrect sequences or modifications that are not caught by current error rate estimation strategies. Even worse, using the wrong haplotype on a protein sequence can result in a match in another protein. The prevalence of such errors in published proteomic datasets is currently unknown.

The dataset of protein haplotypes provided by Spooner et al. [1] was created using the genome assembly version GRCh37, which is now deprecated by Ensembl. During PepQuery analysis, we noted that a substantial share of variant peptides in GRCh37 would be canonical in GRCh38. For results that are fully up to date, a reanalysis of the data provided by the 1000 Genomes project on the current genome assembly is necessary. Limitations also come with the dataset of phased genotypes, as phasing may be inaccurate in regions with low linkage disequilibrium or in repetitive regions, resulting in an overestimation of haplotype frequencies [1]. Finally, the methods for the scoring of confidence of peptide-spectrum matches are not well suited to distinguishing between multiple candidate sequences with a high degree of similarity. In the literature, the identification of variant peptides is validated by generating reference spectra using synthetic peptides [36, 37], but such an approach presents a substantial cost and low throughput. In the present work, we used retention time and fragmentation predictors to generate the reference spectra *in silico* and used these to evaluate the matches. Predictors can instead be directly coupled to Percolator, as implemented in MS2Rescore [38], and hence provide features that can improve the discrimination power between very similar peptides.

In conclusion, accounting for protein haplotypes in the search space for mass spectrometry–based proteomic identification improves the ability to cover relevant regions of the proteome and holds the potential to be utilized in the medical context, given that the database of protein haplotypes is complete and up to date, and novel methods of quality control are developed.

Materials and Methods

Database of protein sequences

The sequence database used for the search was built using data provided by Spooner et al. [1], who generated a database of protein haplotypes using their tool Haplosaurus, available as a part of the Ensembl Variant Effect Predictor [39]. The haplotypes were generated using phased genotype data from the 1000 Genomes project Phase 3, obtained using methods described in [2]. The haplotype analysis was performed using the transcript database Ensembl version 83 [40], human reference genome assembly version GRCh37 [1]. The data provided by Spooner et al. [1] can be found at [41]. For this work, we selected only protein haplotypes generated from minor alleles with frequency at least 1% worldwide. This database was appended with the list of canonical protein sequences in the corresponding version of Ensembl and a list of common sample contaminants, obtained from [42]. The resulting search space contains 104,736 reference sequences, assembly version GRCh37, 290,080 protein haplotype sequences obtained as described above, and 116 sequences of sample contaminants. In total, 394,959 decoy sequences were generated using the algorithm DecoyPyrat [43], provided by the tool py-pgatk [12]. The final protein sequence database in the FASTA format is available as supplementary material (SD1, SD2).

Classification of peptides

We classified peptide sequences as canonical, single-variant, or multivariant based on the number of amino acid substitutions they contain. If a peptide is canonical with respect to one protein sequence and single-variant or multivariant with respect to another protein sequence, it is classified as canonical. Similarly, if a peptide is a single-variant peptide with respect to one protein sequence and multivariant with respect to another protein sequence, it is classified as a single-variant peptide. Substitutions mapping to a peptide that has been "downgraded" in such manner are not considered as discovered, or discoverable.

Public data reanalysis

We used this database to perform a reanalysis on a subset of data published and initially analyzed by Wang et al. [17]—108 fractions from 3 samples of healthy tonsil tissue digested by trypsin, fragmented using higher-energy collisional dissociation (HCD) (MS experiment IDs: P013107, P010694, P010747).

The search was performed using the command-line interface of SearchGUI v. 4.1.16 [44], employing the X!Tandem search algorithm [19], allowing for the oxidation of methionine as a variable modification and carbamidomethylation of cysteine as a fixed modification, with the "quick acetyl" and "quick pyrolidone" options of X!Tandem enabled. PeptideShaker v. 2.2.20 [45] was used for postprocessing of the search results and export of the PSMs to Percolator v. 3.5 [20], which was used to evaluate the confidence of the matches and threshold using an FDR analysis [46]. The list of PSMs was filtered to retain matches with a *q*-value below 0.01 (i.e., FDR is lower than 1%). If a peptide matched to a contaminant sequence, it was removed from further analysis. As some of the canonical protein sequences in Ensembl contain multiple stop codons, the stop codon symbols were removed from their sequences for compatibility with X!Tandem. Peptides that would contain a stop codon were removed from further analysis.

Quality control

To provide supporting evidence for the confidence of the PSMs, chromatographic retention times were predicted by DeepLC v. 1.0.0 [21], and expected peptide fragment ion intensities were predicted using MS2PIP v. 3.6.3 [22]. Peptides passing the 1% FDR threshold were used for calibration of the DeepLC predictions. The absolute distance between the centered and scaled predicted and observed retention times was computed. The MS2PIP predictions were used to measure the distance between the predicted and observed spectrum. The peaks are scaled so that the median intensity in the observed spectrum corresponds to the median intensity in the prediction. A peak in the observed spectrum is considered matching to a peak in the prediction if it differs in *m/z* by no more than 10 ppm. The distance between the matched predicted peaks and the observed ones is expressed as their angular similarity, calculated by the formula in Equations 1 and 2:

$$C\,(M, P) = \frac{\sum_{i=1}^{n} m_i p_i}{\sqrt{\sum_{i=1}^{n} m_i^2}\sqrt{\sum_{i=1}^{n} p_i^2}} \qquad (1)$$

$$A\,(M, P) = 1 - \frac{arccos\,(C\,(M, P))}{\pi} \qquad (2)$$

where $M = (m_1, \ldots, m_n)$ is the set of intensities for the matched measured peaks, and $P = (p_1, \ldots, p_n)$ is the set of intensities for the matched predicted peaks, and n is the number of matched peaks in the spectrum. $C(M, P)$ denotes the cosine similarity between M and P, and $A(M, P)$ denotes the angular similarity between M and P.

Predicted and observed spectra were also displayed as mirror plots for visual comparison in selected PSMs. The peaks in the observed spectrum matching to a predicted peak are highlighted in blue. As the intensity prediction for certain ion fragments by MS2PIP is missing, peaks matching those ions are highlighted in pink. The remaining measured peaks are displayed in gray. Peaks of the predicted spectrum are shown as negative values and labeled by the corresponding fragment ion. The predicted peaks that match a measured peak are displayed in green, and unmatched predicted peaks are displayed in red.

PepQuery analysis

The variant PSMs passing 1% FDR at PSM level using X!Tandem were further validated using PepQuery (v2.0.3) [33]. The following parameters were used: fixed modifications, carbamidomethylation of C; variable modifications, oxidation of M, ammonia loss of C, Glu→pyro-Glu of E, Gln→pyro-Glu of Q, acetylation of peptide N-term; precursor ion mass tolerance, 20 ppm; MS/MS mass tolerance, 0.05 Da; enzyme specificity, trypsin; maximum missed cleavages, 2; allowed isotope range: −1,0,1,2. The parameter "-hc" was also set in the analysis. The human protein database from

GENCODE Release 43 (GRCh37) was used as the reference protein database in the validation. The PSMs that passed all the filtering steps in PepQuery were considered confident. The filtering process is described in detail in [33]. Amino acid substitution modifications were not considered in the filtering process. PSMs classified as low scoring were assigned a score above the threshold of 12 by the Hyperscore algorithm, as is the default; see [33] for details. A complete list of variant PSMs with possible alternative peptide candidates suggested by PepQuery is available as supplementary material (SD5).

Source Code and Requirements

The pipeline to reproduce the postprocessing steps and a further description of the resulting files are provided in https://github.com/ProGenNo/FindingHaploSignatures [47].

- Project name: Finding Haplotypic Signatures in Proteins
- Project homepage: https://github.com/ProGenNo/FindingHaploSignatures
- Operating system(s): Platform independent
- Programming language: Python
- Other requirements: Snakemake v. 7.0.0 or higher, Anaconda 2022.10 or newer
- License: MIT

Additional Files

Supplementary Table S1. PSMs matching to the multivariant peptide covering a region of the most common haplotype of the IGQAP2 protein and their respective confidence measures. The posterior error probability and *q*-value as obtained from Percolator are listed along with retention time difference to prediction as obtained from DeepLC, as well as spectrum similarity with prediction as obtained from MS2PIP.

Supplementary Table S2. Search parameters used for the X!Tandem implementation in SearchGUI.

Data Availability

Supplementary data can be downloaded from figshare [47]. Other data further supporting this work are openly available in the *GigaScience* repository, GigaDB [48].

We provide the following files:

Supplementary Data 1: FASTA file including all target protein sequences (Ensembl reference proteome, protein haplotype sequences, contaminant sequences), excluding decoys.

Supplementary Data 2: FASTA file including all target and decoy sequences.

Supplementary Data 3: List of all peptide-to-spectrum matches (PSMs), resulting from the first run of X!Tandem without the refinement procedure, with all related metadata and quality control measures.

Supplementary Data 4: List of substitutions identified, along with IDs of corresponding PSMs.

Supplementary Data 5: List of variant PSMs and peptide candidates suggested by PepQuery, along with confidence scores for each peptide candidate.

Abbreviations

FDR: false discovery rate; HCD: higher-energy collisional dissociation; LC: liquid chromatography; LD: linkage disequilibrium; MS: mass spectrometry; MS/MS: tandem mass spectrometry; PEP: posterior error probability; PSM: peptide-to-spectrum match; PTM: post-translational modification.

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by the Research Council of Norway (project #301178 to MV), the University of Bergen, and the Novo Nordisk Foundation (project NNF20OC0063872 to S.J.). The funding bodies had no influence on the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

This research was funded, in whole or in part, by the Research Council of Norway 301178. A CC BY or equivalent license is applied to any Author Accepted Manuscript (AAM) version arising from this submission, in accordance with the grant's open access conditions.

Authors' Contributions

J.V. and D.S. developed the software. J.V., D.S., K.G.K., B.W., L.K., and M.V. analyzed the data. J.V. and B.W. quality controlled the results. J.V. wrote the initial draft of the manuscript. S.J., P.R.N., S.B., L.K., and M.V. oversaw the project and edited the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The computations were performed on the Norwegian Research and Education Cloud (NREC), using resources provided by the University of Bergen and the University of Oslo (https://www.nrec.no).

References

1. Spooner W, McLaren W, Slidel T, et al. Haplosaurus computes protein haplotypes for use in precision drug design. Nat Commun 2018;9:4128. https://doi.org/10.1038/s41467-018-06542-1.
2. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature 2015;526:68–74.
3. Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. Nucleic Acids Res 2022;50:D988–95. https://doi.org/10.1093/nar/gkab1049.
4. Xuan Y, Bateman NW, Gallien S, et al. Standardization and harmonization of distributed multi-center proteotype analysis supporting precision medicine studies. Nat Commun 2020;11:5248. https://doi.org/10.1038/s41467-020-18904-9.
5. Aebersold R, Mann M. Mass spectrometry-based proteomics. Nature 2003;422:198–207. https://doi.org/10.1038/nature01511.
6. Verheggen K, Raeder H, Berven FS, et al. Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. Mass Spectrom Rev 2020;39:292–306. https://doi.org/10.1002/mas.21543.
7. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics MCP 2005;4:1419–40. https://doi.org/10.1074/mcp.R500012-MCP200.
8. Pagel O, Loroch S, Sickmann A, et al. Current strategies and findings in clinically relevant post-translational modification-

specific proteomics. Expert Rev Proteomics 2015;12:235–53. https://doi.org/10.1586/14789450.2015.1042867.

9. Menschaert G, Fenyö D. Proteogenomics from a bioinformatics angle: a growing field. Mass Spectrom Rev 2017;36:584–99. https://doi.org/10.1002/mas.21483.
10. Vizcaíno JA, Kubiniok P, Kovalchik KA, et al. The Human Immunopeptidome Project: a roadmap to predict and treat immune diseases. Mol Cell Proteomics MCP 2020;19:31–49. https://doi.org/10.1074/mcp.R119.001743.
11. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat Methods 2014;11:1114–25. https://doi.org/10.1038/nmeth.3144.
12. Umer HM, Audain E, Zhu Y, et al. Generation of ENSEMBL-based proteogenomics databases boosts the identification of non-canonical peptides. Bioinformatics 2022;38:1470–2. https://doi.org/10.1093/bioinformatics/btab838.
13. Levitsky LI, Kuznetsova KG, Kliuchnikova AA, et al. Validating amino acid variants in proteogenomics using sequence coverage by multiple reads. J Proteome Res 2022;21:1438–48. https://doi.org/10.1021/acs.jproteome.2c00033.
14. Choong W-K, Wang J-H, Sung T-Y. MinProtMaxVP: generating a minimized number of protein variant sequences containing all possible variant peptides for proteogenomic analysis. J Proteomics 2020;223:103819. https://doi.org/10.1016/j.jprot.2020.103819.
15. Lobas AA, Karpov DS, Kopylov AT, et al. Exome-based proteogenomics of HEK-293 human cell line: coding genomic variants identified at the level of shotgun proteome. Proteomics 2016;16:1980–91. https://doi.org/10.1002/pmic.201500349.
16. Lobas AA, Pyatnitskiy MA, Chernobrovkin AL, et al. Proteogenomics of malignant melanoma cell lines: the effect of stringency of exome data filtering on variant peptide identification in shotgun proteomics. J Proteome Res 2018;17:1801–11. https://doi.org/10.1021/acs.jproteome.7b00841.
17. Wang D, Eraslan B, Wieland T, et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. Mol Syst Biol 2019;15:e8503. https://doi.org/10.15252/msb.20188503.
18. Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. Nat Methods 2013;10:186–7. https://doi.org/10.1038/nmeth.2369.
19. Fenyö D, Beavis RC. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem 2003;75:768–74. https://doi.org/10.1021/ac0258709.
20. Käll L, Canterbury JD, Weston J, et al. Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nat Methods 2007;4:923–5. https://doi.org/10.1038/nmeth1113.
21. Bouwmeester R, Gabriels R, Hulstaert N, et al. DeepLC can predict retention times for peptides that carry as-yet unseen modifications. Nat Methods 2021;18:1363–9. https://doi.org/10.1038/s41592-021-01301-5.
22. Degroeve S, Martens L. MS2PIP: a tool for MS/MS peak intensity prediction. Bioinformatics 2013;29:3199–203. https://doi.org/10.1093/bioinformatics/btt544.
23. Michalski A, Cox J, Mann M. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC−MS/MS. J Proteome Res 2011;10:1785–93. https://doi.org/10.1021/pr101060v.
24. Houel S, Abernathy R, Renganathan K, et al. Quantifying the impact of Chimera MS/MS Spectra on peptide identification in large-scale proteomics studies. J Proteome Res 2010;9:4152–60. https://doi.org/10.1021/pr1003856.
25. Alves G, Ogurtsov AY, Kwok S, et al. Detection of co-eluted peptides using database search methods. Biol Direct 2008;3:27. https://doi.org/10.1186/1745-6150-3-27.
26. Dorfer V, Maltsev S, Winkler S, et al. Boosting peptide identifications by chimeric spectra identification and retention time prediction. J Proteome Res 2018;17:2581–9. https://doi.org/10.1021/acs.jproteome.7b00836.
27. Cifani P, Li Z, Luo D, et al. Discovery of protein modifications using differential tandem mass spectrometry proteomics. J Proteome Res 2021;20:1835–48. https://doi.org/10.1021/acs.jproteome.0c00638.
28. O'Bryon I, Jenson SC, Merkley ED. Flying blind, or just flying under the radar? The underappreciated power of de novo methods of mass spectrometric peptide identification. Protein Sci 2020;29:1864–78. https://doi.org/10.1002/pro.3919.
29. Chernobrovkin AL, Kopylov AT, Zgoda VG, et al. Methionine to isothreonine conversion as a source of false discovery identifications of genetically encoded variants in proteogenomics. J Proteomics 2015;120:169–78. https://doi.org/10.1016/j.jprot.2015.03.003.
30. Everett LJ, Bierl C, Master SR. Unbiased statistical analysis for multi-stage proteomic search strategies. J Proteome Res 2010;9:700–7. https://doi.org/10.1021/pr900256v.
31. Creasy DM, Cottrell JSU. Protein modifications for mass spectrometry. Proteomics 2004;4:1534–6. https://doi.org/10.1002/pmic.200300744.
32. Helsens K, Timmerman E, Vandekerckhove J, et al. Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results. Mol Cell Proteomics 2008;7:2364–72. https://doi.org/10.1074/mcp.M800082-MCP200.
33. Wen B, Zhang B. PepQuery2 democratizes public MS proteomics data for rapid peptide searching. Nat Commun 2023;14:2213. https://doi.org/10.1038/s41467-023-37462-4.
34. Diament BJ, Noble WS. Faster SEQUEST searching for peptide identification from tandem mass spectra. J Proteome Res 2011;10:3871–9. https://doi.org/10.1021/pr101196n.
35. Li K, Vaudel M, Zhang B, et al. PDV: an integrative proteomics data viewer. Bioinformatics 2019;35:1249–51. https://doi.org/10.1093/bioinformatics/bty770.
36. Johansson HJ, Socciarelli F, Vacanti NM, et al. Breast cancer quantitative proteome and proteogenomic landscape. Nat Commun 2019;10:1600. https://doi.org/10.1038/s41467-019-09018-y.
37. Kuznetsova KG, Kliuchnikova AA, Ilina IU, et al. Proteogenomics of adenosine-to-inosine RNA editing in the fruit fly. J Proteome Res 2018;17:3889–903. https://doi.org/10.1021/acs.jproteome.8b00553.
38. Declercq A, Bouwmeester R, Hirschler A, et al. MS2Rescore: data-driven rescoring dramatically boosts immunopeptide identification rates. Mol Cell Proteomics 2022;21:100266. https://doi.org/10.1016/j.mcpro.2022.100266.
39. ensembl-vep. GitHub. 2023. https://github.com/Ensembl/ensembl-vep
40. Yates A, Akanni W, Amode MR, et al. Ensembl 2016. Nucleic Acids Res 2016;44:D710–6. https://doi.org/10.1093/nar/gkv1157.
41. McLaren W, Spooner W. 2017. https://doi.org/10.6084/m9.figshare.5545084.v1.
42. cRAP protein sequences. https://www.thegpm.org/crap/. Accessed 20 October 2022.
43. Wright JC, Choudhary JS. DecoyPyrat: fast non-redundant hybrid decoy sequence generation for large scale proteomics. J Proteomics Bioinform 2016;9:176–80. https://doi.org/10.4172/jpb.1000404.

44. Vaudel M, Barsnes H, Berven FS, et al. SearchGUI: an open-source graphical user interface for simultaneous OMSSA and X!tandem searches. Proteomics 2011;11:996–9. https://doi.org/10.1002/pmic.201000595.
45. Vaudel M, Burkhart JM, Zahedi RP, et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. Nat Biotechnol 2015;33:22–24. https://doi.org/10.1038/nbt.3109.
46. Käll L, Storey JD, Noble WS. Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry. Bioinformatics 2008;24:i42–8. https://doi.org/10.1093/bioinformatics/btn294.
47. Vasicek J, Skiadopoulou D, Kuznetsova KG et al., Supplementary data: Finding haplotypic signatures in proteins. 2022. https://doi.org/10.6084/m9.figshare.21408117.v3.
48. Vašíček J, Skiadopoulou D, Kuznetsova KG, et al. Supporting data for “Finding Haplotypic Signatures in Proteins.” GigaScience Database. 2023. https://doi.org/10.5524/102458.

Received: April 13, 2023. **Revised:** September 24, 2023. **Accepted:** October 8, 2023