ABSTRACT Cancers arise from the acquisition of DNA mutations, such as substitutions, deletions, amplifications, and rearrangements. Understanding the distribution and correlation of such mutations in cancer may aid the characterization of the disease and subsequent identification of biomarkers for diagnosis and treatment. The proteogenomics database (dbPepVar) created here combines genetic variation information from dbSNP with protein sequences from NCBI’s RefSeq. Public mass spectrometry datasets (Ovarian, Colorectal, Breast, and Prostate) were used to perform a pan-cancer analysis, allowing the identification of unique genetic variations. As a result, 3,726 variant peptides were identified in samples from patients with ovarian cancer, 2,543 in prostate, 2,661 in breast and 2,411 in colon-rectal cancer patients. Data resulting from the proteogenomics approach employed and connected to other biological databases is now available in an intuitive and dynamic web portal where novice users can explore general aspects of the dataset in graph or table format, or dive in to filter the data with click and select options or using more advanced queries with regex. All data can be downloaded in csv or pdf format. In perspective, the web portal developed may direct studies to identify new therapeutic targets for different cancers, and one can also use our database for characterization of variants in samples of unknown genetic background, such as archived samples.

INDEX TERMS Cancer proteomics, genetic variation, proteogenomic database.

I. INTRODUCTION Mass spectrometry (MS) based proteomics has become the primary method for comprehensive protein detection and characterization. Peptide identification is often based on challenging experimental peptide MS spectra against theoretical peptide data created from a protein sequences database such as RefSeq, Uniprot, or Gencode [1]. Those databases for protein identification do not take into consideration genetic variations in populations. This genetic variability gives individuals unique phenotypic characteristics, vulnerability to diseases, and profiles of responses to drugs. Variations in genomes might affect protein-coding sequences, producing not only a single amino acid change but also changing the reading frame, originating abnormal sequences, or removing whole portions of the protein through a premature termination codon insertion [2], [3]. Thus, peptides whose exact sequences are not found in the databases remain unidentified. Some missing sequences might have a central biological role in non-annotated protein-coding regions, specific variations of individuals, or for a specific disease mutation. So the characterization of those new proteoforms is essential for
understanding human biology [1]. To grasp full information on protein variations, it is usual to apply proteogenomics methods, a field on the intersection between genomics and proteomics.

Proteogenomics assists in the identification of new coding variants by searching the MS/MS spectra against a database of customized proteins. The protein sequences in these databases are constructed using genomic and transcriptomic information that is absent in conventional protein databases. The approach improves the annotation at the protein level, refines gene models, characterizes protein isoforms, restricts the location of the translation start and end sites, identifies splice sites, and alternative forms of splicing [4], [5], [6], [7]. In precision medicine, proteogenomics is widely used as an alternative to detect mutations in different cancers to find new biomarkers for tumors, analyze differences in levels of gene expression, compare variants to patient survival and support the development of new drugs [8], [9].

The customization of such proteogenomics databases is achieved according to the research goal. For instance, one of the widely used alternatives is the creation of a database generated from the translation of the six reading frames. However, this approach is limited because of the increase in the database search space and the failure to capture peptides present at the junction between exons [3], [6]. Other approaches develop a database incorporating mutations into populations of selected reference sequences and report the variant sequence with a special character [10]. In another case, a database of tryptic peptides was created from information found in NCBI’s dbSNP [11] by adding to the database only the mutated tryptic peptides and the corresponding reference peptides, an approach distinct from [10], which maintained the complete reference sequence. Applications of such databases include the investigation of neglected tropical diseases [12], Alzheimer’s disease [13], and other complex conditions, like cancer.

Cancer, as a complex disease, arises from combinations of mutations on the same cells over time [14], [15], triggered by external and endogenous factors [16], [17]. Distinct cancer types present different combinations of mutations [14], [15]. Mutations may alter the efficiency of molecules by hindering their stability and activity. Current techniques seek to identify these changes and determine the impact generated on gene products such as RNAs and proteins, which play a major role in the cellular functions of an organism’s processes. In cancer, aberrant proteins stimulate initiation, progression, and response to treatment. The abundances of protein and mRNA molecules are partially correlated and determining how the flow of information culminates in proteomic changes in tumors is a major under-explored issue in cancer biology [18]. An exclusive set of alterations can define a subtype profile in a cancer type [14], [19]. The Single nucleotide polymorphism (SNPs) play a fundamental role in distinct responses to the treatment of cancer patients, and also might characterize the risk of low survival outcomes [14], [15], [20].

The presence of mutations in coding regions might affect cellular signaling pathways, as well as the levels of oncogenic and tumor suppressor proteins [20].

Researchers have been developing solutions for better integrating variant discovery into proteomics studies. The Cancer Proteome Variation Database (CanProVar) integrates public data information on protein variations, provides access to known variations in proteins related to cancer types, and evaluates the impact of these variations on their functional characteristics [21], [22]. The CanProVar provides a base rich in genetic variations related to different cancer, incorporating missense, nonsense mutations, and single-base insertions and deletions derived from specific cancer bases, such as TCGA, HPI, COSMIC, and OMIM. Although there is a great diversity of variations incorporated, the base does not include untranslated region (UTR) mutations. The database Swiss-CanSAAVs was developed using the unique variant sequences of the Humansavar database, which contains only missense mutations, and integrating the MS-CanProVar database with the canonical protein database UniProtKB/Swiss-Prot [23]. For each single amino acid variation, an independent tryptic peptide with two missing cleavage sites around the central variation site was extracted from the protein sequences, and an identifier prefixed with SAAV was adopted to differentiate from the canonical protein sequence. For instance, the database Swiss-CanSAAVs was used to identify profiles of amino acid variants of subpopulations in the breast cancer cell line MCF-7 [24], identifying protein sequences [23]. The dbSAP database presents a set of variants derived from eight different SNP databases and was used to characterize mutations in various types of cancer [25]. Similar work has been done by [26] using a combination of publicly available population variants (dbSNP and UniProt) and somatic variations in cancer (COSMIC), along with sample-specific genomic and transcriptomic data to examine the variation in the proteome within and across 59 cancer cell lines.

Those databases kept variant and reference sequences within a single file. Most peptide search engines use algorithms based on statistical analysis. Such an algorithm might not assign a mutated peptide correctly because of minimal differences in each score value between the variant and reference sequences [27]. Furthermore, the approaches described above are limited to some mutations, allowing only databases with missense mutations and small insertions/deletions mutations. The mutations with the most significant impact on the sequence and protein function, such as frameshifts, variation of start translation, and loss of stop codon also are not included [28], [29], [30].

In this article, we present a variant database called dbPepVar, which contains mutated peptides built from a proteogenomics perspective. The main objective of this work is to assist in the identification of genetic variants associated with cancer at the protein level by providing a ready-to-use web portal containing processed datasets. Compared to other approaches as those described above, our database reports a greater diversity of variants, including mutations that alter
the translational starting site. Using public MS data from four types of cancer, a majority of SNPs were identified, but cancer-shared mutations were also present in a lower amount.

II. MATERIALS AND METHODS

A. DATA SOURCE

To generate the dbPepVar variant database, we used the protein RefSeq data and the dbSNP, both available on the NCBI portal at (https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/) and (https://ftp.ncbi.nlm.nih.gov/snp/). To carry out identifying variants in different types of cancer, we used experimental mass spectrometry data derived from samples from four studies: ovarian cancer [31], prostate cancer [32], colorectal cancer [9], and breast cancer [33]. All MS raw data is available at the ProteomeXchange repository (https://www.proteomexchange.org).

B. DATA PREPROCESSING

Initially, the dbSNP data were preprocessed to remove redundancies, inconsistencies, and incompleteness. The mutations of Leu to Ile or Ile to Leu were discarded since both amino acids have the same molecular mass and are not distinguished by mass spectrometry. We also removed mutations leading to alternative splicing and synonymous mutations. At the end of this process, three new files were generated containing information about the SNP, indel in-frame, and frameshift mutations. From the file containing the SNPs, four new files were generated, separated by the categories of stop-loss, untranslated region Variation (UTR variation), rare SNPs (Minor Allele Frequency <5%), and common SNPs (Minor Allele Frequency ≥ 5%) [34]. For each type of mutation, Perl scripts were implemented to create the new proteoforms according to each mutation type described. All files processed from dbSNP contain the RefSeq identifier of the protein (NP Accession), the reference amino acid, the mutated amino acid, the position of the mutation, and the SNP identifier (RefSNPs, rs).

C. CREATION OF THE PROTEOGENOMICS BASE

The first step in creating the variant database is to generate a multi-fasta file containing the proteoforms according to the dbSNP information and then extract the variant tryptic peptides. To generate the multi-fasta file, six scripts were developed in the Perl language, referring to each type of mutation: rare SNPs, common SNPs, indel in-frame, indel frameshift, stop-loss, and UTR variation. The multi-fasta files were used as input to a second script that performs the search for the tryptic peptides variant of each proteoform. The third script generates a file in the multi-fasta format, concatenating all the variant tryptic peptides for each protein. In this process, we discard variant peptides that had less than or equal to 7 amino acid residues and that were greater than or equal to 35 residues [35]. Due to post-translational processes, a protein can undergo internal cleavages, such as the removal of the initial methionine (Met), to generate a mature product of smaller size [36]. Therefore, for any N-terminal tryptic peptides, we report a duplicate with and without the initial Met. Similarly, in cases where the protein contains a C-terminal peptide with over one mutation or had more than one nonsense mutation, we generate a new entry in the fasta file for any additional peptide variation. Otherwise, the identification software could interpret them as a single peptide, as the C-terminal peptide could lack a tryptic cleavage site (Arg or Lys). For the peptides in which the mutation has removed or added a cleavage site, we verify whether the number of residues remained within the range established in our approach.

The SNPs detected from a particular protein-coding region by NGS technologies in a tumor tissue sample are possibly derived from heterogeneous cells [37]. Furthermore, non-random genetic variations tend to occur together, as haplotypes inherited from a single parent are linked on the same chromosome [37]. Therefore, peptides that showed more than one mutation were reported according to the number of possible combinations. This criterion is made only for peptides SNP-type mutations with a common allelic frequency (Minor Allele Frequency ≥ 5%) to reduce computational complexity. Using this criterion only for common mutations also allows us to reduce the search space and to avoid undesirable combinations with other types of mutations. The number of combinations that a peptide can present is according to the formula $2^n - 1$, where n is the number of mutations. For instance, for a peptide that has 12 mutations, the number of combinations will be $2^{12} - 1 = 4,095$ peptides that will be generated. This process is described in Fig. S3.

For frameshift and stop-loss mutations, we developed a script that performs the mutation in the mRNA and translates it to the respective protein. After this process, we report the variant peptide sequence that starts at the point where the mutation occurred. Missense mutation can also alter the mRNA’s untranslated region (UTR variation) by replacing the base in the initial codon. In this way, the translation machinery scans up to the next initial ATG nucleotide codon triplet [38]. In this case, we report the tryptic peptide corresponding to the new start of the translation of the protein.

We also implemented scripts to extract information such as the protein identifier (NP accession), the SNP identifier (RefSNP, rs), the position of the mutation in the sequence protein, the reference tryptic peptide, and the mutated peptide. This information was useful for missed cleavage analysis and for classifying the peptides according to their mutation type.

D. MS IDENTIFICATION AND ANALYSIS OF THE IDENTIFIED PEPTIDES

The LC-MS data in RAW format was analyzed by MaxQuant (version 1.6.14), using previously described parameters [39]. The peptide identification process was carried out using a two-stage strategy, where the MS data were sought first using the human proteome Refseq and, later, were searched using the personalized dbPepVar database. After each stage, the
“evidence.txt” files are generated regarding the identification of the peptides according to the search base used. The evidence file combines all information about the identified peptide spectrum matches (PSM) and is usually the only file needed for processing the results.

Initially, we remove false positives and contaminants identifications. For data provided from the Super-SILAC quantification protocol (Breast and Prostate cancer), peptides with an intensity value of $L = 0$ (i.e., only identified in the reference Super-SILAC cells) are removed, as it shows that the peptide was not detected in the non-reference sample. The next step is to analyze the quality of the peptide scores and select those that have the best value. All peptides identified in dbPepVar with a score lower than 50 were removed as they indicate low-quality identification.

We compared the two evidence files obtained from the two RefSeq and dbPepVar databases to check if there are two different sequences identified to the same MS spectrum. For this, the fields of the evidence file Raw File and MS/MS Scan Number are used to generate a unique identifier of the PSM. A Perl script was implemented to check and add to a file the peptides that were identified on both databases. In cases of a conflict, the script selects the variant peptide only when its score is 20% higher than the score of the RefSeq sequence identified for the same spectrum.

During the proteolytic digestion process, the enzyme can fail to cleave in one or more tryptic sites. Therefore missed cleavages are often considered during peptide identification. In the evidence file obtained from dbPepVar, we verify whether the peptides that showed missed cleavages were possible false positives. We considered false-positive variant tryptic peptides with missed cleavage whose location contradicted the actual position in the original protein. This occurs because the way our database is created, mutated peptides are concatenated even though they are not necessarily neighbouring peptides in the reference protein. For this analysis, we used the field Missed cleavages in the evidence table, where values greater than 0 indicate the presence of a missed cleavage. The registration file was used to check the position of the peptide in the source protein. In this filtering process, we also discard the variant peptides that were also present in the RefSeq base. Moreover, for mutations of the UTR variation type, we discard peptides that had a cleavage site before methionine, avoiding an erroneous identification of an enzymatic cleavage as a false new start of translation. We then classify the variant peptides according to the type of mutation. The information was obtained from the log files containing information about the mutated peptide, such as its SNP identifier, the location of the mutation in the protein, and the sequence of the reference peptide (see Fig. S2). This information allows the validation of genomic variants at the proteomic level, the location of the type of mutation that affects the peptide, the association of the variant peptide with the corresponding SNP, and the analysis of heterozygosity, through the screening of samples that show the identification of mutated peptides and variants.

Peptide identifications were carried out using publicly available MS data, by performing MS/MS searches in each dbPepVar and Refseq databases separately. The results obtained were submitted to a filtration step that resulted in the data shown in Table S1. The table shows the number of unique peptides identified for the respective types of cancer and the search database used. A removal criterion based on identification scoring was rigorously applied to peptides derived from the dbPepVar database to guarantee the reliability of the identification. The identification of the variant peptide was only considered if: i) the MS spectrum was exclusively identified in the dbPepVar database; ii) the same spectrum providing conflicting identifications in each database, the dbPepVar result must have a score value higher than 20% compared to the identification derived from RefSeq.

For example, the peptide sequences TEIQGIIGQIDEVSIEK (dbPepVar: NP_004864.1) and AAAVSVPEIEIAER (Refseq: NP_065761.1), were identified for the same MS spectrum (MS/MS Scan Number = 89008 and Raw File = 20131021_EXQ3_FaCo_SA_CL_ME180_B) in the ovary sample CL-ME180, with score values of 107.4 and 96.1.
FIGURE 1. Workflow of dbPepVar creation and the analysis of the results. Database creation step: The first step consists of generating a multi-fasta transcript file containing information about the position of the beginning of the reading frame, the transcript identifier, and the reference protein. This information was obtained from CCDS (Consensus Coding Sequence) and RefSeq Link files, which contain the association between CCDS identifiers, RefSeq protein, and transcribed RefSeq. The modified transcript Refseq was used to generate the sequences with frameshift and stop-loss mutations.

Identification step: The search process uses each base individually (dbPepVar and RefSeq) to identify the peptides.

Filtering step: In this step, the identified peptides are checked in both bases (dbPepVar and RefSeq) and verified if they have the same MS spectrum. In a redundancy case, the variant peptide with a higher score was kept. The variant peptides with scores less than 50 and false positives were removed. The false positives are the variant peptides with enzymatic cleavage error whose position differs from the reference protein.

Classification step: In this step, variant peptides are classified according to the type of mutation.

Visualization step: In the last step, the evidence tables of each cancer analyzed are available at the portal dbPepVar, and the data is integrated with the mass spectra visualization tool, Proteogenomics Viewer.

respectively. In this case, the reference peptide was kept and the variant peptide was deleted from the analysis. The adopted criterion provides an additional level of certainty in the identification of the variants present in the samples. A range of 220,405 to 341,906 variant identified peptides were found to have conflicting MS scans with reference sequences by cancer type. After analyzing the scores, 1,429 to 9,735 variant peptides remained. To remove experimental error introduced by the concatenation of variant peptides necessary to the database construction, peptides with missed tryptic cleavage
sites and whose location differed on the reference proteins were regarded as false positives and excluded. Variant peptides found on the reference database were discarded as well. All identified peptides in RefSeq and dbPepVar and their MS identification features (score, mass, mass error, and others) are available on the portal by accessing the “evidence tables” tab. When also considering peptide spectrum matches (PSMs) that were identified only in the dbPepVar search, the number of variations were detected: 3,726 in the ovarian cancer samples, 2,543 in the prostate cancer samples, 2,661 in the breast cancer samples, and 2,411 in the colorectal cancer samples (Table S1). We estimated the number of peptides specific to the types of mutations used in the construction of dbPepVar to verify the proportion concerning the different types of cancer. As expected, most mutations identified are missense SNPs (Table 1), but there are also peptides with frame-shift indels (0.5%–1%), small in-frame indels (3–4%), and stop loss (1%).

### Table 1. Number of unique peptides identified for the types of cancer: ovary (OvCa), prostate (PrCa), breast (BrCa), and colorectal (CrCa).

| Cancer Samples | SNP | INDEL | Frameshift | Stop Loss | UTR variation |
|----------------|-----|-------|------------|-----------|---------------|
| OvCa           | 3,456 | 197   | 34         | 16        | 19            |
| PrCa           | 2,340 | 139   | 29         | 10        | 26            |
| BrCa           | 2,501 | 117   | 20         | 6         | 18            |
| CrCa           | 2,266 | 104   | 22         | 8         | 11            |

After classifying and counting the identified variants, those were organized as unique or shared between samples. This last step required the use of the SNP identifier (rs, reference sequence) as a unique key to each mutation. The shared and unique counts according to the SNP identifier can be seen in Fig. 2. Fig. 2a shows that ovarian cancer has most of the identified mutations with 3,684 SNPs (horizontal bar graph), of which 2,281 were unique to the sample (vertical bar graph). From all SNPs identified, there are 365 shared by all selected colorectal, prostate, breast, and ovarian samples, as shown by the connecting dots at the bottom of Fig. 2a (sixth bar from the left). Prostate and breast cancer samples share the highest number of common SNPs, with 437 entities. Prostate and colorectal cancer samples have the least, with 81 SNPs in common. There are 248 entities shared for three types of cancers: ovarian, prostate, and breast. The prostate samples have less exclusive SNPs, but share most SNPs with other cancers. A similar pattern arises among less frequent mutations (< 5%) (Fig. 2B). While through this type of analysis it is not possible to discriminate specific cancer mutations from those that were already present in the donors genomic background, comparing unique and shared SNPs in such samples might raise interesting hypotheses about the clinical condition under study.

### B. THE WEB PORTAL

The data built into dbPepVar offers a wide range of potential opportunities for data mining and analysis. Our database, built using Shiny R, is available at https://bioinfo.imd.ufrn.br/dbPepVar/ and can be used by life science researchers who do not have command line experience that may benefit from a guided-tour of each section and tab of the main page. Here, we present a glimpse of the potential that dbPepVar has for the discovery of new data (Fig. 3 and 4). However, this paper does not cover the full extent of the data or all potential applications of the platform, which is available as an open resource for the researcher to use in their investigations.

The first menu (“dbPepVar”) contains a summary of the data accessible through the portal (Fig. 3). The graphical displays were separated by section according to the type of data and analysis that can be performed. Broadly, the initial section reports the distribution of samples, peptide sequences, and unique polymorphisms filtered by cancer type or by variant type. The latter sections summarize different aspects of the database in graphical and table format. More specifically, dbPepVar users can view graphs of the distribution of peptides and SNPs by cancer type and mutation classification (SNPs graph only). In the second section, users can explore and visualize the count of the most mutated genes, segregated by cancer type and with a responsive table explicitly showing the displayed data. As with all graphs in the portal, Plotly tools (i.e. lasso or box select) are available and allow comparing data, filtering by cancer type and gene groups from a threshold that can be defined by counting SNPs identified per sample. The responsive table also allows to filter and visualize the number of samples that have a mutation in a specific gene according to the type of cancer. Similar analysis can be done with the graph and table provided in the following sections.

The third section of the first menu exhibits the number of SNPs per gene, which may be used to build a mutational panel for each cancer type and gene of interest. The fourth and fifth sections are dedicated to amino acid change counts by sample and by SNP, respectively. In this way, it is possible to observe, at the proteomic level, the most frequent amino acid exchanges for different cancers and SNPs, which may help understand which mutations propagate from the genome to the proteome. Two additional sections summarizing other layers of integrated information are then displayed, without tables: one with chemical property changes of amino acids sorted by cancer type, where ‘Multiple’ refers to samples with frame-shift mutations, and another showing the distribution of mutated genes by chromosomal location. Thus, users can interactively perform two tasks: (i) filter and visualize the most frequent changes in amino acids according to cancer type, and (ii) filter and visualize the common exchanges between chemical groups of amino acids.

The second menu (“Variants”) shows the actual dataset in an interactive format, where users can perform data mining and generate insights for their research (Fig. 4a). This action can be done by selecting all or single rows with up to 27 columns that describe each mutation. The table includes links to GeneCards, NCBI protein, and dbSNP. Users can filter on any of the provided columns using plain text and
FIGURE 2. Variant distribution by sample types and frequency of unique and shared variants from dbPepVar found in ovary (OvCa), prostate (PrCa), breast (BrCa) and colorectal (CrCa) cancer samples. The vertical bars show the number of identified variants for each sample combination, which is indicated by the vertical line and dots in the lower portion of the figure. (a) Data where missense mutations (orange) are considered. (b) Missense mutations are removed from the data, red bars indicate indel observations and blue bars indicate UTR and stop loss mutations.

Different proteogenomic approaches have developed web portals for data availability and analysis, using different criteria. Therefore, we listed the major databases for variant proteins and compared them with dbPepVar to highlight the unique features of our approach. The result of this comparison can be seen in Table 2.

In recent years, many new biological databases of mutant proteins have been developed and published. However, all published databases have distinct and particular scopes, and to our knowledge, no databases have been published reporting the variants for cancer proteogenomics data using our reverse engineering methodology, i.e., identifying genetic mutations from altered proteins. In particular, the dbPepVar uses more refined criteria to detect peptides that accurately represent the actual peptide, such as changes in cleavage sites, peptide size, and peptides with combined mutations.

IV. DISCUSSION

The characterization of genetic mutations in their protein products is a key step to understanding their role in diseases such as cancer. However, MS-based approaches do not routinely allow the identification of polymorphisms in samples of interest. In this study, a database of variant peptides (dbPepVar) to be used in proteomics was created combining information of proteins from dbSNP and RefSeq. The dbPepVar identified genetic changes at the protein level in MS samples from four different types of cancer. In proteomics, the identification of genetic variants depends on the presence of such variants in the database used during MS spectrum matches. Many publications had suggested diverse approaches to improve such identification coverage.
This includes adding variant tryptic peptides concatenated to the reference sequence entry, such as observed in CanProVar [21], [22], Swiss-CanSAAVs [23] and dbSAP [23], [25], which may lead to false-positive identifications. Adding variant and reference peptides to the same fasta file can increase the probability of matching an inappropriate, but high-scoring peptide among the large number of available sequences. For instance, a variant peptide may be correctly assigned to an isobaric reference peptide, according to the spectrum, but still correspond to a different reference protein. Also, a peptide variant can be mismatched to a spectrum because the change in mass caused by a mutation coincides with the change in mass associated with a post-translational change in a different peptide. In dbPepVar, the variant peptides are incorporated in a single fasta file and the search is performed separately, allowing to distinguish between mutated and reference peptides during the identification process. In addition, a set of filters based on MS scores, removal of redundant sequences, and analysis of cleavage errors were developed to ensure that the identified peptides match the reported protein. The database built by Alfaro and coworkers [26] presents a similar approach. However, they only consider the minimum size of the peptide incorporated in its variant base (7 residues); dbSAP also considers only the minimum size of the peptide (10 residues). Mass spectrometry-based proteomics has some limitations, including the difficulty of identifying peptides that have very small or large sizes. For the peptides to be identified with greater precision, they must have a size in the range of 7-35 amino acid residues [44]. To avoid losing the identification of variants, dbPepVar also considered this interval as a parameter to determine the number of amino acid residues of the peptides present in the base. Swiss-CanSAAVs and the database proposed by Alfaro and coworkers have peptides with two missing cleavage sites, while dbPepVar has only fully tryptic peptides. Including peptides with missing cleavages in protein quantification does not produce significant differences in precision, accuracy, specificity, and sensitivity compared to the use of fully tryptic peptides [45].

The dbPepVar also performs the N-terminal methionine processing for peptides that have two or more mutations, as the sequences are generated by the concatenation of the peptides. This process is done to ensure the identification of variant peptides from proteins where the N-terminal...
methionine has been cleaved by co-translation by the enzyme methionine aminopeptidase [46]. The approaches presented in Table 2 assume that all digested peptides cannot have more than a single mutation. These features reduce search time, avoiding the exhaustive search for all possibilities, but naturally prevent coverage of all possible variant peptides at the same time [47]. A key advantage of dbPepVar lies in its ability to identify multiple combinatorial variants, considering all possible mutations contained in the same peptide. To avoid increasing the search space, combinations were made only for mutations with an allelic frequency greater than 5%. It is known that non-random genetic variations of a haplotype tend to occur together [37]. Therefore, the discovery of peptides with multiple mutations can be interpreted as a disease-associated haplotype, because the altered phenotypes often result from a combination of multiple factors [48]. For instance, peptides with multiple variations have been reported in ovarian and lung cancer samples [8], [37].

dbPepVar was also customized to add mutation types that affect coding regions, such as SNPs, indels, variation of the translation initiation codon, and stop-loss. SwissCanSAAVs and dbSAP have only missense mutations; CanProVar and the database developed by Alfaro et al. have almost the same types of mutations incorporated as dbPepVar, except for gene fusion (Alfaro et al.) and splice site (CanProVar) mutations.

The dbPepVar differs from these approaches by the addition of the UTR-variation/start codon variation mutations. This type of mutation affects the initial methionine, generating changes in the translation start and the untranslated region of the protein [49]. Thus, in the approach proposed by dbPepVar, the peptides were generated from the search for a new alternative translation start methionine. Clinical genetic testing has identified two variants related to endometrial and breast cancers likely to affect native translational initiation on the MLH1 and BRCA2 genes [49]. Although few peptides generated by this type of mutation have been identified, this finding highlights the existence of isoforms that are being expressed by the cell in diverse cancerous environments.

Another advantage of dbPepVar is the possibility of an association between mutations identified in cancer and genetic variations in populations, which can be made from information available in public databases, such as the TCGA. In this way, this information can be used to investigate the predisposition of individuals to the disease and how this variation propagates over generations. This data expands the scope of investigation of an individual’s predisposition to cancer development, given their genetic makeup. Recently, a study was conducted showing that the genotypes of patients with congenital heart disease may be responsible for the increased risk of cancer [50]. Thus, the recognition that genotypes
Table 2. Comparison of dbPepVar with the characteristics of databases proposed by literature.

| Databases       | Separate peptide files | Mutation types                                                                 | Chemically guided filtering                                                                 | Online Database | Data sources                                      |
|-----------------|------------------------|--------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|----------------|--------------------------------------------------|
| CanProVar       | No                     | Missense, insertions, and deletions of single amino acid, frameshifts, stop-loss, splice-site | Not reported                                                                             | Yes            | COSMIC, HPI, TCGA, OMIM, BIOMART, Sjoblom et al. [41], Greenman et al. [42], Ding et al. [43] |
| Swiss-CanSAAVs  | No                     | Missense                                                                        | Excluded Ile / Leu-Leu-Ile (post-processing), variant tryptic peptides with two missed cleavage sites | Yes            | UniProtKB/ Swiss-Pro, CanProVar, humsavar          |
| dbSAP           | No                     | Missense                                                                        | Peptide length (> 9 Aa), redundant sequence excluded                                      | Yes            | Uniprot, PMD, HPMID, MSIP, COSMIC, dbSNP, Ensembl, CanProVar |
| Alfaro et al. [2017] | Yes                   | Missense, insertions, deletions, frameshifts, stop-loss, and fusions.           | Peptide length (>6Aa), variant tryptic peptides with two missed cleavage sites, redundant sequence excluded (pre-processing) | No             | Uniprot, COSMIC, dbSNP, Exome set, RNA set        |
| dbPepVar        | Yes                    | Missense, UTR variation, stop-loss, frameshift, insertions, deletions.          | Peptide length (> 6 Aa and <36 Aa), redundant sequences excluded (post-processing), fully tryptic variant peptides, in silico cleavage of the initial methionine into peptides with two or more mutations, combinatorial analysis of mutations in peptides with more than one variant. | Yes            | dbSNP, RefSeq/NCBI                                |

Influence cancer risk can promote early clinical care and interventions and further promote lifelong health in patients.

The dbPepVar portal contains all the information presented in this work but is not limited to these findings. Each researcher can use it according to their research needs. The results described can be found by navigating to the “variants” tab and selecting the fields referring to the type of search intended. The direct link with dbSNP makes it possible to verify (i) whether the mutation identified in dbPepVar is associated with other congenital or acquired diseases throughout life and (ii) the frequency of a specific variant in different populations. dbPepVar’s variant menu also has a field indicating the remaining percentage of the protein sequence due to amino acid loss in a protein with a premature termination codon (PTC). For example, by selecting the field “PTC gene” and filtering by “TRUE”, it is possible to obtain the variant peptide sequences that cause protein shortening, as well as information associated with the quality of identification by MS and the relationship with other databases such as GeneCards, NCBI and dbSNP. Thus, this information may be useful in investigating the impact of the mutation concerning the reduction of the protein’s polypeptide chain and its relationship with some disease or alteration of its biological function. CanProVar and dbSAP present a portal with some similar characteristics to dbPepVar. CanProVar has the option to visualize the alteration of KEGG biological pathways in cancer and links that direct the user to information on genetic ontology and protein-protein interaction networks. Swiss-CanSAAVs presents in its article a link that directs to the portal, but it is inactive. dbSAP has an exclusive tab for viewing post-translational modifications and another for viewing variant and reference peptide sequences according to tissue type or cell lineage. In dbPepVar, it is also possible to visualize the PTM through the evidence tables resulting from the identification by MS. These tables also served as input for the construction of theoretical protein sequences used to visualize the information presented in the Proteogenomics Viewer. Integrating this platform to dbPepVar is unique to our approach, so the user will access the expression of the peptides and their respective mass spectra, besides being able to visualize their exonic location in the genome. All sections on the main page of dbPepVar are presented in a guided-tour, as well as the tabs described above. This feature allows users to quickly familiarize with the portal in a first encounter. To the best of our knowledge, none of the other databases ease first-user experience with any similar approaches.

dbPepVar presents a proteomics overview for several samples from different types of cancer, allowing researchers to search for information on the set of mutations that affect specific groups of samples, analyze the most frequent mutations and changes in amino acid residues, and have direct access to information regarding each type of mutation. Approaches based on mass spectrometry gain their limitations of the technique, for example, the absence of the mutant peptide in the identification due to size. In that case, the mutant tryptic peptide may be relatively small (e.g., less than six
amino acids) and therefore difficult to reliably match the corresponding MS/MS spectrum.

V. CONCLUSION
This work presents a new proteogenomic approach for building a database of variant peptides that helps identify genetic protein variations with mass spectrometry. The dbPepVar reports nonsense, nonsense, frameshift, indel, stop loss, and UTR variation mutations absent in major protein databases such as RefSeq/NCBI and UNIPROT. Furthermore, the peptides available in dbPepVar were obtained upon careful consideration of the number of amino acids in the sequence, alterations in cleavage sites, and post-translational modifications, which are essential biological characteristics for the reliability of the identification. We have also developed a web portal https://bioinfo.imd.ufrn.br/dbPepVar/ for the database in which provide information on samples of four cancer types: ovarian, colon-rectal, breast, and prostate. Our portal has different filters that help the user search for information on the genetic variations identified for each type of cancer. We also integrate our data into a platform to visualize mass spectrometry-based peptide data and the corresponding genome alignments.

In the future, we aim to expand our database by adding other types of mutations and integrating them with other databases of genetic variations. Forthcoming research may investigate the relationship of the variants reported to the different types of cancer. Finally, we intend to add new features that will allow users to submit their own data for analysis visualization.

VI. THE dbPepVar WEB PORTAL
The web portal https://bioinfo.imd.ufrn.br/dbPepVar/ was implemented with the R package ‘shiny’ (v1.6.0). The packages required are ‘tidyverse’ (v1.3.1) for data preprocessing, ‘vroom’ (v1.5.5) for efficient reading data, ‘plotly’ (v4.9.4.1) for interactive visualizations, and the ‘DT’ (v0.19) for building interactive tables.

VII. DATA AVAILABILITY
Publicly available datasets were analyzed in this study. Code used for analyses and to produce the figures is publicly available at: https://github.com/terrematte/dbPepVar. A containerized version of the web portal is also available at GitHub, with instructions for building the image. Users may also download the container image at: https://hub.docker.com/r/fiuzatayna/dbpepvar.

VIII. ABBREVIATIONS
dbPepVar: new proteogenomics database; MS: Mass spectrometry; OvCa: ovary cancer of dbPepVar; PrCa: prostate cancer of dbPepVar; BrCa: breast cancer of dbPepVar; CrCa: colorectal cancer of dbPepVar; TCGA: The Cancer Genome Atlas;

IX. AUTHOR CONTRIBUTIONS
Lucas Marques da Cunha: conceptualization. Lucas Marques da Cunha, Patrick Terrematte: code implementations. Lucas Marques da Cunha, Patrick Terrematte, Vanedelcêlio Lira da Silva, José Eduardo Kroll: methodology. Lucas Marques da Cunha, Patrick Terrematte, Tayná da Silva Fiúza: composition and draft preparation. Lucas Marques da Cunha, Patrick Terrematte, Tayná da Silva Fiúza, Gustavo Antônio de Souza: composition, reviewing, and editing. Sandro José de Souza, Gustavo Antônio de Souza: supervision, funding, and infrastructure. All authors contributed to the article and approved the submitted version.

X. CONFLICT OF INTEREST
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGMENT
The data processing was conducted with computational resources of the High-Performance Computing Center (NPAD) (https://npad.ufrn.br) and the Bioinformatics Multidisciplinary Environment (BioME) at the UFRN (https://bioinfo.imd.ufrn.br).

REFERENCES
[1] G. M. Sheynkman, M. R. Shortreed, A. J. Czesnik, and L. M. Smith, “Proteogenomics: Integrating next-generation sequencing and mass spectrometry to characterize human proteomic variation,” Annu. Rev. Anal. Chem., vol. 9, no. 1, pp. 521–545, Jun. 2016, doi: 10.1146/annurev-anchem-071015-041722.
[2] A. Végvári, “Mutant proteogenomics,” in Proteomics (Advances in Experimental Medicine and Biology), vol. 926, Sep. 2016, pp. 77–91. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-42316-6_6, doi: 10.1007/978-3-319-42316-6_6.
[3] S. Renuse, R. Chaerkady, and A. Pandey, “Proteogenomics,” Proteomics, vol. 11, no. 4, pp. 620–630, Feb. 2011, doi: 10.1002/pmc.201000615.
[4] G. Menschaert and D. Fenyo, “Proteogenomics from a bioinformatics angle: A growing field,” Mass Spectrometry Rev., vol. 36, no. 5, pp. 584–599, Sep. 2017, doi: 10.1002/mas.21483.
[5] K. V. Ruggles, K. Krug, X. Wang, K. R. Clauser, J. Wang, S. H. Payne, D. Fenyo, B. Zhang, and D. R. Mani, “Methods, tools and current perspectives in proteogenomics,” Mol. Cellular Proteomics, vol. 16, no. 6, pp. 959–981, Jun. 2017, doi: 10.1074/mcp.MR117.000024.
[6] A. I. Nesvizhskii, “Proteogenomics: Concepts, applications and computational strategies,” Nature Methods, vol. 11, no. 11, pp. 1114–1125, Nov. 2014, doi: 10.1038/nmeth.3144.
[7] N. Castellana and V. Bafna, “Proteogenomics to discover the full coding content of genomes: A computational perspective,” J. Proteomics, vol. 73, no. 11, pp. 2124–2135, Oct. 2010, doi: 10.1016/j.jprot.2010.06.007.
[8] S. Xu, S. W. Cha, S. Na, C. Guest, T. Liu, R. D. Smith, K. D. Rodland, S. Payne, and V. Bafna, “Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data,” Proteomics, vol. 14, nos. 23–24, pp. 2719–2730, Dec. 2014, doi: 10.1002/pmc.201400206.
[9] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, and S. R. Davies, “Proteogenomic characterization of human colon and rectal cancer,” Nature, vol. 513, no. 7518, pp. 382–387, Sep. 2014, doi: 10.1038/nature13438.
[10] S. Schandorff, J. V. Olsen, J. Bunkenborg, B. Blægoev, Y. Zhang, J. S. Andersen, and M. Mann, “A mass spectrometry-friendly database for cSNP identification,” Nature Methods, vol. 4, no. 6, pp. 465–466, 2007, doi: 10.1038/nmethod0607-465.
LUCAS MARQUES DA CUNHA received the B.S. degree in information systems from the University Center of Patos (UNIFIP), Paraíba, Brazil, and the M.S. degree in computer science from the Federal University of Paraíba (UFPB), Paraíba, in 2016. He is currently pursuing the Ph.D. degree in bioinformatics with the Federal University of Rio Grande do Norte (UFRN), Brazil.

He is also working as an Assistant Professor with the Faculty of Computer Science, Federal University of Rondônia (UNIR), Porto Velho, Rondônia. He is an expert on biologicals database development and forensic analysis on digital images. He developed a method for detecting tampering in digital images for different compression ratios. He worked in remote learning at the Federal University of Paraíba and supervised several researches on technologies applied to education. He works with teaching programming for mobile devices for android using the Java language. He has over six years of research and teaching experience. His research interests include artificial intelligence, explainable machine learning, bioinformatics, proteogenomics, and computational modeling.

PATRICK TERRERATTE received the B.S. degree in systems analysis from the Federal Institute of Education, Science and Technology of Rio Grande do Norte (IFRNR), Rio Grande do Norte, Brazil, in 2011, the M.S. degree in systems and computation from the Federal University of Rio Grande do Norte (UFRN), Rio Grande do Norte, in 2013, and the Ph.D. degree in bioinformatics from the Bioinformatics Multidisciplinary Environment (BioME), UFRN, in 2022.

From 2016 to 2022, he was an Assistant Professor of computing engineering at the Federal Rural University of Semi-arid, Department of Engineering and Technology (DETEC), Brazil. He is currently an Assistant Professor at the Metropolis Digital Institute (IMD), UFRN. He has published on survival prediction for clear cell renal cell carcinoma. His research interests include artificial intelligence, explainable machine learning, information theory, feature selection, bioinformatics, survival analysis, systems biology, computational modeling, and logic.

TAYNA DA SILVA FUÍZA received the B.S. degree in biotechnology from the Federal University of Ceará (UFC), Ceará, Brazil, in 2017, and the M.S. degree in bioinformatics from the Federal University of Rio Grande do Norte (UFRN), Rio Grande do Norte, Brazil, in 2019, where she is currently pursuing the Ph.D. degree in bioinformatics.

She worked with isolating proteins and carbohydrates from marine algae, evaluating their biological activities in animal models, animal cell culture and transfection. Her research interest includes the development of computational approaches for the identification of potential vaccine candidates in the genomes of pathogenic microorganisms and applying immunoinformatics tools for neglected diseases.

GUSTAVO ANTONIO DE SOUZA received the B.S. degree in computer science from the State University of Rio Grande do Norte (UERN), Rio Grande do Norte, Brazil, the M.S. degree in genetics from the Federal University of Pará (UFPA), Pará, Brazil, and the Ph.D. degree in bioinformatics from the Federal University of Rio Grande do Norte (UFRN), Rio Grande do Norte.

He was a Bioinformatics Specialist at Duna Bioinformatics (2019–2020) and a Postdoctoral Researcher at the Laboratory of Human Genetic Diversity (LDGH), Federal University Minas Gerais (2020–2022). He is currently a Postdoctoral Researcher at the Beneficiência Portuguesa Hospital, Sao Paulo, Brazil. He has experience in computer science, focusing on databases. He has experience with cloud computing and AWS. His research interests include piRNAs, gastric tissues, microarray, and drosha.

JOSE EUDARDO KROLL received the B.S. degree in pharmacy and the M.S. degree in biotechnology from the University of Mogi das Cruzes, Sao Paulo, Brazil, in 2006, and the Ph.D. degree in bioinformatics from the Federal University of Sao Paulo, Sao Paulo.

He was a Senior Bioinformatician at the Institute for the Heart (2020–2021), a Chief Technology Officer at Duna Bioinformatics (2019–2020), and a Postdoctoral Researcher at the Brain Institute (2014–2018). He currently works as a Bioinformatics Specialist at a Hospital & Health Care Company (DASA), Sao Paulo. His research interests include diagnostics, genomic profiling for disease risk, proteogenomics, and cancer bioinformatics.

SANDRO JOSÉ DE SOUZA received the degree in biology from the Federal University of Parana, Parana, Brazil, in 1989, and the Ph.D. degree in biochemistry from the University of Sao Paulo, Sao Paulo, Brazil, in 1993.

From 1995 to 1998, he was a Pew Latin American Fellow at Harvard University. He was one of the pioneers of genomics and bioinformatics in Brazil. He was an Associate Member of the Ludwig Institute for Cancer Research, from 1999 to 2012. He was elected by the World Economic Forum as a Young Global Leader, in 2009. He was a Tinker Visiting Professor at the University of Chicago, in 2011. He is currently a Full Professor at the Brain Institute, UFRN. He is also a Researcher 1B at the Brazilian National Council for Scientific and Technological Development (CNPq).