A tool for integrating genetic and mass spectrometry-based peptide data: Proteogenomics Viewer

PV: A genome browser-like tool, which includes MS data visualization and peptide identification parameters

José Eduardo Kroll[1)2)3)], *Vandeclécio Lira da Silva*[2)3)], *Sandro José de Souza*[2)3)] *and Gustavo Antonio de Souza*[2)3)4)*]

In this manuscript we describe Proteogenomics Viewer, a web-based tool that collects MS peptide identification, indexes to genomic sequence and structure, assigns exon usage, reports the identified protein isoforms with genomic alignments and, most importantly, allows the inspection of MS2 information for proper peptide identification. It also provides all performed indexing to facilitate global analysis of the data. The relevance of such tool is that there has been an increase in the number of proteogenomic efforts to improve the annotation of both genomics and proteomics data, culminating with the release of the two human proteome drafts. It is now clear that mass spectrometry-based peptide identification of uncharacterized sequences, such as those resulting from unpredicted exon joints or non-coding regions, is still prone to a higher than expected false discovery rate. Therefore, proper visualization of the raw data and the corresponding genome alignments are fundamental for further data validation and interpretation.

Also see the video abstract here: http://youtu.be/5NzyRvuk4Ac

Keywords:
alternative splicing events; genome browser; mass spectrometry; peptide identification; proteogenomics

Additional supporting information may be found in the online version of this article at the publisher's web-site.

Introduction

Protein identification by mass spectrometry (MS) is routinely performed through probabilistic scoring of observed ion fragmentation spectra with theoretical fragmentation data from sequences in a protein database of choice [1]. Therefore, the majority of the proteomics researchers will favor peptide search procedures that use reference datasets containing "known" sequences and as little entry redundancy as possible, in order to control the search space size and reduce the chance of false-discoveries [2, 3]. In contrast, proteogenomics studies investigating alternative splicing or non-coding regions require custom made databases containing "novel" sequences predicted either from genome or transcriptome (ESTs, RNASeq, etc.) data [4–7]. However, the identification of such novel peptide variants in proteogenomics studies are still susceptive to higher than expected false-discovery rates (FDR) [8, 9], mostly due to the larger search

DOI 10.1002/bies.201700015

[1)] Institute of Bioinformatics and Biotechnology, Natal – RN, Brazil
[2)] Brain Institute, Universidade Federal do Rio Grande do Norte, Natal – RN, Brazil
[3)] Bioinformatics Multidisciplinary Environment, Instituto Metrópole Digital, UFRN, Natal-RN, Brazil
[4)] Department of Immunology and Centre for Immune Regulation, Oslo University Hospital HF Rikshospitalet, University of Oslo, Oslo, Norway

***Corresponding author:**
Gustavo Antonio de Souza
E-mail: gadsouza@neuro.ufrn.br

Abbreviations:
FDR, false-discovery rate; **MS,** mass spectrometry.

space of the customized databases (and consequently higher chance of spurious matching).

This has led to arguments that such peptide variants should have supporting evidence that is stronger than the criteria used for the remaining of the identified dataset [10, 11]. This was clearly illustrated in the recent two independent publications of the human proteome draft [12, 13], whose authors claimed the identification of peptides from over 17,000 genes in various tissues and cell types, in addition to peptide identification evidence for hundreds of novel genes that are located on long non-coding RNA and other non-coding regions accordingly to Ensembl. A validation of the data by other researchers [11, 14, 15] clearly demonstrated that possibly a third of the identified protein products from known genes and the vast majority of the identifications belonging to novel regions were incorrectly assigned. This was shown to be either an incorrect interpretation of very poor quality spectra or the result of poor protein inference decisions (e.g. a novel peptide from a non-coding region had an amino acid sequence that was not unique and was also shared with a known, highly expressed protein present in the sample). While both proteome draft publications released online tools regarding their data analysis, to our knowledge only Wilhelm and co-workers [13] made MS2 PSM identification publicly available.

This demonstrates the relevance of proper validation and visualization of MS-based peptide identification as a central feature that determines proper standards for publication and sharing of proteomics datasets. While designing data repositories that allow such features of MS2 spectra collected independently is not a new concept (and it has been performed elegantly elsewhere) [16, 17], to our knowledge most of such tools are designed to integrate MS-based datasets to reference protein sequence databases such as Uniprot or Ensembl. Therefore, only very few efforts have been done regarding genomic integration and characterization of MS data from uncharacterized novel sequences from proteogenomics approaches [18]. For example, the prediction of novel isoforms has been mostly performed through transcriptomics, but the improvement in data availability for MS-based proteomics in the past years [19, 20] now allows the characterization of such events through a proteogenomics perspective, i.e. integrating genome and MS data. Furthermore, currently available proteogenomic tools either does not allow spectra inspection or does not report all identification features necessary for proper proteogenomics validation (as guidelined by [10]).

Thus, Proteogenomics Viewer presented here was designed to fulfill that need. It is an easy-to-use and straightforward tool to visualize and assert MS2 quality of any proteomic dataset, with the aim to map and integrate peptide data to genome structure. While it can be used to any dataset that is not necessarily acquired under a proteogenomic effort, its strength is that it was designed as a tool to help proteogenomic researchers to rank the quality of novel peptide forms based on better integration and accessibility of parameters such as spectra quality, coverage of the identified fragment ions in the MS2, peptide uniqueness, posterior error probability scoring, and other possible inference issues.

Methods

Generation of input data for Proteogenomics Viewer

MS raw files

We reanalyzed the following project raw files: PXD000561 [12] and PXD000865 [13], publicly available at ProteomeXchange repository [20]. We have also used data from 11 cell lines [21] obtained from a public repository that is currently discontinued. This dataset was downloaded prior to repository shutdown and was used here with the owner's permission (Mathias Mann, personal communication).

RNASeq data

Sixteen tissues from the Human Body Map Project V2 were downloaded from the ArrayExpress Portal (http://www.ebi.ac.uk/arrayexpress, experiment accession E-MTAB-513). Human genome (hg38) and the RefSeq reference transcriptome, needed for the mapping and assembly, were downloaded from the UCSC Genome Portal (http://genome.ucsc.edu).

Database creation and search parameters

The FASTA database (here called MasterSet) was created using Uniprot database as reference, and any additional isoforms including unique alternative splicing events (ASE) variants from Ensembl (unique as not already present in Uniprot), unique entries predicted from RNASeq data from the Human Body Map V2 project (unique as not present in either Uniprot or Ensembl), or unique entries predicted from Splooce [22] (unique as not present in any of the previous three sequence sources). All of those unique sequences were added to the Uniprot database (complete entries, downloaded from Uniprot in August 2015), generating a total set of 228,153 entries. All ".RAW" files were submitted to peptide/protein search using the Andromeda search engine (MaxQuant v. 1.5.2.8) [23, 24] using standard parameters as described in the Supporting Information.

Any data input in .txt format, from any search engine and not only MaxQuant/Andromeda suit, can be loaded into Proteogenomics Viewer if the correct identifiers are used, as listed in the Supporting Information. In order to follow PSI standards, an mzid converter is fully integrated into Proteogenomics Viewer, therefore datasets in .mzid format can also be used as input. At the moment, the converter is compatible for data searched using X!Tandem, MS-GF+, and OMSSA engines.

Data processing prior to Viewer analysis

RNASeq proteins prediction

Next-generation data from the Human Body Map Project V2 were mapped against the human genome hg38 through a Tophat 2.1.0-based pipeline using default parameters [25] and using a RefSeq GTF file as a reference transcriptome. Tophat

results were carried out through the Cufflinks 2.2.1 [26] using the RefSeq transcriptome for guiding the transcriptome assembly. The resulting GTF files were then converted to amino acid sequences using an in-house script, which has been used before [22, 27]. Briefly, all GTF entries were converted to nucleotide sequences considering only the exon coordinates. As a result, a FASTA file of mRNAs was created and the longest CDS of each entry was converted to an amino acid sequence.

Peptide alignment

An exhaustive algorithm based on regular-expressions was developed to map the set of peptides to the protein database. It allowed the interchange between Isoleucine and Leucine amino acids by converting both their characters (I and L) to X. For the alignment, the first four amino acid sequences found after all tryptic sites and from the beginning of the proteins were indexed using a hash table, in order to speed up the process since thousands of peptides and protein entries were expected as input. The peptides were then checked against the hash table, which returned a list of candidate proteins. In the final step, a regular expression was used to confirm the perfect match between the peptides and the proteins, and to recheck the tryptic sites.

Peptide abundance

Identical peptide sequences identified from different peptide spectrum matches (PSMs) (different charges, modifications, etc.) were reduced to decrease data redundancy, but all calculated AUC intensities were kept and used to infer the peptide quantitation using their median distribution. For each sample, the average and standard deviation were calculated for the peptides intensity, previously converted to a log2 scale. The normalized peptides expression was then obtained by a *z*-score test, where each peptide intensity was compared to its sample intensity distribution. The data are also provided as density box plot containing all PSM quantitation for each peptide, in order to demonstrate distribution prior to removal of PSM redundancy.

Variants identification

A binary matrix showing the presence/absence of matching peptides for each protein was built. Proteins showing peptide entries fully contained within another protein were removed. In cases in which different protein entries shared exactly the same peptide, only one was selected by the algorithm, prioritizing the proteins from Uniprot, Ensembl, Splooce and finally those generated through RNASeq (Human Body Map V2), in that order. This was done to avoid ranking annotated peptides as novel sequences.

Proteins alignment

Proteins were aligned against the human genome version hg38 through the Exonerate aligner [28]. A heuristic search was set and only the best alignment was reported (command "-n1 -m protein2genome"). An exhaustive approach was performed when the heuristic search did not find any result (command "-E -n1 -m protein2genome:bestfit"). The results were processed and all peptides and proteins were indexed to genome coordinates. Peptide sequences with imperfect or absent alignment to the reference genome (e.g. sequences resulted from fused genes, showing long insertions, etc.), will fail to index and were automatically excluded from the analysis based on the alignment score. Only splicing variants and sequences showing small variations were expected as input for the viewer.

Gene annotation

A strategy was created for the annotation of genes, which was based on the comparison of fragments of peptides (k-mers) between the unannotated user protein dataset and an annotated reference protein database. As an advantage, it showed to cluster proteins in an efficient way without the need of other expensive processing alignments. Reference proteins from the Ensembl repository were indexed by splitting 12 k-mer sequences at each three amino acids. The k-mers were stored in a hash tree, relating each of them to a gene name, and those found in more than one gene were removed from the analysis. Unannotated proteins from the user dataset were then split using the same previously k-mer size at each amino acid. Each resulting k-mer was directly checked against the hash table, and the unannotated protein entries were assigned to genes showing the highest k-mer correspondence.

Viewer implementation

Proteogenomics Viewer was written in HTML and JavaScript. The proteome data were stored in JSON format, properly compressed using the LZMA algorithm and directly saved to the disk. For loading the data into the Viewer, a simple AJAX protocol was chosen. Third-party software was avoided minding the simplification of its installation process. For more details, Proteogenomics Viewer source code and the scripts needed for processing the proteome data were made available at: https://bitbucket.org/jekroll/pb_base. The results from this analysis can be accessed at http://www.bioinformatics-brazil.org/protviewer/.

Signal sequences and transmembrane domains prediction

To identify changes in the number of signal sequences or transmembrane domains from sequences from Splooce, when compared to the Uniprot reference, we submitted sequences present at the MasterSet database into the TMHMM predictor tool version 2.0 stand-alone [29]. After filtering the sequence entries with or without predicted domains, we quantified the number of domains for each sequence entry using a script developed in Perl. Examples in which Splooce sequences had a loss of one or more domains when compared to Uniprot reference sequence were selected and used for further analysis. Finally, such entries were filtered to contain only those with collected MS evidence, and manual inspection of MS data were performed to report unpredicted isoforms with good data quality.

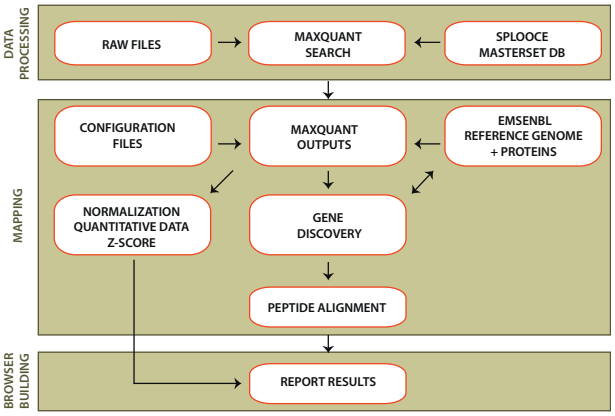

Figure 1. Bioinformatics workflow. Raw data from MS datasets is submitted to peptide/protein identification searches using Maxquant/Andromeda engine and a database in FASTA format containing sequences from Uniprot (reference), Ensembl, Splooce, and unique sequences from the Human Body Map project. Peptide identification information from MaxQuant is used as an input for genomic alignment using Hg18 reference genome (and its translated proteins). Quantitative data are also processed and reported together with alignment information into Proteogenomics Viewer.

Results

Generation of data input for Proteogenomics Viewer

Proteogenomics Viewer can process peptides and PSM lists in .txt format from any search engine, as long as the correct identifiers are used, or .mzid files from compatible search engines (see Supporting Information). For this analysis, we implemented the Viewer using public MS datasets [12, 13, 21], which were processed by MaxQuant [23, 24] using an in-house developed protein sequence database [27]. We generated such database containing protein sequences predicted by Splooce [22], a tool that retrieves and presents alternative splicing events found in public expressed sequence datasets and also those predicted from RNASeq data from the Human Body Map Project V2. To decrease the chances of incorrectly claiming an annotated isoforms as novel, a non-redundant set of human protein sequences from Uniprot and Ensembl were also included (Fig. 1). Therefore, novel isoforms predicted by Splooce can only be identified when the identified peptide is unique and not shared between a Uniprot or Ensembl sequence. Supporting Information Figure S1 shows an example of a splicing variant (as predicted by Splooce) for the gene GFAP (glial fibrillary acidic protein) aligned to its Uniprot reference. It is also important to note that the majority of tryptic peptides present in the entries sources used are also present in the Uniprot entries. In spite of the fact that the database size is approximately 3.2× larger than Uniprot (increasing the processing time of the search engine), the search space itself, which could interfere in peptide identification efficiency, increased by only 11% when all unique peptides are considered in addition to those already in Uniprot alone (data not shown).

The MaxQuant output identified a total of 14,020,392 peptide spectra matches, resulting in the identification of 332,673 non-redundant, unique peptides, leading to the identification of 20,845 protein groups that were the product of the expression of 13,786 genes. From those peptides, 328,680 are observed in the Uniprot reference database, 3,073 are peptides that are uniquely observed in entries from Splooce, 755 are unique to Ensembl and 165 are unique to the Human Body Map (Fig. 2). As discussed [8, 9], false-discovery rates within novel peptide identifications are expected to be higher. If a PEP threshold of 0.05 (5%) is applied to the identified peptides, 318,415 Uniprot peptides will be confirmed (95.71%). However, only 2,079 (67.65%) Splooce-unique peptides have PEP score below that threshold, confirming that higher rates of false-discoveries are assigned to "novel" peptide identifications. Similar trends were observed for Ensembl unique peptides (87.1%) and Human Body Map unique peptides (64.24%).

The MaxQuant output was also checked for the identification of olfactory receptors. These genes have a very low expression breadth and their protein products are expected to be detected solely in the airway epithelium and olfactory neurons (none is represented in the public datasets used here). Therefore, those proteins could be used to benchmark incorrect protein assignment by the search engine used by us. Using the protein identification protocol discussed above, our search engine identified seven olfactory receptors (based on unique peptides). Five of those had PEP scores between 0.05 and 0.01, and two (OR2M7 and OR51B4) had a PEP score just below 0.01, and in principle could be considered good identifications. However, the MS2 of those peptides clearly showed poor spectra, indicating the possibility they are cases of misidentification (Supporting Information Fig. S2). This demonstrates that such receptors were also misidentified in the original human proteome drafts reports [12, 13], as further analysis of the drafts indicated [15].

Proteogenomics Viewer and data processing

Proteogenomics Viewer is a web-tool written in HTML and JavaScript that uses some popular libraries. Proteogenomics Viewer data was created using in-house Perl scripts and external tools, such as the Exonerate protein-genome aligner [28]. The analysis pipeline, detailed in Fig. 1, depends on few files created by the search engine software, which were reorganized for the proper identification and visualization of protein variants.

Using MaxQuant generated datasets, the data processing involved only two output files: msms.txt that stores the peptides, spectra and scores information, and the evidence.txt file that stores the peptides intensities. The data were processed through three major steps: i) peptides alignment, ii) variants identification, and iii) protein-genome alignment. Peptides were aligned to the respective protein database using an exhaustive algorithm based on regular expressions. It allowed the interchange between Isoleucine and Leucine amino acids, besides a recheck of all tryptic sites, in order to avoid matching incorrect proteins showing identical peptide sequences which are not preceded by a Lys or an Arg. Protein

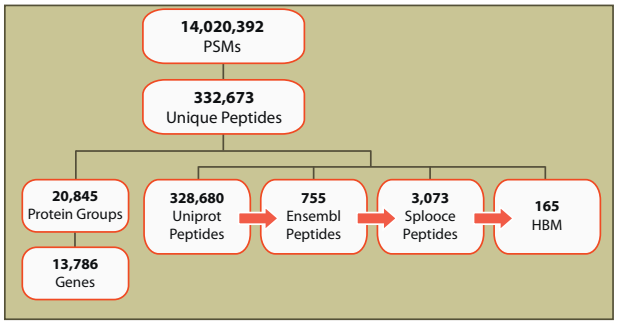

Figure 2. Results overview of the data identified using MasterSet database. Red arrows indicate the peptide numbers identified per source (Uniprot, Ensembl, etc.) as they were prioritized during the construction of the MasterSet database (see the Methods Section).

variants showing unique peptides were identified and then mapped to the reference human genome using the Exonerate aligner whose algorithm is able to identify exon/exon borders. All information related to the alignment itself and the source data were stored and compressed to be used with Proteogenomics Viewer interface.

Proteogenomics Viewer interface

Proteogenomics Viewer is divided in two sections, named Control and Viewer (Fig. 3, panels A and B, respectively). In a gene centric manner, the user can load specific genes using the Control section, which will be loaded and visualized in the Viewer section. The Control section also allows the selection of specific peptides identified for the protein translated by the gene, and the visualization of the MS/MS spectrum and the quantitation of the selected peptide for all samples under analysis. Control also allows filtering of all identified peptides under an error probability score (as calculated by the search engine). Figure 4 illustrates these for peptide QNESLER from the protein GFAP, a peptide only possible when the border of exon 2 joints with the border of exon 6 (see also Supporting Information Fig. S1). By clicking on Peptide Spectrum, a graphical representation of the annotated MS2 spectrum of the best scoring PSM for that peptide will be shown (Fig. 4A). The function Peptide Quantitation loads a table listing all tissues/samples where that peptide was identified. A colored box indicating a normalized abundance level for the peptide follows each tissue name as well as a density plot showing abundance data distribution for that peptide used to calculate the median (see the Methods Section for more details) (Fig. 4B). In this particular case, this peptide characterizing a GFAP isoform is highly abundant in fetal brain. Furthermore, this section also shows the gene names where the peptide is present indicating how unique is the peptide/gene assignment (lower part of Fig. 4B).

Characterizing the loss of targeting signals and transmembrane domains

An important feature that is related to Proteogenomics Viewer implementation is sequence indexing. This allows further implementations regarding, for example sequence domain analysis. To illustrate this, we collected the identified Splooce peptide alignments characterizing newer protein isoforms, retrieved their complete sequence, performed a TMHMM prediction of signal sequences and transmembrane domains, and compared the data to the same predictions against the reference Uniprot proteins. After manually inspecting individual peptide identifications from Splooce entries with loss of domains, we detected with high confidence 58 Splooce isoforms (from 49 genes) (Supporting Information Table S1). From those, 32 represented exon skipping (including those resulting in shorter protein isoforms with alternative TSS – such as shown in Fig. 5), 11 alternative 3′ splicing site, 8 intron retention, 4 alternative 5′ splicing site, and 3 with both alternative 3′ and 5′ splicing sites. Figure 5 illustrates genome alignment and peptide evidence for two shorter isoforms, with alternative TSS, for gene/protein PDIA4. The signal sequence domain is located at position 1–24 in the Uniprot sequence, while the two other sequences from Splooce start at either position 279 or position 488 in the Uniprot reference. The identifications of those N-terminal peptides are further validated not only by the good scoring of the Met containing peptides (MIEQSGPPSK and MEPEEFDSDTLR, both non-tryptical in the reference sequence), but also by the identification of peptides (IEQSGPPSK and EPEEFDSDTLR) where the Met was cleaved, indicating N-terminal methionine amino peptidase activity, a common feature observed in peptides located at the protein N-terminal. Although the evidence presented in Fig. 5 strongly suggests that alternative splicing mediated the use of alternative start codon, we cannot rule out the possibility of alternative translation (not mediated by alternative splicing) as the cause of such pattern.

Discussion

There is a lack of integration between MS/PSM data visualization and peptide/genomic alignment visualization. Currently, data visualization for MS data are largely available and is mostly offered by the search engines themselves, to help users to check their own MS2 identifications post data processing. For external researchers to check data from the original dataset, however, they are required to have access to both original raw files as well as processed files from the search engine that was used. Processed files are not always available in public repositories, therefore data visualization is not always straightforward. Furthermore, many of the most used search engines available require commercial licenses. Public data visualization tools based on published MS2 outputs that do not require in house analysis is less available, and mostly developed to fulfill publication requirements and are made public to demonstrate data from specific publications, such as ProteomicsDB [13].

To our knowledge, most of the proteogenomics tools available (as found in https://omicstools.com/proteogenomics-category, e.g.) are mostly designed to align peptide sequence (text input) to genomic sequences. Notable exceptions are tools like PGTools [30], PG Nexus [31], iPiG [32], proBAMsuite [33], and other pipelines (for example, the one described in [34]), which were designed not only to

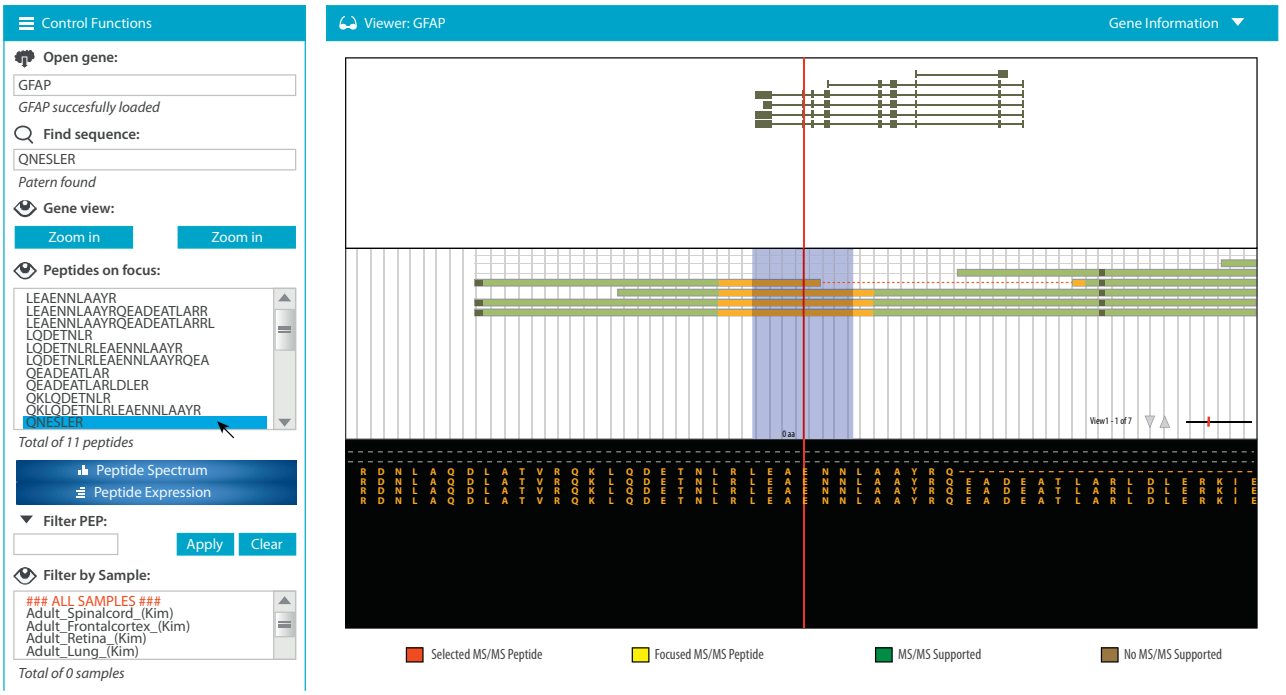

Figure 3. Proteogenomics Viewer Interface. The interface is divided basically into two sections: the Control Functions (**A**) where specific genes or peptide sequences can be searched, peptide sequence information is loaded and MS2 spectra, peptide quantitation and scoring/tissue filters can be applied; the Viewer section (**B**) where the alignment and graphical representation of the genome structure, the protein structure coverage and the protein sequence, in that order, are shown.

provide a genome browser-based visualization of MS-based peptide identifications, but in some cases also to provide scripts for data processing, peptide search, FDR calculations, among others, starting from complete datasets of raw MS files or PSM inputs, on a truly global proteomic approach. However, in our opinion some of such tools decreases the user freedom to apply search engines and approaches that they prefer rather than the one(s) employed at the referred tools (in addition, as we described above, PG Nexus, e.g. uses Mascot engine, which requires a commercial license). We also argue that, a tool that can process MS data post-engine search is more suitable for meta-analysis of different projects from different sources. Finally, we believe that these tools do not allow for MS inspection and does not report critical peptide identifications features (scoring, sequence uniqueness, etc.), which we argue are a major drawback for several issues recently observed in proteogenomics projects, as we discuss below.

Arguably, one of the best and most complete MS2 peptide identification repositories is the PeptideAtlas Project [16, 17], which catalog and report identified peptides from thousands of publications and is often reviewed and updated. Proteins can be individually searched and when identified by a proteomics study, the Atlas reports tissues/samples where the protein was observed, protein sequence coverage based on identified peptides, frequency of each peptide based on number of studies that independently identified it, and protein quantitation. It has been shown that such peptide frequency reports were instrumental to facilitate targeted proteomic studies using methods such as SRM and MRM, because PeptideAtlas can show beforehand what are the most common and easily detected peptides from a specific protein, which can then be used as targets for further detection and quantitation [35, 36].

However, it is also clear that datasets re-analysis can add further issues that can limit how the PeptideAtlas operate. For example, it is intriguing the presence of non-tryptic peptides (unspecific cleavages) reported in the PeptideAtlas, some even at very high frequencies, from MS data collected from samples where authors had clearly claimed the use of high-quality trypsin, which does not produce peptides with unspecific cleavage sites [37], and the use of buffers containing protease inhibitors during protein extraction, to avoid unspecific cleavage from enzymes present in the sample. Furthermore, it is well known and expected that search engine merging strategies have higher FDRs than reported [2], as a penalty for improved identification coverage [38]. Finally, since as now the PeptideAtlas only catalogs proteins that are annotated in UniprotKB or Ensembl, it does not investigate novel sequences so far. All these issues encouraged us to design a simplified tool aimed specifically at genomic alignment analysis. As the PeptideAtlas, we were equally careful to provide our own analysis of the raw MS data, rather than only merging and reporting what the original researchers provided, to minimize FDR. Other tools have implemented similar harmonization approaches successfully, such as SpliceVista [39]. Similarly to our Viewer, SpliceVista was built to process PSM information

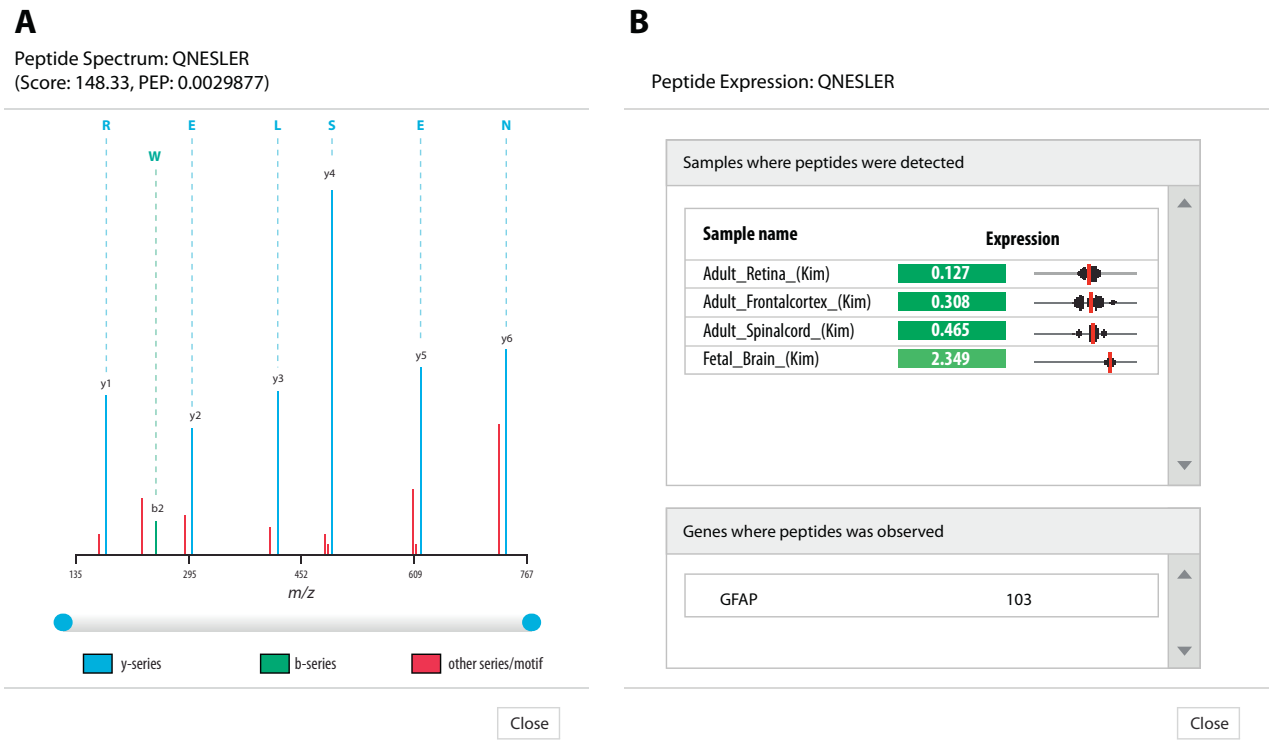

Figure 4. A: Visualization of MS2 spectrum. Once a peptide is selected in the "Peptide on Focus" area, the MS2 spectrum of the best scoring PSM that lead to that peptide identification can be viewed by clicking the action "Peptide Spectrum". A graphical representation of the annotated ion fragments is shown including the search engine scoring and the error probability. The peptide QNESLER characterizes an exon skipping detected by the Splooce tool in the protein GFAP. **B:** Peptide expression and uniqueness. A selected peptide can also have its quantitative data checked by the action "Peptide Expression". The upper panel reports the samples in which that peptide was identified and at which intensity based on the area under curve calculations performed at the MS raw data. Due to the different nature of the fractionation used by all datasets, all peptide quantitation (including all PSM detected for a given peptide) was normalized using a *z*-score approach, which is reported in two manners: to the right, a median of all PSM *z*-scores is numerically given, green boxes representing scores that are in the upper end of the data distribution, and in red (not shown in this example) in the lower end of the distribution; to the right is a density boxplot of all PSM *z*-scores used to calculate the median. The lower panel shows all genes/proteins where the peptide exists as a tryptic/LysC peptide, and the total number of PSMs identified for that protein. In this example QNESLER can only be identified as tryptic for protein entries in the database that align to the gene GFAP.

from a search engine and to map them into exon-intron structure based on transcript data present at the Evidence Viewer Database [40]. A major difference between our Viewer and SpliceVista is that SpliceVista has a strong quantitative feature based on its PQPQ suit [41], which, while very powerful, limits to our knowledge the analysis of splicing events found within single projects. Its data visualization as well is only limited to the gene structure, similarly to our Viewer section. We had, on the other hand, focused in descriptive and qualitative discovery of ASEs, which allows Proteogenomics Viewer to act as a repository tool by combining the analysis of different projects and updated with splicing information, similarly to the PeptideAtlas. Our visualization is also focused on MS2 data interpretation to make easier for users to assess the confidence level of the identification of a peptide that characterizes a known or new isoform.

The recent publications of the human proteome draft [12, 13], and some of the issues raised by others regarding the peptide identification of non-coding regions [14] made it clear to the community that there is a need to establish guidelines to report data and to design tools that can help proteogenomic researchers to visualize and validate novel identifications in a manner not only restricted to FDR thresholds from the search engine used. By developing Proteogenomics Viewer, we not only aimed to fulfill that demand, but we have also created a tool that will allow the development of a repository where MS-based peptide identification can be processed and aligned to genomic data in order to report possible new splicing events.

We had the concern to implement into Proteogenomics Viewer certain features to comply to proteogenomic guidelines which were suggested by Nesvizhskii [10], including: i) a database in FASTA format is made available in this publication and also at Proteogenomics Viewer web address; ii) our approach not only query the novel peptide identified to reference database(s), but the Viewer option allows direct comparison with the reference protein(s) (and not only by reporting the accession number of the reference as suggested in the guideline); iii) MS2 spectra annotation is provided, meaning that the quality of the identification can be assessed without the need to trust FDR thresholds (even though the approach still performs the whole search under a

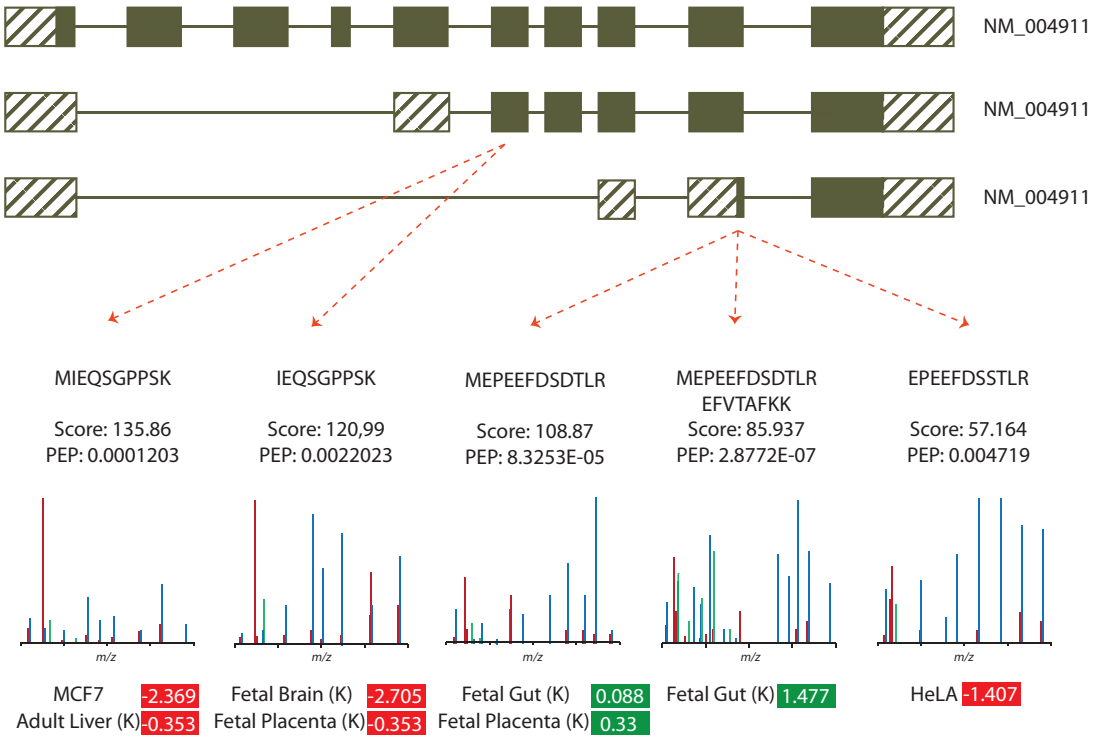

Figure 5. Loss of targeting signal in Splooce entries identified by MS. A graphic representation of splicing variants detected by Splooce for gene/protein PDIA4 (NM004911), with exon/intron structure shown. Dashed areas represent non-coding regions of the cDNA, and filled areas represent coding regions. The translated protein P13667 from Uniprot reference sequence posses a signal sequence from amino acids 1–24. Two isoforms detected by Splooce align to internal regions of the reference Uniprot entry, characterizing exon skipping events that lead to alternative N-terminal sites. This is strongly supported by the fact that both exon skipping events generate a frameshift in case the original Met is used. Although unlikely, alternative translation (not due to alternative splicing) cannot be ruled out. The MS2 spectra, their scoring and expression profile of all five unique tryptic peptides that lead to the identification of those Splooce entries are also shown. (K) means the tissue belongs to the dataset from [12].

FDR threshold nonetheless); and iv) peptides are clearly marked regarding its uniqueness or whether they map to multiple genome locations. The only guideline suggestion that we have not implemented so far is the check for possible isobaric false positives. For example, the novel peptide might be a false positive if it can be explained by the fact that the identification is actually a known peptide with a post-translational modification with the same mass shift and that was not considered during the database search.

Much of such guidelines were created to prevent positive peptide/protein identifications from issues that could have been easily avoided. The publication of the two human proteome drafts clearly illustrated that proteins could be wrongly identified due to protein inference errors, for example a “rare” protein being identified by a tryptic peptide that is not unique to the database and is shared with another entry from a more commonly observed protein that was also identified. Our data analysis showed that many of such errors were avoided in our dataset, since many of the false-positives detected later by independent authors were largely absent in our re-analysis of the data. For example, Kim and co-workers [12] and Wilhelm and co-workers [13] had identified 108 and 200 olfactory receptors, respectively, even though none of the two articles had analyzed tissues from the nasal airway. Our re-analysis of the MS data had identified only a tiny fraction of that, seven receptors, and those identifications were easily verified as false-positives. We argue that, if the articles reporting the drafts of the human proteome had access to a tool such as Proteogenomics Viewer, some if not all of the incorrect protein assignments regarding olfactory receptors on such samples could had been easily avoided.

A further feature, tackled with Proteogenomics Viewer, is MS-based peptide indexing. Indexing sequence information is critical to allow correlative investigation, and while it has been fully explored in nucleic acid datasets since the initial releases of the human genome [42, 43], it is rarely executed in proteomics datasets. Therefore, we provide with this manuscript the indexing of all peptides identified in our analysis and exemplified its use by investigating the loss of signal sequence and transmembrane domains in identified isoforms present in Splooce when compared to the Uniprot reference. The majority of the 58 isoforms characterized (43 isoforms) (one of which is shown in Fig. 5) represented shortened protein isoforms due to alteration of the N-terminal region, resulting in the loss of the signal sequences.

Finally, providing peptide-centric relative quantitation per se is straightforward since it can be directly retrieved from AUC measurements provided by the search engine.

However, the lack of a standardized sample preparation effort and the difference of instrument performance between labs (even when similar instruments are used) still pose a challenge for sample comparison at peptide level. This can be partly explained by the fact that all datasets used in our analysis had a striking difference of performance. While the human proteome drafts had identified on average 3,000–5,000 proteins per tissue, the cell lines characterized [21] had on average almost 9,000 proteins identified per cell type. However, due to the fact that the proteome drafts analyzed far more unique samples, their combined genomic coverage was very similar to the combined cell line dataset coverage (around 11,500 to 12,000 proteins per dataset, from which most of those are shared between them). So there is a discrepancy between low identification coverage per sample versus high coverage of sample types, which resulted in different datasets that, while "similar" in a numerical sense regarding proteins identified, had actually a lower coverage of peptides identified in all of them (see Supporting Information Fig. S3). This is obviously a challenge for the future implementation of proper meta-analysis approaches to be performed in proteomics projects in general and yet to become available at Proteogenomics Viewer.

Conclusion

There has been an increase in the use of probabilistic-based MS peptide identifications to detect novel protein isoforms and protein products from non-coding regions. It has been shown, however, that such novel sequences are more prone to be incorrectly identified. We developed a tool that not only align and index peptide and genomic information, but also provides easily accessible peptide identification quality features for user validation. Such a tool can process multiple datasets collected from independent projects, regardless if using different search engine strategies. Normalization of the peptide quantitative data from independent datasets will allow future implementation of meta-analysis approaches.

Acknowledgment

We would like to thank Gisele Tomazella for helpful discussions and for providing assistance with figure design.

Funding

This work was supported by a grant (23038.004629/2014-19) to SJS and GAS from the Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior (CAPES); grant (305233/2015-7) from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) to SJS. Support from the Ludwig Institute for Cancer Research to SJS is also acknowledged.

The authors declare that there are no conflicts of interest.

References

1. **Aebersold R, Mann M**. 2003. Mass spectrometry-based proteomics. *Nature* **422**: 198–207.
2. **Reiter L, Claassen M, Schrimpf SP, Jovanovic M**, et al. 2009. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol Cell Proteomics* **8**: 2405–17.
3. **Nesvizhskii AI**. 2010. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics* **73**: 2092–123.
4. **Brosch M, Saunders GI, Frankish A, Collins MO**, et al. 2011. Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome. *Genome Res* **21**: 756–67.
5. **Severing EI, van Dijk AD, van Ham RC**. 2011. Assessing the contribution of alternative splicing to proteome diversity in *Arabidopsis thaliana* using proteomics data. *BMC Plant Biol* **11**: 82.
6. **Sheynkman GM, Shortreed MR, Frey BL, Smith LM**. 2013. Discovery and mass spectrometric analysis of novel splice-junction peptides using RNA-Seq. *Mol Cell Proteomics* **12**: 2341–53.
7. **Tress ML, Bodenmiller B, Aebersold R, Valencia A**. 2008. Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biol* **9**: R162.
8. **Blakeley P, Overton IM, Hubbard SJ**. 2012. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J Proteome Res* **11**: 5221–34.
9. **Zhang K, Fu Y, Zeng WF, He K**, et al. 2015. A note on the false discovery rate of novel peptides in proteogenomics. *Bioinformatics* **31**: 3249–53.
10. **Nesvizhskii AI**. 2014. Proteogenomics: concepts, applications and computational strategies. *Nat Methods* **11**: 1114–25.
11. **Wright JC, Mudge J, Weisser H, Barzine MP**, et al. 2016. Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nat Commun* **7**: 11778.
12. **Kim MS, Pinto SM, Getnet D, Nirujogi RS**, et al. 2014. A draft map of the human proteome. *Nature* **509**: 575–81.
13. **Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A**, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* **509**: 582–7.
14. **Ezkurdia I, Calvo E, Del Pozo A, Vazquez J**, et al. 2015. The potential clinical impact of the release of two drafts of the human proteome. *Exp Rev Proteomics* **12**: 579–93.
15. **Ezkurdia I, Vazquez J, Valencia A, Tress M**. 2014. Analyzing the first drafts of the human proteome. *J Proteome Res* **13**: 3854–5.
16. **Deutsch EW, Sun Z, Campbell D, Kusebauch U**, et al. 2015. State of the human proteome in 2014/2015 As viewed through PeptideAtlas: enhancing accuracy and coverage through the AtlasProphet. *J Proteome Res* **14**: 3461–73.
17. **Desiere F, Deutsch EW, King NL, Nesvizhskii AI**, et al. 2006. The PeptideAtlas project. *Nucleic Acids Res* **34**: D655–8.
18. **Khatun J, Yu Y, Wrobel JA, Risk BA**, et al. 2013. Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions. *BMC Genomics* **14**: 141.
19. **Perez-Riverol Y, Alpi E, Wang R, Hermjakob H**, et al. 2015. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* **15**: 930–49.
20. **Vizcaino JA, Deutsch EW, Wang R, Csordas A**, et al. 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* **32**: 223–6.
21. **Geiger T, Wehner A, Schaab C, Cox J**, et al. 2012. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* **11**: M111 014050.
22. **Kroll JE, Galante PA, Ohara DT, Navarro FC**, et al. 2012. SPLOOCE: a new portal for the analysis of human splicing variants. *RNA Biol* **9**: 1339–43.
23. **Cox J, Mann M**. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**: 1367–72.
24. **Cox J, Neuhauser N, Michalski A, Scheltema RA**, et al. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**: 1794–805.
25. **Trapnell C, Pachter L, Salzberg SL**. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–11.
26. **Trapnell C, Williams BA, Pertea G, Mortazavi A**, et al. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511–5.

27. **Kroll JE, de Souza SJ, de Souza GA**. 2014. Identification of rare alternative splicing events in MS/MS data reveals a significant fraction of alternative translation initiation sites. *PeerJ* **2**: e673.
28. **Slater GS, Birney E**. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.
29. **Krogh A, Larsson B, von Heijne G, Sonnhammer EL**. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**: 567–80.
30. **Nagaraj SH, Waddell N, Madugundu AK, Wood S**, et al. 2015. PGTools: a software suite for proteogenomic data analysis and visualization. *J Proteome Res* **14**: 2255–66.
31. **Pang CN, Tay AP, Aya C, Twine NA**, et al. 2014. Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. *J Proteome Res* **13**: 84–98.
32. **Kuhring M, Renard BY**. 2012. IPiG: integrating peptide spectrum matches into genome browser visualizations. *PLoS ONE* **7**: e50246.
33. **Wang X, Slebos RJ, Chambers MC, Tabb DL**, et al. 2016. ProBAMsuite, a bioinformatics framework for genome-based representation and analysis of proteomics data. *Mol Cell Proteomics* **15**: 1164–75.
34. **Weisser H, Wright JC, Mudge JM, Gutenbrunner P**, et al. 2016. Flexible data analysis pipeline for high-confidence proteogenomics. *J Proteome Res* **15**: 4686–95.
35. **Deutsch EW, Lam H, Aebersold R**. 2008. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* **9**: 429–34.
36. **Farrah T, Deutsch EW, Kreisberg R, Sun Z**, et al. 2012. PASSEL: the PeptideAtlas SRMexperiment library. *Proteomics* **12**: 1170–5.
37. **Olsen JV, Ong SE, Mann M**. 2004. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics* **3**: 608–14.
38. **Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW**. 2013. Combining results of multiple search engines in proteomics. *Mol Cell Proteomics* **12**: 2383–93.
39. **Zhu Y, Hultin-Rosenberg L, Forshed J, Branca RM**, et al. 2014. SpliceVista, a tool for splice variant identification and visualization in shotgun proteomics data. *Mol Cell Proteomics* **13**: 1552–62.
40. **Wang ET, Sandberg R, Luo S, Khrebtukova I**, et al. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–6.
41. **Forshed J, Pernemalm M, Tan CS, Lindberg M**, et al. 2008. Proteomic data analysis workflow for discovery of candidate biomarker peaks predictive of clinical outcome for patients with acute myeloid leukemia. *J Proteome Res* **7**: 2332–41.
42. **Yuan J, Liu Y, Wang Y, Xie G**, et al. 2001. Genome analysis with gene-indexing databases. *Pharmacol Ther* **91**: 115–32.
43. **Sadakane K, Shibuya T**. 2001. Indexing huge genome sequences for solving various problems. *Genome Inform* **12**: 175–83.