Bioinformatics, 36(1), 2020, 205–211
doi: 10.1093/bioinformatics/btz530
Advance Access Publication Date: 27 June 2019
Original Paper

OXFORD

Gene expression

MiTPeptideDB: a proteogenomic resource for the discovery of novel peptides

Elizabeth Guruceaga[1,2,†], Alba Garin-Muga[3,4,†] and Victor Segura [1,2,*]

[1]Bioinformatics Platform, Center for Applied Medical Research, University of Navarra, Pamplona 31008, Spain, [2]IdiSNA, Navarra Institute for Health Research, Pamplona 31008, Spain, [3]eHealth and Biomedical Applications Department, Vicomtech, San Sebastian 20009, Spain and [4]Biodonostia Health Research Institute, (Bioengineering Area), eHealth Group, San Sebastian 20014, Spain

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Janet Kelso

Received on January 14, 2019; revised on May 17, 2019; editorial decision on June 21, 2019; accepted on June 25, 2019

Abstract

Motivation: The principal lines of research in MS/MS based Proteomics have been directed toward the molecular characterization of the proteins including their biological functions and their implications in human diseases. Recent advances in this field have also allowed the first attempts to apply these techniques to the clinical practice. Nowadays, the main progress in Computational Proteomics is based on the integration of genomic, transcriptomic and proteomic experimental data, what is known as Proteogenomics. This methodology is being especially useful for the discovery of new clinical biomarkers, small open reading frames and microproteins, although their validation is still challenging.

Results: We detected novel peptides following a proteogenomic workflow based on the MiTranscriptome human assembly and shotgun experiments. The annotation approach generated three custom databases with the corresponding peptides of known and novel transcripts of both protein coding genes and non-coding genes. In addition, we used a peptide detectability filter to improve the computational performance of the proteomic searches, the statistical analysis and the robustness of the results. These innovative additional filters are specially relevant when noisy next generation sequencing experiments are used to generate the databases. This resource, MiTPeptideDB, was validated using 43 cell lines for which RNA-Seq experiments and shotgun experiments were available.

Availability and implementation: MiTPeptideDB is available at http://bit.ly/MiTPeptideDB.

Contact: vsegura@unav.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the last few years, the field of Proteomics has become a promising approach in Molecular Biology and precision medicine due to its suitability for the study of proteins and their implication in biology and disease. The widespread use of human high-throughput proteome studies has been possible mainly thanks to the advances made in mass spectrometry (MS/MS-based proteomics) (Nilsson *et al.*, 2010). The new specifications of the high resolution mass spectrometers allow the identification and quantification of the proteins in a given biological sample with a high coverage based on the huge number of high quality spectra detected by these instruments in a single run (Nagaraj *et al.*, 2011). Although the dimensionality of the datasets obtained is still lower than the ones obtained in Genomics and Transcriptomics using both microarrays and Next Generation Sequencing (NGS) technologies, the analyses of the shotgun experiments performed in the last years have achieved results that are worth highlighting. The characterization of proteomes (Kim *et al.*, 2014; Wilhelm *et al.*, 2014), the identification of new clinical biomarkers (Halvey *et al.*, 2014) and the discovery of novel functional biological elements as the small open reading frames (smORFs) (Couso and Patraquim, 2017) and the microproteins (Zhang *et al.*, 2017) have increased our knowledge about the molecular mechanisms of human diseases (Tamborero *et al.*, 2013; Zhang *et al.*, 2014).

One of the main driving forces behind the advances in Proteomics in the last decade has been the Human Proteome Project (HPP) (Legrain *et al.*, 2011), an international project launched in 2010 and supported by the Human Proteome Organization (HUPO). The essential goal of this project is the deciphering of the complete human proteome, an ambitious task divided into two initiatives: the C-HPP (chromosome-based HPP) and the BD-HPP (biology and disease HPP). The first one is responsible for the detection of all the human proteins and their functional characterization using MS/MS or antibody technologies (Paik *et al.*, 2012a), while the latter is responsible for studying the implications of the proteins in the cellular processes and the human diseases (Lam *et al.*, 2016). The integration of Genomics, Transcriptomics and Proteomics has been widely used by the HPP groups to achieve these goals (Tabas-Madrid *et al.*, 2015). This proteogenomic approach has also been applied to detect peptides that are not present in the proteomic reference databases and therefore the generation of custom databases is required (Nesvizhskii, 2014; Zhu *et al.*, 2018). This is the case of single aminoacid polymorphisms (SAPs) (Garin-Muga *et al.*, 2016; Zhang *et al.*, 2014) and novel peptides, smORFs or microproteins derived from non-coding genes or novel genes detected by NGS experiments (Choi *et al.*, 2018; Li *et al.*, 2018).

Cancer is one of the priority research areas of the BD-HPP, and the cancer transcriptome has been studied in detail for cataloguing all the transcripts expressed in this disease. The MiTranscriptome consensus human transcriptome (Iyer *et al.*, 2015) was obtained from the analysis of 7256 RNA-Seq experiments and it was key to associate lncRNA transcription with carcinogenesis. The complete catalog contains 384 066 different transcripts, some of them are not described in other databases as GenBank, Ensembl or GENCODE. In the case of known lncRNAs, it has been noted that they can be associated with ribosomes and the fact that many of them appear to have arisen relatively recently in evolution indicates that they could be an important source of new peptides (Ruiz-Orera *et al.*, 2014). This fact brings new opportunities for the research field of Proteogenomics (Choi *et al.*, 2018; Li *et al.*, 2018).

In this manuscript we present MiTPeptideDB as a resource for the detection of novel peptides using a proteogenomic approach. First, custom databases of these peptides were generated with the predicted translations of known lncRNAs and novel transcripts annotated using GENCODE and MiTranscriptome assembly. Then, a peptide detectability study was included to filter those peptides with a low probability of being detected by MS. This step is one of the novelties of our approach being especially useful to decrease the size and increase the quality of the custom databases derived from RNA-Seq experiments instead of RIBO-Seq experiments, in which only translated transcripts are sequenced (Choi *et al.*, 2018). Finally, a bioinformatic pipeline was proposed and applied to a set of public shotgun experiments of cell lines obtained from the NCI60 project demonstrating the feasibility of our method. The expression and function of the detected peptides must be subsequently validated using the proper experiments of molecular biology.

2 Materials and methods

2.1 Bioinformatic workflow

We developed a bioinformatic pipeline to detect novel peptides in two stages. First, proteomic databases were generated based on the transcripts described in the human MiTranscriptome assembly (Iyer *et al.*, 2015) using a proteogenomic approach including for the first time the filtering of peptides based on their detection probability by MS (Fig. 1A). After the creation of the databases, we analyzed some of the shotgun experiments available in the NCI60 project corresponding to 43 cell lines (Fig. 1B). The analyses of the shotgun experiments were performed using three sequential proteomic searches with Mascot search engine and removing the assigned spectra from the datasets before each new search. This approach can be easily generalized to include additional search engines.

2.2 MiTranscriptome annotation and generation of custom proteomic databases

The MiTranscriptome initiative represents a great computational and biological effort to describe for the first time the whole complexity of the human transcriptome, especially in the case of cancer samples. In this study, a total of 25 datasets with 7256 poly(A) +

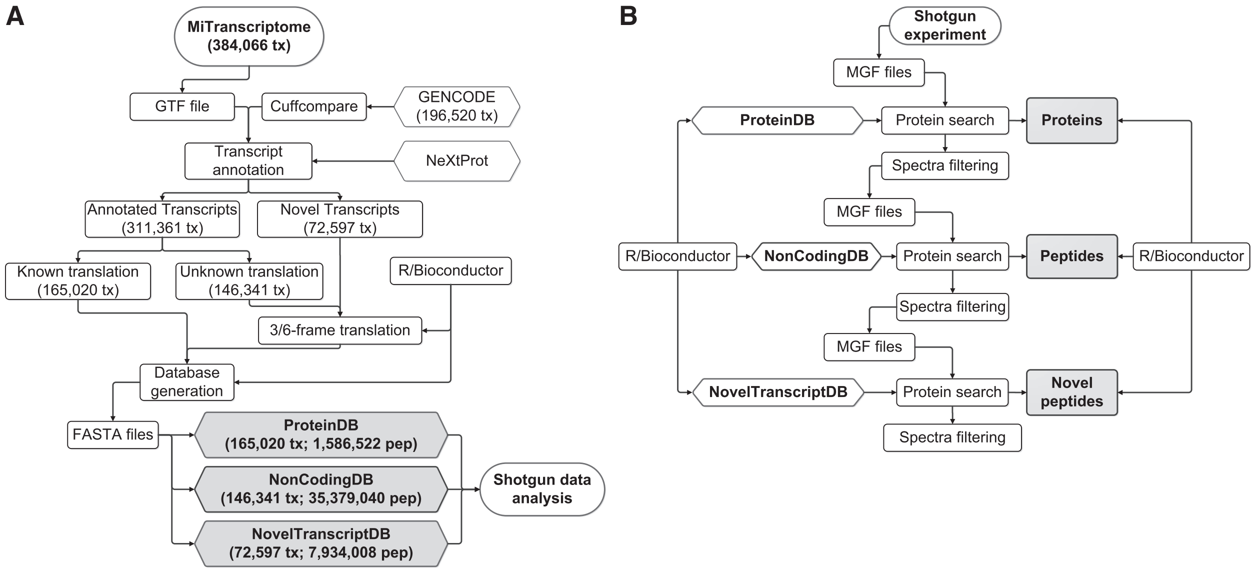

Fig. 1. Bioinformatic workflow developed for the identification of novel peptides in cancer using the MiTranscriptome human assembly. (**A**) Generation of custom proteogenomic databases. The number of transcripts (tx) and peptides (pep) are shown. (**B**) Shotgun data analysis

RNA-Seq samples from different resources (including the TCGA and the ENCODE projects among other public datasets) were processed to define a set of 6503 high quality experiments. As a result, the MiTranscriptome human assembly contains 384 066 predicted transcripts that are available as a GTF file in its web page (http://mitranscriptome.org/).

Once we downloaded the transcripts of this assembly, we first annotated them in order to identify the already known transcripts corresponding to both protein coding and non-coding genes. For this purpose, we used the GENCODE version 19 annotation of the human genome that includes 57 820 genes (196 520 transcripts), 20 345 of them being protein coding genes (81 814 transcripts). The annotation was performed with Cuffcompare software from the Cufflinks analysis suite and a total of 311 361 transcripts were identified. However, only 165 020 of these transcripts had a known amino acid sequence in GENCODE database. This analysis approach enabled us to define three different FASTA files for the proteomic searches: (i) **ProteinDB**, a database with 1 586 522 peptide sequences obtained from GENCODE representing 165 020 transcripts of MiTranscriptome; (ii) **NonCodingDB**, a database with information about 146 341 transcripts of MiTranscriptome annotated using GENCODE, whose peptide sequences (35 379 040 entries) were obtained using the 3 or 6-frame translation method (Castellana and Bafna, 2010) taking their genome sequence as input; (iii) **NovelTranscriptDB**, a database of transcripts of the MiTranscriptome human assembly that were not present in GENCODE. In this case the sequences of the 72 597 transcripts were also obtained using the previously mentioned 3 or 6-frame translation algorithm depending if we knew the strand of the transcript or not (7 934 008 peptide sequences). All the entries of the FASTA files correspond to peptides of between 9 and 30 amino acids, obtained using the *in silico* digestion of Proteogest software (Cagney *et al.*, 2003) for the proteins with at least one tryptic peptide. We applied the standard rules of cleavage for trypsin enzyme digestion and allowed oxidation of methionine and one missed cleavage. Finally, all the amino acid sequences in NonCodingDB that matched a sequence in ProteinDB were removed from NonCodingDB, and all the amino acid sequences in NovelDB that matched a sequence in NonCodingDB were removed from NovelDB. These FASTA files and their annotation are part of MiTPeptideDB and can be downloaded from http://bit.ly/MiTPeptideDB. In addition, we have included the pipeline used for the generation of the FASTA files (Supplementary Material S1).

2.3 Database filtering using peptide detectability

One of the main statistical problems to deal with in proteogenomic studies is the inevitable increased size of the custom databases when the 6-frame translation is used to infer the amino acid sequences of the peptides derived from the genomic or transcriptomic experiments. The main effect of searching MS spectra against these huge databases is the difficulty estimating the False Discovery Rate (FDR) and, therefore, the high probability of false spectra assignments to peptides that are not present in the sample under study. The identification and removal of the false positive peptide identifications is particularly challenging when the sequence of novel peptides is predicted based on RNA-Seq experiments (Choi *et al.*, 2018; Olexiouk *et al.*, 2018) due to the higher level of transcriptional noise in the signal and the identification of non-translated transcripts. Besides, the rate of validation of these findings is very low (Choi *et al.*, 2018; Couso and Patraquim, 2017; Samandi *et al.*, 2017) although their implications in key biological functions and in the development of diseases have been experimentally proven (Olexiouk *et al.*, 2018; Samandi *et al.*, 2017; Zhang *et al.*, 2017).

For these reasons, we recommend the use of more astringent FDR thresholds to select the best candidates of these rare events and a new approach to improve the quality of the databases similar to the one previously used in Guruceaga *et al.* (2017). We designed a peptide detectability classifier based on the information about the number of peptide observations stored in GPMDB database (Craig *et al.*, 2004) and the evaluation of more than 550 physicochemical and biochemical properties calculated for each tryptic peptide using *seqinr* R package (Gentleman *et al.*, 2004). First, we filtered these characteristics with a sampling-based t-test to obtain 106 non-redundant properties that were used to generate a Random Forest classifier implemented using *caret* R package (Gentleman *et al.*, 2004). In the case of MiTPeptideDB, we filtered the peptides of the custom databases based on their detectability obtaining custom databases of proteotypic peptides: **FilteredProteinDB** (363 896 entries, 77% smaller than ProteinDB), **FilteredNonCodingDB** (8 240 422 entries, 76.7% smaller than NonCodingDB) and **FilteredNovelDB** (1 816 397 entries, 77.1% smaller than NovelDB). These FASTA files and their annotation are part of MiTPeptideDB and can be downloaded from http://bit.ly/MiTPeptideDB. The R script used for the mean value calculation of peptide properties and the classification of each peptide into detectable or non-detectable peptide (Supplementary Material S2), and the physicochemical and biochemical properties of all the entries for each MiTPeptideDB database (Supplementary Material S3) are also provided.

2.4 Proteomic and transcriptomic datasets

The NCI60 dataset was developed in the late 1980s by the US National Cancer Institute (NCI), and it contained cell lines from nine distinct tumour types. The proteomic experiments from the NCI60 cell lines selected (Supplementary Table S1) were downloaded from the NCI60 database (http://129.187.44.58: 7070/NCI60/main/index) and analyzed using MiTPeptideDB (Supplementary Methods). The transcriptome guided analysis used the experiments of the CCLE project (https://portal.gdc.cancer.gov) corresponding to the 43 cell lines in common with the NCI60 dataset. We processed these data to obtain the normalized expression values of the transcripts defined in MiTranscriptome (Supplementary Methods).

The detection of novel peptides in this cancer cell lines was performed using the two sets of databases previously generated. First, the databases obtained from the MiTranscriptome assembly without any additional processing were used (ProteinDB, NonCodingDB and NovelDB). In this case, when the search database was ProteinDB, the FDR at peptide-spectrum match (PSM), peptide and protein level was calculated using Mayu (Reiter *et al.*, 2009) with a criteria of protein FDR $< 1\%$. In the following searches, against NonCodingDB and NovelDB, only FDR at PSM level was calculated to select those results with PSM FDR $< 0.01\%$. In order to provide high statistical evidence to these unusual and difficult to validate events an extremely low value of FDR was used to select the PSMs assigned to novel peptides. In a second batch of searches, the filtered databases that include the proteotypic peptides predicted by our classifier of peptide detectability (Guruceaga *et al.*, 2017) were used. The FDR of the results obtained with FilteredProteinDB database was calculated at PSM, peptide and protein level using in house scripts in R programming language (Gentleman *et al.*, 2004), and a threshold of protein FDR $< 1\%$ was applied. Then, PSM FDR at 0.01% was fixed for the analysis of FilteredNonCodingDB and FilteredNovelDB results. The R code and data necessary to test MiTPeptideDB are provided in http://bit.ly/MiTPeptideDB (Supplementary Material S4).

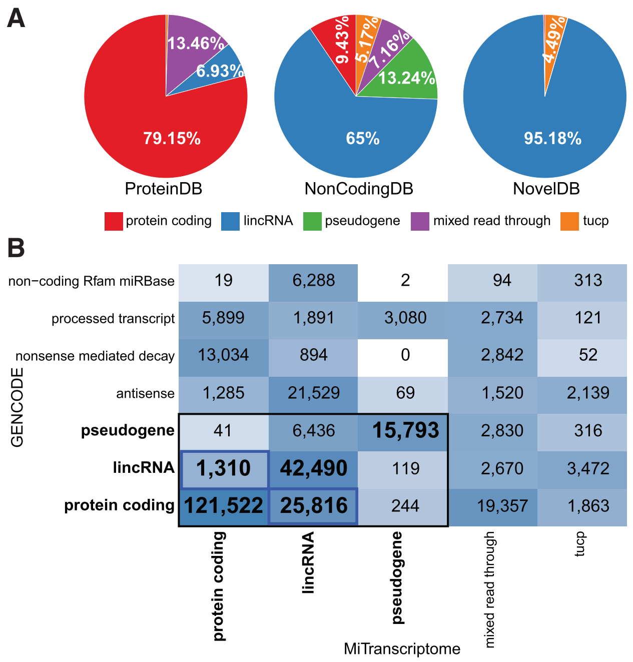

Fig. 2. (**A**) Distribution of the number of transcripts in ProteinDB, NonCodingDB and NovelDB databases as a function of the transcript biotype assigned by MiTranscriptome. Percentages are shown for the biotypes with a greater number of transcripts (>1%). (**B**) Heatmap with the comparison of the biotypes assigned by MiTranscriptome and GENCODE. Common biotypes to both annotation systems are remarked in bold and misclassifications of interest are framed in blue

3 Results and discussion

3.1 Annotation of the MiTranscriptome assembly

The annotation process of the MiTranscriptome data using GENCODE as reference resulted in the generation of three FASTA databases: ProteinDB with peptides of 165 020 transcripts, NonCodingDB with 146 341 and NovelDB with 72 597. Since MiTranscriptome provides its own coding potential study independent to the GENCODE transcript biotype, the transcripts of these databases were classified as: protein coding, lncRNAs, pseudogenes, mixed read through and transcripts of unknown coding potential (tucp). In Figure 2A we represented the distribution of the MiTranscriptome biotypes in each of the generated databases. In the case of ProteinDB we found 130 532 transcripts corresponding to protein coding genes, 22 215 mixed read throughs, 11 444 lncRNAs, 784 tucps and 45 pseudogenes. NonCodingDB was constituted by 95 131 lncRNAs, 19 369 transcripts of pseudogenes, 13 805 protein coding, 10 477 mixed read throughs and 7559 tucps. Finally, the NovelDB database contained 69 096 lncRNAs, 3260 tucps, 188 transcripts of protein coding genes, 30 mixed read through and 23 pseudogenes. It is clear from this biotype distribution that most of the transcripts of MiTranscriptome previously annotated in GENCODE and with known amino acid sequences corresponded to protein coding genes, in NonCodingDB lncRNAs and pseudogenes were the most abundant genes and finally, the transcripts of MiTranscriptome not included in GENCODE (NovelDB) were mainly classified as lncRNAs.

We also compared the global transcript biotypes assignments of MiTranscriptome and GENCODE for all the common transcripts, that are 311 361. This comparison was performed considering the most important categories common to both annotation databases: 'protein coding', 'lncRNA' and 'pseudogenes' (Fig. 2B). In general, we found a great consistency in these assignments with discrepancies in around the 15.89% (33 966 transcripts). The most abundant transcript categories were protein coding with 121 522 common transcripts and lncRNA with 42 490. However the misclassified transcripts between protein coding genes and lncRNAs are of special interest: transcripts defined as protein coding in GENCODE that were considered lncRNAs in MiTranscriptome (25 816 transcripts) and transcripts defined as lncRNAs in GENCODE that were considered protein coding transcripts in MiTranscriptome (1310 transcripts). Both gene sets are framed with a blue border in Figure 2B. Therefore, the expected number of detected peptides in the proteomic searches of shotgun experiments would be higher using our non-canonical bioinformatic pipeline based on a comprehensive transcriptome obtained from thousands of experiments. While we assume that some of the protein coding genes were misclassified as lncRNAs by one of the annotation references, the mentioned transcripts are not considered using standard reference databases. In addition, this novel approach for the detection of new peptides can be easily applied to other human transcriptomes obtained experimentally (Supplementary Material S1).

3.2 Detection of novel peptides in cancer cell lines

The proof of concept for the detection of novel peptides of tumoral origin was conducted in the experiments of the cell lines shared between the CCLE and the NCI60 datasets (Supplementary Table S1). The analysis was initially performed using the complete protein databases obtained after the proteogenomic analysis of MiTranscriptome (Fig. 1) and following the restricted criteria of FDR defined by the HPP guidelines for the identification of proteins using MS/MS experiments (Paik *et al.*, 2012b), and even a more restricted value of FDR for the candidate novel peptides (FDR < 0.01%).

The results for the CCLE RNA-Seq samples are summarized in Supplementary Table S2. A total and mean number of transcripts per database expressed in the samples under study were: 161 536 and 137 912 transcripts, respectively, with a coefficient of variation (CV) of 2.47% for ProteinDB, 127 785 and 70 087 transcripts in the case of NonCodingDB (CV = 8.34%) and 50 143 and 13 729 transcripts for NovelDB (CV = 17.92%). Interestingly, the larger the number of lncRNAs and novel peptides in the considered database, the fewer the number of transcripts expressed in the cell lines and the higher the observed CV. When the results of the protein searches were analyzed we found a similar trend, with more MS/MS detections using ProteinDB and a huge reduction of this number with NovelDB (Supplementary Table S2). However, the number of peptide identifications was more heterogenous across samples and tissues compared with the large degree of uniformity found in the transcriptomes of the same cell lines (Supplementary Fig. S1A–F). A total and mean number of detected ProteinDB peptides were 51 210 and 11 128 (47 800 and 15 695 transcripts with a CV of 21%), 325 and 40 (1239 and 139 transcripts with a CV of 67.82%) of NonCodingDB and 36 and 2 (131 and 7 transcripts with a CV of 88.08%) using NovelDB. The potential novel peptides were identified using the NonCodingDB and NovelDB databases with a total number of 361 detections that comply with a very restrictive FDR threshold given the high level of expected noise in these assignments. However, it is already known that the number of false detections, even with a strict control of the number of false positives, is very high in these cases. Therefore biological validation is required to characterize the function in the cell of the detected novel peptides and their viability as biomarkers (Zhang *et al.*, 2017).

The observed difference in magnitude of the results in the CCLE and the NCI60 datasets is related to a technological issue. In the case of the CCLE dataset RNA-Seq technology has been used and the obtained transcriptome data are quite complete (Supplementary Table S2). It is well known the difficulty in quantifying the low-expressed transcripts but in this kind of experiments we obtain transcriptome data for more than 200 000 transcripts. In contrast, the NCI60 dataset consists of shotgun experiments that identify a proteome of 8000 proteins at most (Supplementary Table S2). Identification process using mass spectrometry is semi-random and undersampling can explain the observed variability (Zhang *et al.*, 2013).

Combining the results of transcriptomic and proteomic experiments it was possible to obtain the expressed transcripts of MiTranscriptome for which any peptide was detected in the shotgun experiment of the same cell line. This transcriptome guided analysis can be a key bioinformatic and biological tool to provide an additional support to the MS/MS identifications obtained from MiTPeptideDB. In Supplementary Figure S1G–I we summarized the results and a lower number of detected peptides that fulfilled this proteogenomic filter can be observed: 50 968 peptides using ProteinDB (46 517 transcripts), 267 from NonCodingDB (891 transcripts) and five for NovelDB (11 transcripts). In terms of mean values of peptides and transcripts per sample we observed 11 040 peptides (14 874 transcripts) in ProteinDB, 32 peptides (110 transcripts) in NonCodingDB and 1 peptide (two transcripts) in NovelDB (Supplementary Table S2). This case study allowed us to identify a total of 272 candidate peptides corresponding to 902 transcripts that could be, after further validation studies, novel peptides with important biological functions and implications in the molecular mechanisms of human diseases (Supplementary Material S5). The number of detected novel peptides is in the range of the results obtained in other studies focused on the identification of smORFs and microproteins (Choi *et al.*, 2018; Samandi *et al.*, 2017), where the authors highlighted the great number of false identifications and the low rate of validations. This can be due to the technical limitations of the MS/MS instruments when dealing with very rare and low expressed peptides, although in the last years the new MS techniques have increased the capacity of Proteomics significantly. These results, despite their biological and clinical interest, have to be properly validated including the expression and tissue specificity of the new transcripts, the coding potential of the novel peptides and their molecular function (Perez-Gracia *et al.*, 2017; Zhang *et al.*, 2014).

Continuing the analysis of these results we tried to evaluate the statistical and biological robustness of the novel peptide identifications. In Figure 3A we showed that the number of identifications decreased according to the coding potential and the annotation quality of the transcripts used to generate the databases and, as expected, the obtained proteome coverage was not as high as the transcriptome coverage. However, the integration of RNA-Seq and MS/MS experiments can be used as a reference to select the best set of novel peptide candidates. The ion score distribution obtained in the Mascot searches for the significant PSMs (FDR $< 1\%$) confirmed that the identification of novel peptides with high ion scores is possible (Fig. 3B). Despite obtaining lower values of this score for NonCodingDB and NovelDB assignments compared with the ProteinDB results, in all the cases this value is higher than 40, a value considered statistically significant in most of the Mascot searches. In addition, the distribution of the MiTranscriptome transcript biotypes for the detected peptides (FDR $< 1\%$) enhanced the credibility of the results (Supplementary Fig. 2A): predominance of peptides derived from protein coding genes using ProteinDB, a mixture of protein coding and lncRNAs with NonCodingDB and predominance of lncRNAs for NovelDB. Using the biotypes of GENCODE the results were very similar, as can be seen in Supplementary Figure 2B for NonCodingDB that is the most complex database due to the presence of coding and non-coding transcripts. Finally, the known higher tissue specificity of the non-coding genes respect to the protein coding genes was also evident in the peptide identifications (Supplementary Fig. 2C–E). A number of detected transcripts shared among different tissues of origin were higher in ProteinDB and smaller in NovelDB, where the peptide identifications were restricted to certain tissues. This fact reinforces our theory that the peptides of NonCodingDB and NovelDB, although fewer in number, could be good novel candidates and deserve further consideration.

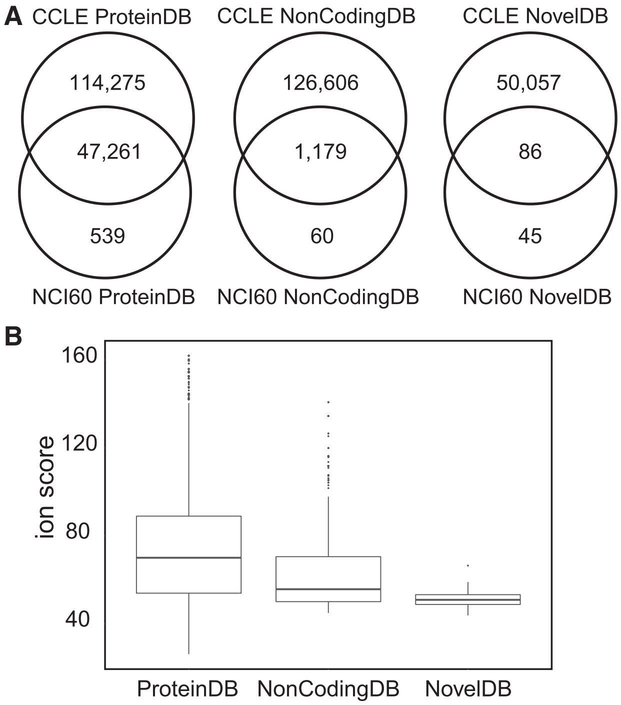

Fig. 3. (**A**) Comparison between the obtained transcriptome and proteome coverages measured in number of transcripts. (**B**) Distribution of the Mascot ion scores for the identified peptides

3.3 Quality improvement in proteogenomic databases using peptide detectability

In order to have fully HPP compliant protein identifications and at the same time avoid ambiguous identifications due to degenerated peptides, we applied a peptide filter based on the peptide uniqueness (Paik *et al.*, 2012b). For that reason, the bioinformatic analysis described in this study is completely suitable for the unambiguous robust statistical detection of novel peptides with the added value of its compatibility with the analysis guidelines of the BD-HPP project. Another important peptide feature that is not considered in the bioinformatic tools currently available in Proteogenomics is peptide detectability, although its importance in high throughput proteomic studies for protein detection and quantification (Li *et al.*, 2010). Peptide detectability of the MiTPeptideDB tryptic peptides was evaluated using a Random Forest classifier (Supplementary Material S2 and S3) obtained using more than 550 peptide properties and trained with the information of the GPMDB database (Guruceaga *et al.*, 2017). Thus, MiTPeptideDB gives the possibility to perform

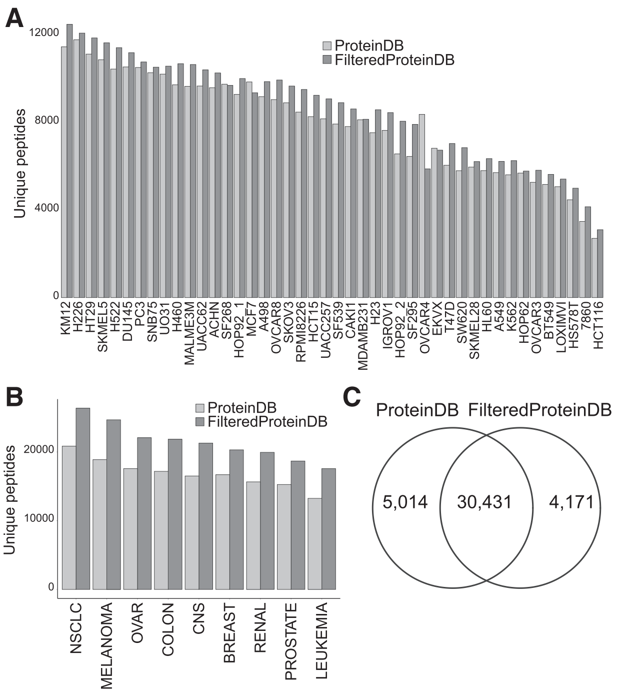

Fig. 4. (**A**) Comparison of the number of detected unique peptides between ProteinDB and FilteredProteinDB per sample and (**B**) tissue of origin. (**C**) Total number of unique peptides detected using ProteinDB and FilteredProteinDB

proteomic analyses using databases in which the peptides with low detection probability by MS/MS are removed. In order to evaluate the differences in the obtained results, the NCI60 proteomic analyses were also performed with the filtered databases: FilteredProteinDB, FilteredNonCodingDB and FilteredNovelDB. The statistical analyses were performed at PSM, peptide and protein level (FDR $<$ 1%) for FilteredProteinDB and only at PSM level for FilteredNonCodingDB and FilteredNovelDB (FDR $<$ 0.01%). In this analysis we identified a total of 26 novel peptides corresponding to 72 transcripts (Supplementary Material S5).

The comparison between the results obtained with the complete and the filtered databases (Fig. 4) is specially interesting when RNA-Seq experiments are used to identify novel peptides. In fact, this database reduction is not only advisable but mandatory in those cases where the decrease of the computational time is necessary or the increase in the accuracy of the FDR calculation is essential to ensure reliable results as occurs with exploratory research, including the definition of new biomarkers and biological findings with low detection probabilities (The *et al.*, 2016). In order to better evaluate the differences only unique peptides were considered in this comparison. As expected, the results using FilteredProteinDB were better in terms of statistical confidence and the observed differences were homogeneous across cell lines (Fig. 4A) and their tissues of origin (Fig. 4B) obtaining a higher number of detected peptides per sample. The mean number of detected unique peptides with FilteredProteinDB per sample was 7% higher than using ProteinDB (Fig. 4C), 26% higher in the case of mean transcripts per sample. We have also analyzed the physicochemical properties that contribute in a higher degree to the detectability of the novel peptides identified in the filtered databases compared with the identifications of the complete databases. The statistically significant differences associated with peptide detectability are related to the ionization capability of the peptides, their complexity, their amino acid composition and their hydrophobicity (Supplementary Table S3), as previously described (Tang *et al.*, 2006). Besides, these candidate novel peptides were searched against the current human assembly (GENCODE 30) using BLASTP (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Considering the 361 novel peptides identified with MiTPeptideDB, 85 would remain being novel while 10 of the 26 detected novel peptides with the filtered databases would still be novel peptides (Supplementary Material S5).

4 Conclusion

Proteogenomics is a promising area of research in several technological and scientific areas, especially in Biology, Biomedicine and, in the last few years, clinical biomarker discovery (Nesvizhskii, 2014) and the detection of functional and regulatory elements such as smORFs and microproteins. In general, the optimal sequencing technology for these analyses is RIBO-Seq. However, considering the huge number of RNA-Seq experiments publicly available, we developed a bioinformatic analysis pipeline capable of handling the high noise level of these datasets. The bases of proteogenomic methods are the creation of custom databases for the proteomic searches and the subsequent statistical analyses for the FDR estimation in the obtained results considering the size effect of these databases (Ansong *et al.*, 2008). Usually, custom databases are generated from NGS experiments using the 3 or 6-frame translation method to infer the protein amino acid sequence of the identified DNA or RNA structures. This approach can be useful for the development of personalized databases in the case of the multiomic study of a single patient of a disease (Garin-Muga *et al.*, 2016).

In this manuscript, we present a new approach which takes as its starting point the MiTranscriptome human assembly (Iyer *et al.*, 2015), a human *de novo* transcriptome assembly based on more than 7000 samples mainly from different tumor types of the TCGA project. We used the transcripts found after the analysis of the RNA-Seq experiments of all these samples to generate different custom databases according to the quality of the transcript annotations and the availability of their amino acid sequence. In addition, we introduced two database filtering steps to reduce the number and improve the quality of these peptides. First, we applied a filter of peptide uniqueness in line with the guidelines proposed by the HPP project, avoiding the ambiguous identifications derived from degenerate peptides. Afterward, we removed those peptides considered non-detectable by MS/MS using a Random Forest classifier that was an innovation in this kind of proteogenomic analyses. In this way, it is possible to increase the computational efficiency of the searches independently of the search engine used and mitigate the effect of their size in the FDR estimation. In fact, the FDR must be even more carefully controlled in the case of the spectra assignments corresponding to candidate novel peptides and FDR $<$ 0.01% was used for these events.

The comprehensive comparison of the MiTransctriptome and GENCODE transcript biotype assignments showed differences between the two annotations. The most interesting ones were those between protein coding and lncRNA biotypes, which suggested the existence of transcripts defined as non-coding in one of the annotation sources that can be considered coding using the other one. The possibility of detecting those peptides was enhanced using our proteogenomic approach.

As a proof of concept, we identified novel peptides in shotgun experiments of different cancer cell lines using MiTPeptideDB with high-throughput experiments from the CCLE and the NCI60 projects. A total of 43 cell lines from different tissue origins were analyzed to determine the expression level of the transcripts from the custom databases and detect proteins and peptides in the shotgun experiments. However, further validation experiments should be designed, for example, using MRM, focused on the common results to both technologies. In this manner, we would increase the likelihood of the peptide detections in a particular cell line. This is especially important in the case of the novel peptides because their characterization is not straightforward although their known implication in biology and diseases. The in-depth study of these findings must include an evaluation of the statistical significance obtained for their detection, their molecular function and specific expression in a certain biological matrix.

In summary, in this manuscript we introduce a bioinformatic workflow for the detection of novel peptides and their proteogenomic analysis. This resource was generated based on a human assembly (MiTranscriptome), a set of custom databases and for the first time a filtering process based on the predicted peptide detectability to deal with the disadvantages of high size databases and noisy RNA-Seq experiments. We analyzed a set of shotgun experiments from cancer cell lines validating the capacity of MiTPeptideDB to deliver good results. We provide the databases in FASTA format, their transcript annotation, the peptide detectability classifier and the code needed to use the bioinformatic pipeline (http://bit.ly/MiTPeptideDB).

Funding

This work was supported by PRBB-ISCIII [PT13/0001/0002], PRB3-ISCIII [PT17/0019/0013], Departamento de Salud of Gobierno de Navarra [33/2015] and Ministerio de Economía y Competitividad [DPI2015-68982-R] cofinanced by FEDER funds. The Bioinformatics Platform of CIMA is member of the ProteoRed-ISCIII platform.

Conflict of Interest: none declared.

References

Ansong,C. *et al.* (2008) Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Brief. Funct. Genomic. Proteomic.*, 7, 50–62.

Cagney,G. *et al.* (2003) In silico proteome analysis to facilitate proteomics experiments using mass spectrometry. *Proteome Sci.*, **1**, 5.

Castellana,N. and Bafna,V. (2010) Proteogenomics to discover the full coding content of genomes: a computational perspective. *J. Proteomics*, **73**, 2124–2135.

Choi,S.W. *et al.* (2018) The small peptide world in long noncoding RNAs. *Brief. Bioinformatics*, bby055.

Couso,J.P. and Patraquim,P. (2017) Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.*, **18**, 575–589.

Craig,R. *et al.* (2004) Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.*, **3**, 1234–1242.

Garin-Muga,A. *et al.* (2016) Proteogenomic analysis of single amino acid polymorphisms in cancer research. *Adv. Exp. Med. Biol.*, **926**, 93–113.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Guruceaga,E. *et al.* (2017) Enhanced missing proteins detection in NCI60 cell lines using an integrative search engine approach. *J. Proteome Res.*, **16**, 4374–4390.

Halvey,P.J. *et al.* (2014) Proteogenomic analysis reveals unanticipated adaptations of colorectal tumor cells to deficiencies in DNA mismatch repair. *Cancer Res.*, **74**, 387–397.

Iyer,M.K. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.

Kim,M.S. *et al.* (2014) A draft map of the human proteome. *Nature*, **509**, 575–581.

Lam,M.P.Y. *et al.* (2016) Data-driven approach to determine popular proteins for targeted proteomics translation of six organ systems. *J. Proteome Res.*, **15**, 4126–4134.

Legrain,P. *et al.* (2011) The human proteome project: current state and future direction. *Mol. Cell. Proteomics*, **10**, M111.009993.

Li,Q. *et al.* (2018) Discovering putative peptides encoded from noncoding RNAs in ribosome profiling data of *Arabidopsis thaliana*. *ACS Synth. Biol.*, **7**, 655–663.

Li,Y.F. *et al.* (2010) The importance of peptide detectability for protein identification, quantification, and experiment design in MS/MS proteomics. *J. Proteome Res.*, **9**, 6288–6297.

Nagaraj,N. *et al.* (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol. Syst. Biol.*, **7**, 548.

Nesvizhskii,A.I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nat. Methods*, **11**, 1114–1125.

Nilsson,T. *et al.* (2010) Mass spectrometry in high-throughput proteomics: ready for the big time. *Nat. Methods*, **7**, 681–685.

Olexiouk,V. *et al.* (2018) An update on sorfs.org: a repository of small ORFS identified by ribosome profiling. *Nucleic Acids Res.*, **46**, D497–D502.

Paik,Y.-K. *et al.* (2012a) The chromosome-centric human proteome project for cataloging proteins encoded in the genome. *Nat. Biotechnol.*, **30**, 221–223.

Paik,Y.-K. *et al.* (2012b) Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res.*, **11**, 2005–2013.

Perez-Gracia,J.L. *et al.* (2017) Strategies to design clinical studies to identify predictive biomarkers in cancer research. *Cancer Treat. Rev.*, **53**, 79–97.

Reiter,L. *et al.* (2009) Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteom.*, **8**, 2405–2417.

Ruiz-Orera,J. *et al.* (2014) Long non-coding RNAs as a source of new peptides. *eLife*, **3**, e03523.

Samandi,S. *et al.* (2017) Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife*, **6**, e27860.

Tabas-Madrid,D. *et al.* (2015) Proteogenomics dashboard for the human proteome project. *J. Proteome Res.*, **14**, 3738–3749.

Tamborero,D. *et al.* (2013) Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.*, **3**, 2650.

Tang,H. *et al.* (2006) A computational approach toward label-free protein quantification using predicted peptide detectability. *Bioinformatics (Oxford, England)*, **22**, e481–e488.

The,M. *et al.* (2016) Fast and accurate protein false discovery rates on large-scale proteomics data sets with percolator 3.0. *J. Am. Soc. Mass Spectrom.*, **27**, 1719–1727.

Wilhelm,M. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582–587.

Zhang,B. *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382–387.

Zhang,Q. *et al.* (2017) The microprotein minion controls cell fusion and muscle formation. *Nat. Commun.*, **8**, 15664.

Zhang,Y. *et al.* (2013) Protein analysis by shotgun/bottom-up proteomics. *Chem. Rev.*, **113**, 2343–2394.

Zhu,Y. *et al.* (2018) Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat. Commun.*, **9**, 903.