non-coding RNA

MDPI

Article

FuncPEP: A Database of Functional Peptides Encoded by Non-Coding RNAs

Mihnea P. Dragomir [1,2,*,†], Ganiraju C. Manyam [3,†], Leonie Florence Ott [1,4,†], Léa Berland [1,†], Erik Knutsen [1,5], Cristina Ivan [1,6], Leonard Lipovich [7], Bradley M. Broom [3] and George A. Calin [1,6,*]

1 Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; leonieflorence.ott@gmail.com (L.F.O.); leaberland370@gmail.com (L.B.); erik.knutsen@uit.no (E.K.); civan@mdanderson.org (C.I.)
2 Department of Surgery, Fundeni Clinical Hospital, Carol Davila University of Medicine and Pharmacy, 022328 Bucharest, Romania
3 Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA; GCManyam@mdanderson.org (G.C.M.); BMBroom@mdanderson.org (B.M.B.)
4 Institute of Tumor Biology, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany
5 Department of Medical Biology, Faculty of Health Sciences, UiT—The Arctic University of Norway, N-9037 Tromsø, Norway
6 Center for RNA Interference and Non-Coding RNAs, The University of Texas MD Anderson Cancer Centre, Houston, TX 77054, USA
7 Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201, USA; llipovich@med.wayne.edu
* Correspondence: mihnea.p.dragomir@gmail.com (M.P.D.); gcalin@mdanderson.org (G.A.C.)
† These authors contributed equally to this work.

Received: 30 August 2020; Accepted: 18 September 2020; Published: 23 September 2020

Abstract: Non-coding RNAs (ncRNAs) are essential players in many cellular processes, from normal development to oncogenic transformation. Initially, ncRNAs were defined as transcripts that lacked an open reading frame (ORF). However, multiple lines of evidence suggest that certain ncRNAs encode small peptides of less than 100 amino acids. The sequences encoding these peptides are known as small open reading frames (smORFs), many initiating with the traditional AUG start codon but terminating with atypical stop codons, suggesting a different biogenesis. The ncRNA-encoded peptides (ncPEPs) are gradually becoming appreciated as a new class of functional molecules that contribute to diverse cellular processes, and are deregulated in different diseases contributing to pathogenesis. As multiple publications have identified unique ncPEPs, we appreciated the need for assembling a new web resource that could gather information about these functional ncPEPs. We developed FuncPEP, a new database of functional ncRNA encoded peptides, containing all experimentally validated and functionally characterized ncPEPs. Currently, FuncPEP includes a comprehensive annotation of 112 functional ncPEPs and specific details regarding the ncRNA transcripts that encode these peptides. We believe that FuncPEP will serve as a platform for further deciphering the biologic significance and medical use of ncPEPs.

Keywords: non-coding RNAs; long non-coding RNAs; ncRNA-encoded peptides; small open reading frames; ncRNA translation; small peptides; micropeptides

1. Introduction

Large-scale transcriptomics efforts, such as the FANTOM Consortium [1], have revealed that most of the mammalian genome is pervasively transcribed. Most of these transcripts are classified

as non-coding RNAs (ncRNAs) [2], and over two-thirds of human genes can be considered to be ncRNAs that do not encode known proteins [3,4]. NcRNAs are defined as transcripts that lack an open reading frame (ORF), in particular those lacking ORFs of >100 amino acids (aa) in size and devoid of ORF evolutionary conservation [5]. More recently, some ncRNA transcripts have been documented by experimental methods, such as mass spectrometry (MS), to encode previously unknown short peptides [6,7].

The common definition of an ncRNA coding for peptides is that of a transcript, initially annotated as non-coding, that subsequently was identified as containing a small open reading frame (smORF) encoding a peptide of less than 100 aa [8]. This definition is consistent with the fact that >95% of conventionally proteins from public databases are longer than this threshold. NcRNA-encoded peptides (ncPEPs) can be predicted by multiple computational methods [5,9] but ultimately require biological validation at the level of their biogenesis—which can be confirmed indirectly by ribosome profiling [10] or directly by MS [6]. Biologically, it is still uncertain if the function, biogenesis, or structural properties of ncPEPs are different from "classic" coding region-derived peptides. In comparison to protein-coding genes, and to short conserved ORFs of known functional peptides encoded by a subset of protein-coding genes, smORFs in ncRNAs lack evolutionary conservation and, without laboratory-based validation, also lack bioinformatics evidence of function [11]. Furthermore, many ncRNAs, and hence their respective ncPEPs, are expressed in a tissue- and species-specific manner [3]; this suggests distinct, perhaps essential, organism- or evolution-specific functions. Indeed, many of the ncPEPs have recently emerged as functional and possess newly characterized fundamental roles in cellular processes and the maintenance of cellular homeostasis [12,13]. The biogenesis of ncPEPs seems to differ from the other peptides as they are often translated from unique small coding sequences, smORFs, while peptides from coding regions are translated from well-known mRNAs that are phylogenetically conserved or are formed through the cleavage of larger peptides/proteins [14]. In particular, human smORFs may lack conservation beyond primates, whereas conventional human proteins typically exhibit at least pan-vertebrate, and sometimes pan-metazoan, conservation [15,16]. The sequence of an ncRNA smORF usually begins with a traditional AUG start codon and frequently does not terminate with typical stop codons. This start AUG codon (ATG codon in cDNA), by using ORF-finding software (such as the NCBI ORF Finder [17] or other similar tools [18]), permits the identification of hypothetical smORFs from various genomic locations, including regions annotated as non-coding. Thousands of potential small peptides have already been predicted in different organisms, but most of these arise from en masse computational identification attempts that automatically find all possible ORFs. Many of these potential peptides are translated from transcripts that had been annotated as ncRNAs, and only for a very few has their potential expression and function been studied. With the development of new identification techniques, we expect that the number of functionally validated ncPEPs will rapidly increase.

Therefore, we posit that a curated database of functionally confirmed peptides arising from ncRNA transcripts is highly necessary. Currently, our database, named **fu**nctional **nc**RNA encoded **pep**tides (FuncPEP), contains 112 peptides encoded by ncRNAs, all of which have been validated indirectly by ribosome profiling of the corresponding "host" ncRNAs, loss-of-function techniques, and/or directly by MS, Western blotting or immunostaining, and are biologically functional, being linked to a physiological or a pathological phenomenon. We decided to include indirectly confirmed ncPEPs, which we hope will be confirmed directly in the future. We are confident that discoveries in upcoming years will allow us to widely expand this database.

NcRNAs are commonly classified as long ncRNAs (lncRNAs) and short non-coding RNAs (sncRNAs). LncRNAs, in contrast to sncRNAs, are defined by a length of >200 nt. Both classes of ncRNAs consist of several different ncRNA species. While long-intergenic ncRNAs (lincRNAs) or transcribed ultraconserved regions (T-UCRs) are part of the lncRNA class, microRNAs (miRNAs), transfer RNAs (tRNAs), and PIWI-interacting RNAs (piRNAs) are examples of sncRNAs. NcRNA species of various lengths, such as circular RNAs (circRNAs) or small nucleolar RNAs

(snoRNAs), can be assigned to both classes of ncRNAs in dependence of their specific size [19,20]. It is widely appreciated that most ncRNA classes, such as circRNAs, miRNAs, and lncRNAs, have distinct, important, and, in many cases, essential functions [21–23]. In cases where these ncRNAs contain experimentally validated smORFs, the ncRNAs functions may be different and independent from, or in exceptional cases even opposite to, the functions of the encoded peptide [24]. Alternatively, the smORF translation may be a non-functional byproduct of the ncRNAs's transcript, and all essential functions might still be carried out solely at the RNA level.

During our systematic literature review, we noticed several intriguing cases where a transcript initially annotated as an ncRNA that encodes a peptide was reclassified as a protein-coding transcript. This exposes several fundamental and paradoxical conundrums of post-genomic biology: Are "ncRNAs" really biologically and empirically, non-coding? How many ncRNAs are translated into peptides? Are some ncRNAs incorrectly annotated in transcriptome databases, being actually protein-coding transcripts as only the coded peptide is functional, or do such ncRNAs harbor both non-coding transcript and small peptide functions? Are some functional ncRNAs only occasionally, unexpectedly and erroneously translated by cellular ribosomes?

More than a decade has passed since these conundrums were raised for the first time, and no definitive answers exist [5]. Clearly, a subset of "lncRNA translation" events is an annotation artifact, and is due to the erroneous mis-annotation of certain mRNAs as "lncRNAs" [7]. Two distinguishing properties characterized this subset: Robust translation at levels comparable to those of other highly expressed known protein-coding genes, and obvious ORF homologies to known proteins in multiple species. Furthermore, considerable annotation ambiguities exist. Certain genes, such as steroid receptor RNA activator (SRA), are inherently bi-functional, with well-validated functions as both ncRNA transcript and as a peptide-codifying gene [25]. This shows that the binary division of transcripts into coding and non-coding could be in some instances a "false dichotomy". Recently, lncRNA proteogenomic was developed, by using direct MS methods, rather than indirect computational or lab-based methods. All human lncRNA genes from the ENCODE Consortium's Gencode human gene catalog were tested for evidence of translation in one normal cell type and one cancer cell line. It was determined that most lncRNAs are very rarely translated, as only approximately 1% of lncRNAs have smORFs whose translation is detectable by MS. However, many of these translational events were singletons (and therefore possible false positives or rare cellular mistakes), and lncRNAs were vastly translationally depleted, compared to known mRNAs, even after normalization for expression-level differences. Intriguingly, a few peptides were "non-singleton hits" (detected as multiple independent events in the mass spectra, meaning that they were real translation products and not one-off artifacts), suggesting that MS is a valuable approach for direct discovery of translated lncRNA ORFs [7]—a direct approach that remains persistently underutilized in an era of continuing overreliance on indirect ribosome-profiling-based detection of putative lncRNA translation.

We constructed a new database in order to centralize and monitor this valuable information, to systematically represent the ncRNAs and the corresponding ncPEPs' main characteristics, and to facilitate the use of this information for precision medicine and post-genomic era therapeutics. The database can be accessed at https://bioinformatics.mdanderson.org/Supplements/FuncPEP/.

2. Data Collection and Computational Methods

2.1. Data Collection and Database Construction

The information stored in our database was obtained through a systematic literature search and review from the NCBI PubMed and Google Scholar databases for the period 28 May 1996, when the first paper describing an ncPEP was published [26] to present (July, 2020). The literature review was conducted by two independent researchers (L.F.O. and L.B.) and disagreements were resolved by discussion with two other investigators (M.P.D. and C.I.). We selected peptides accurately identified to be encoded by ncRNAs by combining the following search terms: At least one of the following terms

regarding non-coding transcripts ("antisense RNA," OR "lincRNA", OR "lncRNA", OR "miRNA", OR "circRNA", OR "rRNA", OR "tRNA", OR "ncPEP") AND at least one of the following terms regarding translation ("smORF," OR "Ribo-seq", OR "ribosome profiling", OR "mass spectrometry", OR "translation"). Furthermore, additional manual search of the published ncRNA literature was regularly performed by a fifth author (G.A.C.).

Three additional filters were employed, and we currently require full compliance with all three conditions:

(i) We included only ncPEPs that were directly (MS, Western blotting, immunohistochemistry, immunofluorescence) or indirectly (ribosome profiling, loss/gain of function studies) experimentally confirmed;
(ii) All included ncPEPs were not merely experimentally confirmed but also functionally characterized (linked to a physiological or a pathological process); and
(iii) For concordance with the definitions discussed above, we considered ncPEPs of only ≤ 100 aa in length.

NcPEPs confirmed by indirect methods are marked by an asterisk "*" in the database and will be screened for further confirmation by direct methods. They currently comprise a major part of the database because ribosome profiling studies (also known as "Ribo-seq", i.e., selective sequencing of only the ribosome-bound RNA fraction instead of all cellular RNAs) remain more prevalent than MS. Of note, solely computationally predicted ncPEPs are not included in the database but will be added if experimentally confirmed.

The FuncPEP database includes the peptide's name assigned by the authors, its symbol, the name assigned by NCBI, as well as the synonyms that we were able to identify. In accordance with data provided by authors, we also extracted the peptides' lengths, molecular weight, sequence, functions, and data about the related ncRNA. In addition, we included data and links from two NCBI portals, Protein and Gene, and links to the Pfam 33.1 database providing predicted protein domains of the ncPEPs, giving additional potential insights into the function of ncPEPs.

As vast sequence data is available in the Pfam protein database, deep learning was utilized to build the prediction model using Python 3.4. The Pfam seed sequence data with over 1.25 million records was divided into training (80%) and testing sets (20%). Keras 2.3.0 with TensorFlow 2.0. was used on this sequential data with a variant of convolution neural network called the residual convolution network to identify the functional classes of proteins [27]. Only families with at least 10 peptide sequences in the database were used in building the model to predict domains of peptides from ncRNA. The final convolution network was generated with a batch size of 256 running for 20 epochs. This residual convolution network model performed well on test data with an accuracy of over 95%. The Pfam domains for non-coding peptides of lengths greater than 20 were predicted using this neural network model. The corresponding domain information is included in the FuncPEP database.

We assigned individual identifiers to the retrieved peptides, as symbols and names utilized across the research community might become ambiguous over time. The website for FuncPEP database was developed using R open source programming language and formatted by Markdown.

2.2. *Sequence and Molecular Weight Prediction*

We collected all the given information from each paper. However, sometimes information, such as the aa sequence or molecular weight, was not given. In order to provide a complete database, we used the NCBI ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/) to predict amino acid sequences when only the DNA sequence of the respective ORF was provided. Additionally, we used the Peptide Molecular Weight Calculator by Selleckchem (https://www.selleckchem.com/peptide-calculator.html) to calculate the molecular weights of peptides without provided weight or when only the approximate weights derived from Western blots were given.

2.3. Comparison of ncPEPs with Known Peptides Encoded by Messenger RNAs of Coding Genes

We compared the composition of human ncPEPs with that of human peptides from coding regions. For this purpose, we downloaded from https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildInfo the sequences of 2,476,065 distinct human tryptic-digest peptides detected by mass spectrometry analysis of several human organs and datasets (all ≤ 100 aa), build Human 2020–01. Several criteria for classifying human peptides from a biochemical perspective were retrieved from [26,27]. The composition of amino acids, respectively, classes of amino acids in peptides sequences, was obtained with Perl and R (version 3.5.1) codes.

The significance of the difference between the proportions of different amino acids in the human peptides encoded by messenger RNAs compared to human ncPEPs was assessed in R (version 3.5.1) with the nonparametric Mann–Whitney–Wilcoxon test with a default correction method Benjamini & Hochberg for multiple testing and with a Chi-Square test. A box-and-whisker plot (Box plot represents first (lower bound) and third (upper bound) quartiles, whiskers represent 1.5 times the interquartile range) was used to visualize the data.

3. Implementation and Results

3.1. Systematic Review

A total of 25,330 articles were identified on PubMed and Google Scholar databases. After removing the duplicates, 18,272 articles were screened by title and abstract, and 302 articles were considered for full-text assessment. Of these, 44 studies, containing 112 validated functional peptides, were included in FuncPEP's database according to our inclusion criteria (Figure 1).

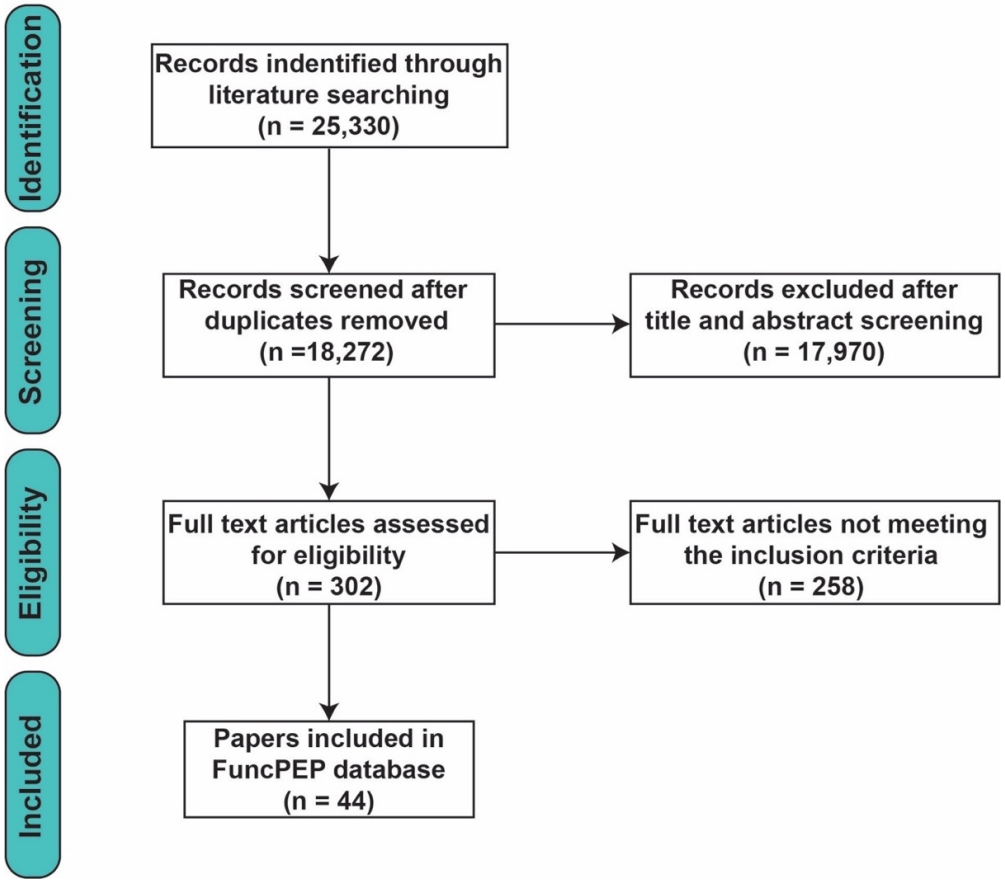

Figure 1. Flowchart describing the process of scientific paper identification, screening, eligibility testing, and inclusion.

3.2. Database Interface

FuncPEP provides a user-friendly interface through a website allowing access to an ncPEP database. The database website can be accessed at: https://bioinformatics.mdanderson.org/Supplements/FuncPEP/.

The database website is divided into four sections: (1) The Home section, where users can find background information on the database as well as on the topic of ncPEP; (2) the Database section (red circle and arrow), where all information on ncPEPs is accessible through a dynamic table browser and providing an overview of the ncPEPs (Figure 2A) and complete information for each ncPEP can be accessed by selecting the respective ncPEPs ID (green dashed circle and arrow) (Figure 2B); (3) the Methods section (blue circle and arrow), describing the methodology used to curate and collect the ncPEPs data (Figure 2C); and (4) the Help section (purple circle and arrow), containing information on how to navigate the site (Figure 2D).

Figure 2. Database interface (**A**) The database section of FuncPEP providing an overview of information provided for each ncPEP. (**B**) Complete information for a selected ncPEP, accessible through a dynamic table browser. (**C**) The methods section of FuncPEP, describing the methodology used to curate and collect the ncPEPs data. (**D**) The help section of FuncPEP containing information on how to navigate the site.

3.3. Characteristics of ncPEPs from the FuncPEP Database

FuncPEP was created as a useful resource for accumulating ncPEP data from published scientific literature. One of FuncPEP's aims is to provide comprehensive information about functionally confirmed ncPEPs. To establish a clear overview of the topic, FuncPEP also contains information about corresponding ncRNAs: their length, the genomic position, symbols from the source and NCBI, and the predicted Pfam domain.

In accordance, ncPEPs included in FuncPEP are all less than 100 aa. The average size of the 112 ncPEPs included in FuncPEP is about 49 aa (Figure 3A). FuncPEP also contains the molecular weight of the retrieved ncPEPs; for human ncPEPs, this ranges between 0.65 and 15 kDa, with an average of 5.7 kDa (Figure 3B).

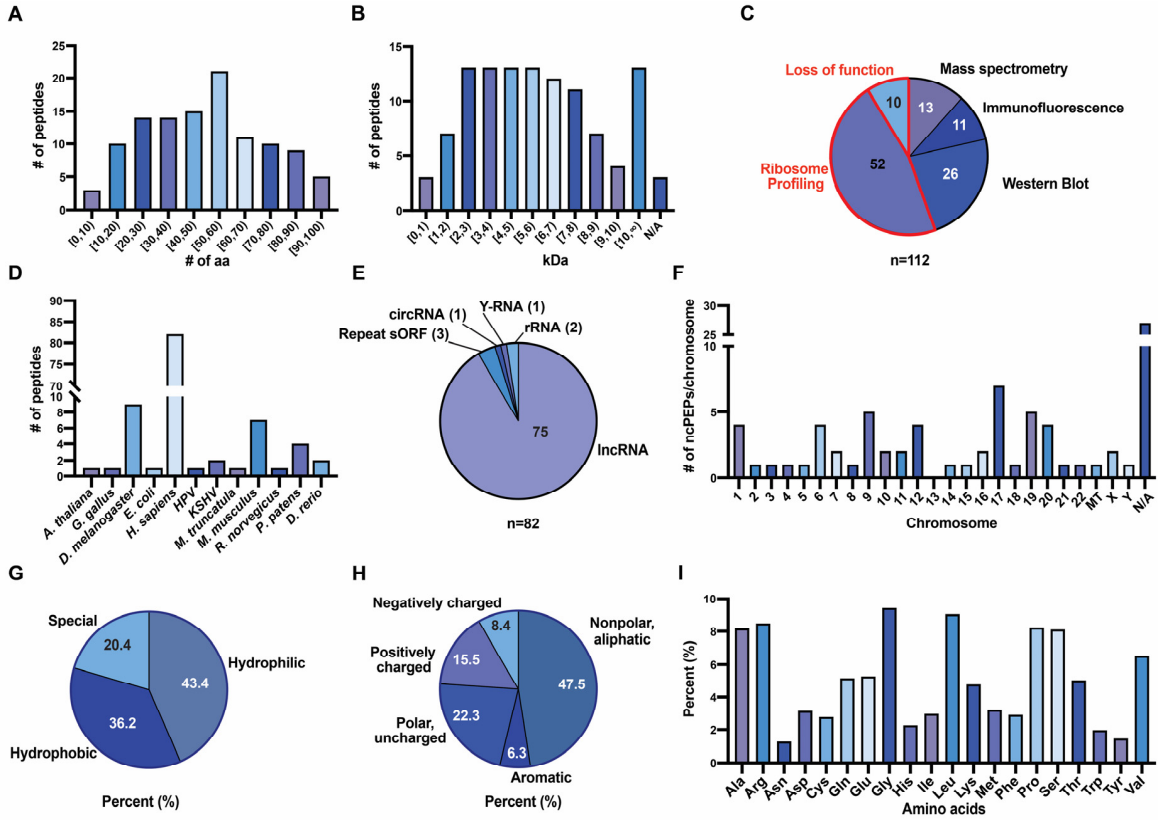

Figure 3. Characteristics of ncPEPs from the FuncPEP database. (**A**) NcPEP´s length in amino acids; (**B**) NcPEP´s molecular weight (kDa); (**C**) Methods used to detect ncPEPs (red border and font—indirect identification methods, black border and fond—direct identification methods); (**D**) Species in which ncPEPs were discovered; (**E**) Classification of human ncPEP´s host ncRNA; (**F**) Chromosomal distribution of human ncPEPs; (**G**) Classification of human ncPEP´s amino acids into hydrophobic, hydrophilic, and special; (**H**) Classification of human ncPEP´s amino acids according to the chemistry of the R groups; (**I**) Amino acids composition of human ncPEPs. #: ncPEPs.

Most of the peptides in our database were not detected by a single but by several different methods. Often, bioinformatically predicted smORFs located in ncRNA were experimentally proven to be translated. Since bioinformatics predictions were regarded as insufficient proof for the existence of an ncPEP, only experimentally confirmed peptides were included. Experimental methods range from indirect ones (red border), such as loss-of-function assays and ribosome profiling, to direct ones (black border), like immunofluorescence staining, Western blotting, or MS, the latter providing the highest sensitivity. Only taking the most reliable detection method into consideration, 13 (11.6%) of the

peptides in our database were detected by MS, 26 (23.2%) by Western blot, 11 (9.8%) by immunostaining methods, 52 (46.4%) by ribosome profiling, and 10 (8.9%) by loss-of-function studies (Figure 3C).

Next, we analyzed in which species the ncPEPs were discovered. Importantly, most ncPEPs have been discovered and studied in *Homo sapiens* (73.2%, 82/112), followed by *Drosophila melanogaster* (8%, 9/112) and *Mus musculus* (6.3%, 7/112) (Figure 3D). Some ncPEPs can be found across different species, such as Myoregulin identified in *Homo sapiens* and *Mus musculus*, or Humanin, detected in *Homo sapiens* and *Rattus norvegicus*. This implies a certain degree of conservation amongst some ncPEPs, and strengthens the case for their potential functionality, even though for lncRNAs at large and presumably for their smORFs—lack of conservation does not imply lack of function [28]. Yet, because the majority of ncPEPs was discovered in humans, we decided to further focus on characterizing human ncPEPs.

We analyzed the biotypes of ncRNAs that harbor human ncPEP smORFs. The majority of human peptides are encoded by lncRNAs (91.5%, 75/82), three by smORFs in a genomic GGGGC repeats (3.7%), two by mitochondrial rRNAs (2.4%), and one each by a Y-RNA (1.2%) and a circRNA (1.2%) (Figure 3E). This data underlines the fact that sncRNA only exceptionally are encoding ncPEPs and usually the longer immature forms of sncRNAs are the ones long enough to contain ncPEP smORFs. The best-known examples of sncRNA encoded peptides were discovered in plants and are two pri-miRNAs: pri-miR-171b (*Medicago truncatula*) and the pri-miR-165a (*Arabidopsis thaliana*), which are known to produce specific peptides miPEP171b and miPEP165a [29]. While in theory all ncRNA are a potential source of ncPEPs and our database contains ncPEPs from multiple ncRNA species, the majority of ncPEPs are derived from lncRNAs. Of note, having a length of up to several hundreds of nucleotides, most circRNAs and pri-miRNAs are also lncRNAs [30,31].

Subsequent analysis of the chromosomes that harbor the respective human ncRNAs revealed that, except for chromosome 13, ncPEPs can be found on all chromosomes, including the X and Y chromosomes (Figure 3F). The potential lack of ncPEPs on chromosome 13 (although we cannot exclude the identification of ncPEP from chromosome 13 ncRNAs), is likely due to the gene-poor nature of this chromosome (this is also the reason why trisomy 13, 18, or 21 are the only live birth survivable trisomies in humans): The fewer genes there are in general, the fewer ncRNAs and hence ncPEPs. Indeed, we also observed only one ncPEP on chromosome 18 and one on chromosome 21.

Furthermore, ncPEPs' structural characteristics presented in the FuncPEP database include all data considered relevant and useful for peptide studies. Hence, we evaluated ncPEPs in regard to their aa chemistry and composition. All the following analyses are taking into consideration only human ncPEPs with a confirmed aa sequence (n = 31). Analyzing the composition [32] of curated ncPEPs, we found that they are similarly composed of hydrophilic (43.4%) and hydrophobic amino acids (36.2%) and the remaining 20.4% are special ones (glycine, cysteine, proline) (Figure 3G). Moreover, evaluating the aa´s charge [33] of ncPEPs, 47.5% of the aa carry non-polar aliphatic R-groups, 6.3% aromatic, 22.3% polar uncharged, 15.5% positively charged, and 8.4% negatively charged R-groups (Figure 3H). Looking more in detail at ncPEP´s aa composition, we observed that glycine, leucine, arginine, proline, and alanine are the five most frequent aa from the structure of human ncPEPs accounting for 43% of all aa, whereas tryptophan, tyrosine, and asparagine are the rarest ones with less than 5% in total (Figure 3I).

3.4. Comparison of Human Functional ncPEPs with Human Peptides from Coding Regions

Subsequently, we compared the physico-chemical properties of human ncPEPs with an aa sequence provided (n = 31) with human tryptic-digest peptides detected by MS and mapping to known protein-coding regions (n = 2,476,065) derived from the PeptideAtlas (https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/buildInfo, [34]). Despite the significantly different sizes of the datasets, we first compared the length of the human ncPEPs with the sequence provided and the human coding peptides. NcPEPs range from 16 to 90 aa in length, with a median length of 52 aa. Human coding peptides range from 7 to 83 aa in length with a median length of 15 aa (Figure 4A). Furthermore, we analyzed the

chemical properties of the aa and noticed a similar content of hydrophobic aa across both peptide groups (p = n.s.). However, they differed significantly in regard to their hydrophilic and special aa content. The ncPEPs contain less hydrophilic ($p = 0.0064$) but more special aa ($p = 0.0037$) (Figure 4B). Moreover, we also noticed several differences between the two groups of peptides regarding the chemistry of their R-groups. NcPEPs have a significantly more nonpolar aliphatic group ($p = 0.0072$) and positively charged groups ($p = 0.0115$) compared to peptides from coding regions. On the other hand, ncPEPs contain significantly less negatively charged R-groups compared to human peptides from coding regions ($p = 3.36 \times 10^{-5}$). Only the content of aromatic R-groups (p = n.s.) and of polar uncharged R-groups (p = n.s.) is similar between the two groups (Figure 4C).

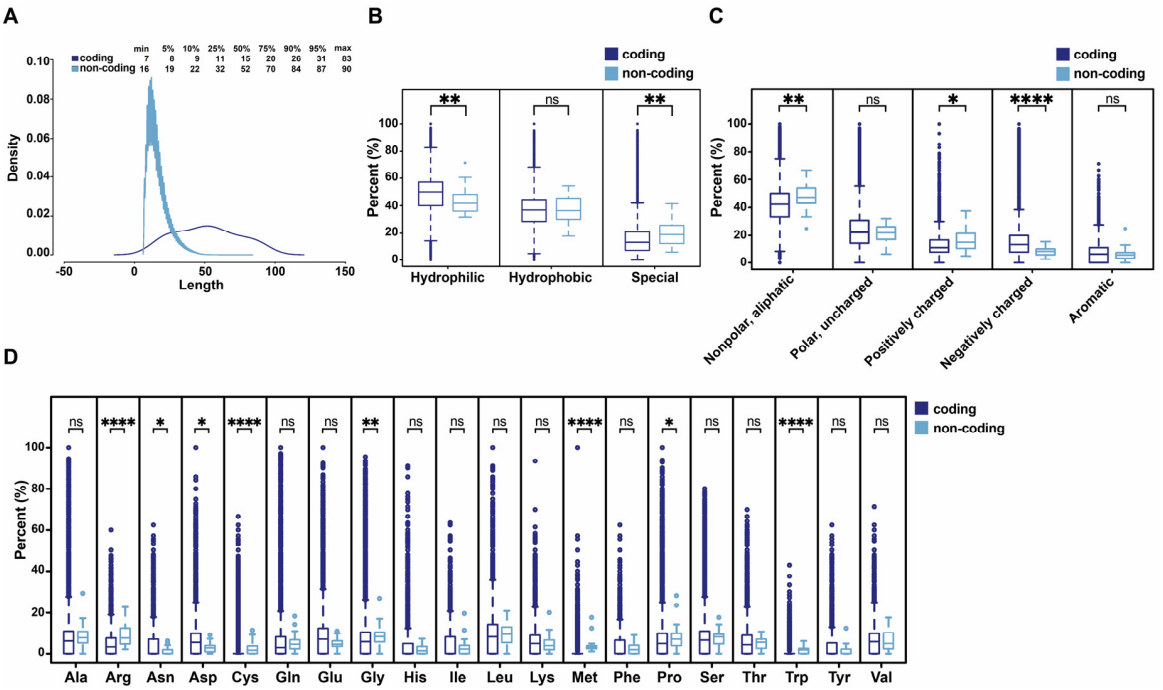

Figure 4. Up-to-date comparison of human functional ncPEPs with human peptides from coding regions. (**A**) Comparison of the size range between human ncPEPs and peptides from coding regions. (**B**) Comparison between human ncPEPs and peptides from coding regions according to the properties of the amino acids (hydrophobic, hydrophilic, and special). (**C**) Comparison between human ncPEPs and peptides from coding regions according to the chemistry of the R groups (aromatic R groups; nonpolar aliphatic R groups; negatively charged R groups; positively charged R groups; and polar uncharged groups). (**D**) Comparison between human ncPEPs and peptides from coding regions according to the amino acid distribution. Due to a not normally distributed dataset, data is presented using box plots with whiskers representing 1.5 times the interquartile range. (Mann–Whitney–Wilcoxon test with Benjamini & Hochberg correction for multiple testing; ns—not significant, * $p < 0.05$; ** $p < 0.01$; **** $p < 0.0001$).

Additionally, we compared the distribution of aa on the peptides within the two sample sets. In human peptides encoded by coding RNAs, the three most abundant aa are leucine, glutamic acid, and serine while methionine, cysteine, and tryptophan are all rather rare aa [32]. Human ncPEP´s aa composition displays some parallels to, but is mainly distinct from, the composition of the peptides from coding regions. NcPEPs contain more arginine ($p = 3.59 \times 10^{-7}$, cysteine ($p = 7.82 \times 10^{-6}$), glycine ($p = 0.0031$), methionine ($p = 8.81 \times 10^{-14}$), proline ($p = 0.0272$), and tryptophan ($p = 1.38 \times 10^{-13}$) compared to human peptides from coding regions. In contrast to that, the prevalence of asparagine ($p = 0.0296$) and aspartic acid ($p = 0.0196$) is significantly lower in ncPEPs. The distribution of the other aa onto the peptides across the two sample sets does not differ significantly (Figure 4D).

However, despite all the similar patterns regarding the physical-chemical properties of ncPEPs and peptides from coding regions, it is noteworthy that the Chi-square test revealed significant differences regarding the general distribution of aa, their physical-chemical properties, and that of their R-groups ($p < 0.0001$, for all three comparisons). Nevertheless, the comparison of ncPEPs with peptides from known proteins is confounded by the fact that groups of very different sizes were compared. To our knowledge, however, this is the first comparison between human peptides from coding regions and ncPEPs, and our preliminary data suggest several notable differences between these two types of peptides.

3.5. Examples of ncPEPs Functions

NcPEPs included in FuncPEP have been shown to play specific roles in many cellular processes, in certain cases even contrary to those played by the ncRNAs that they are translated from. Indeed, the lncRNA, HOXB-AS3, one of the many antisense lncRNAs overlapping the human homeobox B (HOXB) cluster of transcription factor genes first characterized by the FANTOM Consortium [1], plays an oncogenic role in myeloid malignancies and its high expression associates with poor prognosis [35]. On the other hand, its corresponding ncPEP has a tumor suppressive function in colorectal cancer: The loss of this ncPEP is a central tumorigenic event, leading to the reprogramming of cancer metabolism by enabling hnRNP A1-dependent aerobic glycolysis [36].

Additionally, we observed that ncPEPs may act as key regulators in many fundamental processes, including calcium homeostasis [12], gastrulation [37], interaction with the mRNA decapping complex [38], development [39], cell proliferation, muscle growth [40], and immunity [41]. For example, Anderson et al. discovered myoregulin (MLN), responsible for SERCA inhibition. SERCA is a membrane pump controlling the uptake of Ca^{2+} into the sarcoplasmic reticulum (SR). MLN impedes Ca^{2+} uptake by direct interaction with SERCA, finally controlling muscle relaxation. In mice, the genetic deletion of this ncPEP enhances Ca^{2+} retention in skeletal muscles improving exercise performance [12]. Cai et al. demonstrated the role of a not yet named ncPEP, encoded by the lncRNA-Six1. This ncPEP activates the Six1 gene, while the corresponding lncRNA carries out a cis-acting regulation of the protein Six1. Taken together, these data suggest that this ncPEP promotes cell proliferation and is involved in muscle growth [40].

An ncPEP encoded by lncRNA 1500011K16Rik (*Mus musculus*) or LINC00116 (*Homo sapiens*) was published in parallel by two different groups. Mitoregulin or micropeptide regulator of β-oxidation (MOXI) is localized at the inner mitochondrial membrane. By association with the trifunctional protein, it positively influences mitochondrial fatty acid β-oxidation. Mitroregulin knock-out mice suffer from a perturbed fatty acid metabolism and reduced mitochondrial respiratory efficiency, resulting in decreased agility of the animals [42,43]. Interestingly, another human peptide was described by two different groups. The non-annotated *p*-body dissociating polypeptide (NBDY) was first described to be involved in the 5′ to 3′ mRNA decay pathway by interacting with mRNA decapping enzymes [38]. Later the same year, the peptide was also detected to be in involved in the immune response upon viral infection [41]. This bifunctionality stresses that ncPEPs are important regulators and not merely translational artefacts.

Given ncPEPs' roles in fundamental cellular mechanisms, their function in oncogenic development and progression appears probable. Indeed, several studies on ncPEPs have shown that these peptides actually play key roles in a number of human cancers. Many ncPEPs have been found to play protumorigenic or antitumorigenic functions by affecting cancer metabolic reprogramming, the stability of oncogenic proteins, or the epithelial-mesenchymal transition (EMT) process. Zheng et al. showed that the circular RNA circPPP1R12A contains an ORF encoding a functional ncPEP, named circPPP1R12A-73aa. They found that this ncPEP promotes colorectal cancer growth and metastasis via the activation of the Hippo-YAP signaling pathway [44]. Another ncPEP, CASIMO1, promotes breast cancer progression by various mechanisms. It is involved in actin reorganization and thereby facilitates cell mobility. Additionally, it positively influences cell cycle progression and interacts

with squalene epoxidase (SQLE), resulting in a modulated steroid synthesis [45]. Other ncPEPs promote tumor progression by stimulating cell proliferation—circE7 [46], influencing immune tolerance in melanoma cells—meloe-3 [47], or inhibiting apoptosis in esophageal squamous carcinoma by interaction with Yin Yang 1—YY1BM [48].

In contrast to that, a number of ncPEPs in our database have tumor-suppressive capacities. CIP2A-BP, for example, encoded by LINC00665, binds directly to the oncogene CIP2A. Upon this interaction, the PIK3A/AKT/NFκB pathway is inhibited, resulting in decreased expression of matrix metallopeptidase (MMP2), MMP9, and Snail that all can promote tumor progression in triple-negative breast cancer (TNBC) [49]. Additionally, in TNBC, a small regulatory peptide of STAT3 (ASRPS), binds to STAT3 and thereby reduces expression of vascular endothelial growth factor (VEGF), resulting in decreased angiogenesis [50]. Furthermore, the micropeptide-inhibiting actin cytoskeleton (MIAC) exerts its tumor-suppressive capacity by inhibiting cell proliferation and metastasis through actin formation suppression in head and neck squamous cell carcinoma [51]. The HOXB-AS3 peptide prevents metabolic reprogramming in colorectal cancer [36], and the ncPEP LINC87aa, derived from a circRNA, inhibits the transcriptional elongation of multiple oncogenes in glioblastoma [52].

An overview of the diverse functions exerted by the ncPEPs in our database reveals that across as all species the majority of ncPEPs is involved in immunity (n = 62, 55.36%). This observation is in accordance with the fact that small peptides are capable to bind and regulate the folding and function of histocompatibility leukocyte antigen (HLA) [53,54]. Therefore, we consider that an ingenious method to discover new functional ncPEPs is to pull-down HLA molecules and to analyze the peptides located in their specific binding grooves. Especially in *Drosophila melanogaster* and several plants, ncPEPs are also involved in development, representing the second largest functional group across all species (n = 15, 13.39%). Other ncPEPs are involved in different muscular processes (n = 7, 6.25%), cancer biology (n = 9, 8.04%), neural processes (n = 4, 3.57%), metabolism (n = 4, 3.57%), and other cellular functions, such as viral infection (n = 2, 1.79%), differentiation (n = 2, 1.79%), transcription (n = 1, 0.89%), translation (n = 1, 0.89%), cell survival (n = 1, 0.89%), cellular proliferation (n = 1, 0.89%), drug resistance (n = 1, 0.89%), and one ncPEP has antioxidative capacities (0.89%) (Figure 5A). Since the majority of the ncPEPs in our database are found in *Homo sapiens*, the distribution of functions in the human ncPEP set is similar. Apart from the majority of human ncPEPs being involved in immunogenic processes (n = 60, 73.17%), several human ncPEPs play roles in cancers, either working as tumor suppressors (n = 5, 6.1%) or exerting oncogenic functions (n = 4, 4.89%). A similar number of human ncPEPs are involved in neural (n = 3, 3.66%) and muscular processes (n = 3, 3.66%), while the others influence development (n = 1, 1.22%), transcription (n = 1, 1.22%) and translation (n = 1, 1.22%), cell survival (n = 1, 1.22%), cellular differentiation (n = 2, 2.44%), and one has antioxidative effects (Figure 5B). The proportion of immunity-associated ncPEPs is greater in human than in all species, but this difference may be due to ascertainment bias (some families of short immune peptide genes are better-studied in humans) not to a genuine functional distinction. This overview demonstrates that ncPEPs are involved in a multitude of different cellular processes in *Homo sapiens* and other species, implying that they represent an important group of molecules that has not been explored yet. Because of their functions and key roles in cancer initiation and progression, ncPEPs may become a new class of biomarkers and, in cases where they are directly functional, may be amenable to repurposing into therapeutic targets or tools.

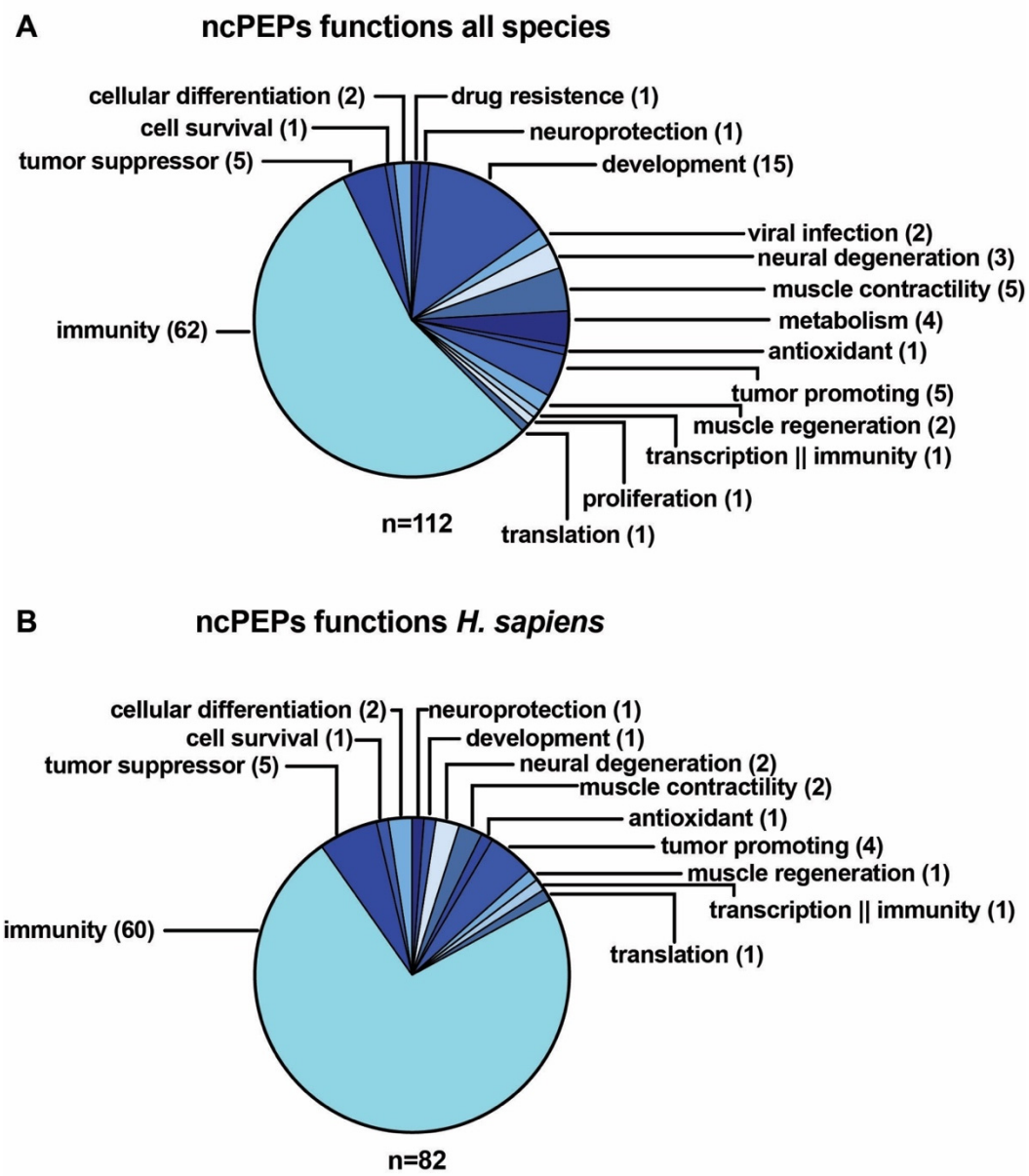

Figure 5. Overview of ncPEP functions. (**A**) Functions of ncPEPs across all species in our database. (**B**) Functions of the human ncPEPs in our database.

4. Discussion

The originality of the FuncPEP database is implicit in its three key aspects. First, all included ncPEPs are experimentally confirmed, and not merely predicted. Several studies that we identified during the screening process only employed computational predictions, or indirect experimental methods, to detect potential ncPEPs en-masse, but those candidate ncPEPs were not further validated. These ncPEPs were hence not included in FuncPEP, and will be added in the future only if they become experimentally confirmed. We consider that prediction tools should only be used as an initial step for detecting potential ncPEPs, which needs following experimental confirmation, and we caution against the widespread overreliance on ribosome profiling as an ncPEP discovery pipeline without validations by complementary independent methods.

We observed, during the screening process, multiple ways to study the protein-coding potential of ncRNAs. The phyloP score is an interesting method to detect potential ncPEPs by analyzing the conservation of the genomic sequence [55]. Given that, for known protein-coding genes, a high

conservation score is correlated with a higher translation potential [56], many researchers have chosen to use the phyloP score to predict if an ncRNA encodes a peptide, even though unbiased genome-wide phyloP surveys of ncPEPs have not yet been undertaken. Another method frequently used to detect potential ncPEPs is PhyloCSF. This technique allows a specific and easy phylogenic classification of small genomic regions [10]. Based on a formal statistical comparison of phylogenetic codon models, PhyloCSF is a comparative genomics method analyzing multispecies nucleotide sequence alignment to determine whether a genomic region is likely to represent a conserved protein-coding region. However, an important limitation of phyloP and PhyloCSF is that, in lineages as diverse as yeast and mammals, most lncRNA genes, dramatically and unlike protein-coding genes, are not conserved [5]. Even between closely related species, such as the great apes and the prosimians, lncRNA conservation is mostly lacking. In fact, two-thirds to three-quarters of human lncRNAs are not conserved beyond primates [57,58]. Therefore, it is crucial to also employ conservation-independent methods for the discovery of smORFs in ncRNAs. Computational ORF finders [17,18] can identify smORFs of 100 aa or fewer that are able to encode peptides. SmORFs are usually excluded from proteome annotation, especially if they are not conserved and do not represent well-known small proteins, even if they are bound by cellular ribosomes or actively undergo translation.

After ncPEPs are predicted, or as an experimental alternative to computational prediction, two key high-throughput whole-genome laboratory-based methods are currently in wide use to empirically confirm smORFs: ribosome profiling and MS. Ribosome profiling is one of the most popular techniques to assess the translational relevance of smORFs [59]. Based on high-throughput sequencing of the polysome-bound RNA fraction (as opposed to all RNA in the cell), followed by bioinformatics analysis of the mapped RNA fragments, ribosome profiling gives a "global snapshot" of the "translatome" at a particular time point [59]. However, this technique provides only indirect evidence of translation, given that ribosome binding of lncRNAs was observed previously and suggested to also regulate translation of mRNAs by occupying ribosomes. Additionally, one tRNA from a bacterium [60] and a small ncRNA from yeast [61] were shown to exert regulatory capacities in translation in stress conditions, indicating that ncRNAs of all sizes can bind to ribosomes without being translated themselves [62]. Hence, the results from ribosome profiling could represent unproductive encounters between ribosomes and RNAs that do not necessarily result in the translation of the RNAs. More recently, a key advance in bioinformatics analysis of ribosome profiling data has enabled the differential identification of ncRNA smORFs that show clearly mRNA-like patterns of ribosome pausing at the E, P, and A sites of each codon, strongly suggesting that translation is actually occurring [63]. Understanding the proportion of ribosome-lncRNA binding events consistent with E/P/A pausing, as opposed to indiscriminate ribosome sponging by lncRNAs, remains a challenge in the ribosome profiling field. On the other hand, MS is considered the gold-standard technique in proteomics, being the sole direct tool to study the protein-coding potential of ncRNA smORFs. Unlike ribosome profiling, MS has not been commonly used for ncRNA smORF analysis, mainly because of the high cost of the procedure, the necessary extensive customization of proteomics analysis, and the decreased likelihood of detection of low-abundance targets [7,64]. Despite these limitations, MS allows ncPEP detection through the unbiased whole-proteome identification of amino acid sequences in tryptic-digest mass spectra [65], which can then be matched to predicted and known ncRNA smORFs from a custom database.

Second, all ncPEP included in FuncPEP were proven to be functional, being linked to a physiological or pathological phenomenon. Hence, beyond experimental confirmation, the subsequent objective to be achieved when working with ncPEPs is to determine their physiological or pathological function. Targeted genome editing is used to confirm transcriptional relevance, but more importantly to study peptide function by either deleting the entire ORF or deleting or replacing specific parts of the ORF. Subsequently, appropriate phenotypic screens in cell culture or in vivo model organisms are performed. For ncRNAs, disabling (but not deleting) the entire smORF, for example by changing the ATG start codon at the genome level into a codon that does not support translation initiation, is a good way to test whether their cellular function depends on the translation of the smORF. Allowing rapid and

efficient genomic editing, CRISPR/Cas9 is one of the most widely adopted "gain and loss of function" techniques [66]. However, also other genome editing techniques are used to analyze ncPEP´s function. For example, Myoregulin, a peptide encoded by a skeletal muscle-specific RNA, previously annotated as an lncRNA, has been studied using Talen (transcription activator-like effector nuclease)-mediated homologous recombination, a powerful and alternative strategy to perform genome editing [12]. Whole-genome high-throughput gene editing functional screens (using CRISPR/Cas9, Zfn, or Talen) in human cell cultures and model organisms is becoming essential for differentiating ncPEP-dependent from RNA-only biological functions of ncRNAs.

Third, we included only ncPEPs that did not exceed 100 aa in length. Our approach is mainly motivated by previous studies that showed that in lncRNAs, most of the ORFs corresponded to peptides less than 100 aa in length [8], as well as by the fundamental definition of an lncRNA as a transcript with solely sub-100-aa ORFs [5]. Moreover, because of their small size, we believe that this class of peptides was overlooked and in the near future will be intensely researched and established as a separate entity.

In particular, putative hormone-like and innate-immunity (defensin-like) functions of ncPEPs, along with the potential of disease-specific ncPEPs to function as signaling molecules and autoantigens, and their possible interactions with pathogens' nucleic acids and proteins during infection, summarily warrant devoting an increased amount of attention toward the lncRNA proteogenomics field. FuncPEP will facilitate elucidation of this still-emerging biology.

5. Conclusions

The FuncPEP database was developed to organize and present the published data on ncPEPs. The first papers on the topic of ncPEPs were published around the year 2000, and we have noticed an exponentially growing interest in ncPEPs, in accordance with the exponentially increasing number of publications and discoveries in the field. Thanks to an easy-to-use interface, FuncPEP allows researchers to find essential, relevant, and up-to-date data on functionally characterized ncPEPs. The database includes details about both the ncPEPs and the related ncRNA transcripts that encode them. Additionally, a link to the original paper reporting the discovery of each ncPEP can be found. Of note, only functionally well-characterized ncPEPs were included in FuncPEP and computationally predicted ncPEPs were excluded and will be included only if their function will be experimentally confirmed. By creating FuncPEP, we aim to provide a useful, easy-to-use, and continually updated web-based bioinformatics tool that allows researchers to quickly and directly access recent discoveries in the area of ncPEPs.

Author Contributions: Investigation, methodology, data curation, formal analysis, visualization, writing—original draft, supervision, M.P.D.; data curation, methodology, software, writing—original draft, G.C.M.; investigation, methodology, data curation, formal analysis, visualization, writing—original draft, L.F.O.; investigation, methodology, data curation, formal analysis, visualization, writing—original draft, L.B.; conceptualization, supervision, writing—review and editing, E.K.; investigation, methodology, data curation, formal analysis, writing—review and editing, C.I.; conceptualization, supervision, writing—review and editing, L.L.; conceptualization, supervision, writing—review and editing, B.M.B.; conceptualization, supervision, writing—review and editing, funding acquisition, G.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: Calin is the Felix L. Haas Endowed Professor in Basic Science. Work in Calin's laboratory is supported by National Institutes of Health (NIH/NCATS) grant UH3TR00943-01 through the NIH Common Fund, Office of Strategic Coordination (OSC), the NCI grants 1R01 CA182905-01 and 1R01CA222007-01A1, an NIGMS 1R01GM122775-01 grant, a U54 grant #CA096297/CA096300–UPR/MDACC Partnership for Excellence in Cancer Research 2016 Pilot Project, a Team DOD (CA160445P1) grant, a Chronic Lymphocytic Leukemia Moonshot Flagship project, a Sister Institution Network Fund (SINF) 2017 grant, and the Estate of C. G. Johnson, Jr. Lipovich was supported by the NIH Director's New Innovator Award.

Conflicts of Interest: The authors declare no conflict of interest.

Availability: FuncPEP is freely accessible at https://bioinformatics.mdanderson.org/Supplements/FuncPEP/.

References

1. The FANTOM consortium; Carninci, P.; Kasukawa, T.; Katayama, S.; Gough, J.; Frith, M.; Maeda, N.; Oyama, R.; Ravasi, T.; Lenhard, B.; et al. The Transcriptional Landscape of the Mammalian Genome. *Science* **2005**, *309*, 1559–1563. [CrossRef]
2. Mudge, J.; Frankish, A.; Harrow, J. Functional Transcriptomics in the Post-ENCODE era. *Genome Res.* **2013**, *23*, 1961–1973. [CrossRef]
3. Derrien, T.; Johnson, R.; Bussotti, G.; Tanzer, A.; Djebali, S.; Tilgner, H.; Guernec, G.; Martin, D.; Merkel, A.; Knowles, D.G.; et al. The GENCODE v7 Catalog of Human Long Noncoding RNAs: Analysis of Their Gene Structure, Evolution, and Expression. *Genome Res.* **2012**, *22*, 1775–1789. [CrossRef]
4. Frankish, A.; Diekhans, M.; Ferreira, A.-M.; Johnson, R.; Jungreis, I.; Loveland, J.E.; Mudge, J.; Sisu, C.; Wright, J.C.; Armstrong, J.; et al. GENCODE Reference Annotation for the Human and Mouse Genomes. *Nucleic Acids Res.* **2018**, *47*, D766–D773. [CrossRef]
5. Dinger, M.E.; Pang, K.C.; Mercer, T.R.; Mattick, J.S. Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. *PLoS Comput. Boil.* **2008**, *4*, e1000176. [CrossRef] [PubMed]
6. Slavoff, S.A.; Mitchell, A.J.; Schwaid, A.G.; Cabili, M.N.; Ma, J.; Levin, J.Z.; Karger, A.; Budnik, B.A.; Rinn, J.L.; Saghatelian, A. Peptidomic Discovery of Short Open Reading Frame–Encoded Peptides in Human Cells. *Nat. Methods* **2012**, *9*, 59–64. [CrossRef]
7. Banfai, B.; Jia, H.; Khatun, J.; Wood, E.; Risk, B.; Gundling, W.E.; Kundaje, A.; Gunawardena, H.P.; Yu, Y.; Xie, L.; et al. Long Noncoding RNAs are Rarely Translated in Two Human Cell Lines. *Genome Res.* **2012**, *22*, 1646–1657. [CrossRef] [PubMed]
8. Ruiz-Orera, J.; Messeguer, X.; Subirana, J.A.; Albà, M.M. Long Non-coding RNAs as a Source of New Peptides. *eLife* **2014**, *3*, e03523. [CrossRef] [PubMed]
9. Makarewich, C.A.; Olson, E.N. Mining for Micropeptides. *Trends Cell Boil.* **2017**, *27*, 685–696. [CrossRef] [PubMed]
10. Lin, M.F.; Jungreis, I.; Kellis, M. PhyloCSF: A Comparative Genomics Method to Distinguish Protein Coding and Non-coding Regions. *Bioinformatics* **2011**, *27*, i275–i282. [CrossRef] [PubMed]
11. Aspden, J.; Eyre-Walker, Y.C.; Phillips, R.J.; Amin, U.; Mumtaz, M.A.S.; Brocard, M.; Couso, J.P. Extensive Translation of Small Open Reading Frames Revealed by Poly-Ribo-Seq. *eLife* **2014**, *3*, e03528. [CrossRef] [PubMed]
12. Anderson, U.M.; Anderson, K.M.; Chang, C.-L.; Makarewich, C.A.; Nelson, B.R.; McAnally, J.R.; Kasaragod, P.; Shelton, J.M.; Liou, J.; Bassel-Duby, R.; et al. A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* **2015**, *160*, 595–606. [CrossRef] [PubMed]
13. Nelson, B.R.; Makarewich, C.A.; Anderson, D.M.; Winders, B.R.; Troupes, C.D.; Wu, F.; Reese, A.L.; McAnally, J.R.; Chen, X.; Kavalali, E.T.; et al. A Peptide Encoded by a Transcript Annotated as Long Noncoding RNA Enhances SERCA Activity in Muscle. *Science* **2016**, *351*, 271–275. [CrossRef] [PubMed]
14. Li, L.-J.; Leng, R.-X.; Fan, Y.-G.; Pan, H.-F.; Ye, D.-Q. Translation of Noncoding RNAs: Focus on lncRNAs, pri-miRNAs, and circRNAs. *Exp. Cell Res.* **2017**, *361*, 1–8. [CrossRef] [PubMed]
15. Ramakrishnan, A.; Janga, S.C. Human Protein-RNA Interaction Network is Highly Stable Across Mammals. *BMC Genom.* **2019**, *20*, 1004–1014. [CrossRef]
16. López-Bigas, N.; De, S.; Teichmann, S.A. Functional Protein Divergence in the Evolution of Homo Sapiens. *Genome Boil.* **2008**, *9*, R33. [CrossRef]
17. NCBI. Open Reading Frame Finder. Available online: https://www.ncbi.nlm.nih.gov/orffinder/ (accessed on 1 April 2020).
18. Jia, H.; Osak, M.; Bogu, G.K.; Stanton, L.W.; Johnson, R.; Lipovich, L. Genome-Wide Computational Identification and Manual Annotation of Human Long Noncoding RNA Genes. *RNA* **2010**, *16*, 1478–1487. [CrossRef]
19. Esteller, M. Non-coding RNAs in Human Disease. *Nat. Rev. Genet.* **2011**, *12*, 861–874. [CrossRef]
20. Hombach, S.; Kretz, M. Non-coding RNAs: Classification, Biology and Functioning. *Adv. Exp. Med. Biol.* **2016**, *937*, 3–17. [CrossRef]
21. Dragomir, M.P.; Knutsen, E.; Calin, G.A. SnapShot: Unconventional miRNA Functions. *Cell.* **2018**, *174*, 1038. [CrossRef]

22. Dragomir, M.; Calin, G.A. Circular RNAs in Cancer-Lessons Learned From microRNAs. *Front. Oncol.* **2018**, *8*, 179. [CrossRef] [PubMed]
23. Geisler, S.; Coller, J. RNA in Unexpected Places: Long Non-coding RNA Functions in Diverse Cellular Contexts. *Nat. Rev. Mol. Cell Boil.* **2013**, *14*, 699–712. [CrossRef] [PubMed]
24. Koerner, M.V.; Pauler, F.M.; Huang, R.; Barlow, D.P. The Function of Non-coding RNAs in Genomic Imprinting. *Development* **2009**, *136*, 1771–1783. [CrossRef] [PubMed]
25. Chooniedass-Kothari, S.; Emberley, E.; Hamedani, M.; Troup, S.; Wang, X.; Czosnek, A.; Hube, F.; Mutawe, M.; Watson, P.; Leygue, E. The Steroid Receptor RNA Activator is the First Functional RNA Encoding a Protein. *FEBS Lett.* **2004**, *566*, 43–47. [CrossRef] [PubMed]
26. Tenson, T.; DeBlasio, A.; Mankin, A. A Functional Peptide Encoded in the *Escherichia coli* 23S rRNA. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 5641–5646. [CrossRef]
27. Bileschi, M.; Belanger, D.; Bryant, D.H.; Sanderson, T.; Carter, B.; Sculley, D.; Depristo, M.A.; Colwell, L. Using Deep Learning to Annotate the Protein Universe. *bioRxiv* **2019**, 626507.
28. Mattick, J.S.; Dinger, M.E. The Extent of Functionality in the Human Genome. *HUGO J.* **2013**, *7*, 1–4. [CrossRef]
29. Lauressergues, M.; Couzigou, J.-M.; Clemente, H.S.; Martinez, Y.; Dunand, C.; Becard, G.; Combier, J.-P. Primary Transcripts of MicroRNAs Encode Regulatory Peptides. *Nature* **2015**, *520*, 90–93. [CrossRef]
30. Cai, X.; Hagedorn, C.H.; Cullen, B.R. Human microRNAs are Processed from Capped, Polyadenylated Transcripts That Can Also Function as mRNAs. *RNA* **2004**, *10*, 1957–1966. [CrossRef]
31. Jeck, W.R.; Sorrentino, J.A.; Wang, K.; Slevin, M.K.; Burd, C.E.; Liu, J.; Marzluff, W.F.; Sharpless, N.E. Circular RNAs Are Abundant, Conserved, and Associated with ALU Repeats. *RNA* **2012**, *19*, 141–157. [CrossRef]
32. Lodish, H.; Berk, A.; Kaiser, C.A.; Krieger, M.; Bretscher, A.; Ploegh, H.; Amon, A.; Martin, K.; Scott, M.P. *Molecular Cell Biology*, 8th ed.; W.H.Freeman & Co Ltd.: New York, NY, USA, 2016; p. 1280.
33. Nelson, D.L.; Cox, M.M.; Lehninger, A.L. *Lehninger Principles of Biochemistry*, 7th ed.; W.H. Freeman and Company: New York, NY, USA; Macmillan Higher Education: Houndmills, Basingstoke, 2017.
34. Desiere, F. The Peptide Atlas Project. *Nucleic Acids Res.* **2006**, *34*, D655–D658. [CrossRef] [PubMed]
35. Huang, H.-H.; Chen, F.-Y.; Chou, W.-C.; Hou, H.-A.; Ko, B.-S.; Lin, C.-T.; Tang, J.-L.; Li, C.-C.; Yao, M.; Tsay, W.; et al. Long Non-coding RNA HOXB-AS3 Promotes Myeloid Cell Proliferation and Its Higher Expression Is an Adverse Prognostic Marker in Patients with Acute Myeloid Leukemia and Myelodysplastic Syndrome. *BMC Cancer* **2019**, *19*, 1–14. [CrossRef] [PubMed]
36. Huang, J.-Z.; Chen, M.; Chen, D.; Gao, X.-C.; Zhu, S.; Huang, H.; Hu, M.; Zhu, H.; Yan, G.-R. A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Mol. Cell.* **2017**, *68*, 171–184.e6. [CrossRef]
37. Pauli, A.; Norris, M.L.; Valen, E.; Chew, G.-L.; Gagnon, J.A.; Zimmerman, S.; Mitchell, A.; Ma, J.; Dubrulle, J.; Reyon, D.; et al. Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science* **2014**, *343*, 1248636. [CrossRef]
38. D'Lima, N.G.; Ma, J.; Winkler, L.; Chu, Q.; Loh, K.H.; Corpuz, E.O.; Budnik, B.A.; Lykke-Andersen, J.; Saghatelian, A.; Slavoff, S.A. A Human Microprotein That Interacts with the mRNA Decapping Complex. *Nat. Methods* **2016**, *13*, 174–180. [CrossRef] [PubMed]
39. Galindo, M.I.; Pueyo, J.I.; Fouix, S.; Bishop, S.A.; Couso, J.P. Peptides Encoded by Short ORFs Control Development and Define a New Eukaryotic Eene Family. *PLoS Boil.* **2007**, *5*, e106. [CrossRef]
40. Cai, B.; Li, Z.; Ma, M.; Wang, Z.; Han, P.; Abdalla, B.A.; Nie, Q.; Zhang, X. LncRNA-Six1 Encodes a Micropeptide to Activate Six1 in Cis and Is Involved in Cell Proliferation and Muscle Growth. *Front. Physiol.* **2017**, *8*. [CrossRef]
41. Razooky, B.S.; Obermayer, B.; O'May, J.B.; Tarakhovsky, A. Viral Infection Identifies Micropeptides Differentially Regulated in smORF-Containing lncRNAs. *Genes* **2017**, *8*, 206. [CrossRef]
42. Makarewich, C.A.; Baskin, K.K.; Munir, A.Z.; Bezprozvannaya, S.; Sharma, G.; Khemtong, C.; Shah, A.M.; McAnally, J.R.; Malloy, C.R.; Szweda, L.I.; et al. MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-Oxidation. *Cell Rep.* **2018**, *23*, 3701–3709. [CrossRef]
43. Stein, C.S.; Jadiya, P.; Zhang, X.; McLendon, J.M.; Abouassaly, G.M.; Witmer, N.; Anderson, E.J.; Elrod, J.W.; Boudreau, R.L. Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Rep.* **2018**, *23*, 3710–3720.e8. [CrossRef]

44. Zheng, X.; Chen, L.; Zhou, Y.; Wang, Q.; Zheng, Z.; Xu, B.; Wu, C.; Zhou, Q.; Hu, W.; Jiang, J.; et al. A Novel Protein Encoded by a Circular RNA circPPP1R12A Promotes Tumor Pathogenesis and Metastasis of Colon Cancer via Hippo-YAP Signaling. *Mol. Cancer* **2019**, *18*, 47. [CrossRef] [PubMed]
45. Polycarpou-Schwarz, M.; Groß, M.; Mestdagh, P.; Schott, J.; Grund, S.E.; Hildenbrand, C.; Rom, J.; Aulmann, S.; Sinn, P.; Vandesompele, J.; et al. The Cancer-Associated Microprotein CASIMO1 Controls Cell Proliferation and Interacts with Squalene Epoxidase Modulating Lipid Droplet Formation. *Oncogene* **2018**, *37*, 4750–4768. [CrossRef] [PubMed]
46. Zhao, J.; Lee, E.E.; Kim, J.; Yang, R.; Chamseddin, B.; Ni, C.; Gusho, E.; Xie, Y.; Chiang, C.-M.; Buszczak, M.; et al. Transforming Activity of an Oncoprotein-Encoding Circular RNA From Human Papillomavirus. *Nat. Commun.* **2019**, *10*, 2300. [CrossRef] [PubMed]
47. Charpentier, M.; Croyal, M.; Carbonnelle, D.; Fortun, A.; Florenceau, L.; Rabu, C.; Krempf, M.; Labarrière, N.; Lang, F. IRES-Dependent Translation of the Long Non coding RNA Meloe in Melanoma Cells Produces the Most Immunogenic MELOE Antigens. *Oncotarget* **2016**, *7*, 59704–59713. [CrossRef] [PubMed]
48. Wu, S.; Zhang, L.; Deng, J.; Guo, B.; Li, F.; Wang, Y.; Wu, R.; Zhang, S.; Lu, J.; Zhou, Y. A Novel Micropeptide Encoded by Y-Linked LINC00278 Links Cigarette Smoking and AR Signaling in Male Esophageal Squamous Cell Carcinoma. *Cancer Res.* **2020**, *80*, 2790–2803. [CrossRef] [PubMed]
49. Guo, B.; Wu, S.; Zhu, X.; Zhang, L.; Deng, J.; Li, F.; Wang, Y.; Zhang, S.; Wu, R.; Lu, J.; et al. Micropeptide CIP 2A- BP Encoded by LINC 00665 Inhibits Triple-Negative Breast Cancer Progression. *EMBO J.* **2019**, *39*. [CrossRef] [PubMed]
50. Wang, Y.; Wu, S.; Zhu, X.; Zhang, L.; Deng, J.; Li, F.; Guo, B.; Zhang, S.; Wu, R.; Zhang, Z.; et al. LncRNA-Encoded Polypeptide ASRPS Inhibits Triple-Negative Breast Cancer Angiogenesis. *J. Exp. Med.* **2019**, *217*. [CrossRef]
51. Li, M.; Li, X.; Zhang, Y.; Wu, H.; Zhou, H.; Ding, X.; Zhang, X.; Jin, X.; Wang, Y.; Yin, X.; et al. Micropeptide MIAC Inhibits HNSCC Progression by Interacting with Aquaporin 2. *J. Am. Chem. Soc.* **2020**, *142*, 6708–6716. [CrossRef]
52. Zhang, M.; Zhao, K.; Xu, X.; Yang, Y.; Yan, S.; Wei, P.; Liu, H.; Xu, J.; Xiao, F.; Zhou, H.; et al. A Peptide Encoded by Circular Form of LINC-PINT Suppresses Oncogenic Transcriptional Elongation in Glioblastoma. *Nat. Commun.* **2018**, *9*, 4475. [CrossRef]
53. Illing, P.T.; Vivian, J.P.; Dudek, N.L.; Kostenko, L.; Chen, Z.; Bharadwaj, M.; Miles, J.J.; Kjer-Nielsen, L.; Gras, S.; Williamson, N.A.; et al. Immune Self-Reactivity Triggered by Drug-Modified HLA-peptide Repertoire. *Nature* **2012**, *486*, 554–558. [CrossRef]
54. Saini, S.K.; Ostermeir, K.; Ramnarayan, V.R.; Schuster, H.; Zacharias, M.; Springer, S. Dipeptides Promote Folding and Peptide Binding of MHC Class I Molecules. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 15383–15388. [CrossRef]
55. Pollard, K.S.; Hubisz, M.J.; Rosenbloom, K.R.; Siepel, A. Detection of Nonneutral Substitution Rates on Mammalian Phylogenies. *Genome Res.* **2009**, *20*, 110–121. [CrossRef] [PubMed]
56. Wang, J.; Zhu, S.; Meng, N.; He, Y.; Lu, R.; Yan, G.-R. ncRNA-Encoded Peptides or Proteins and Cancer. *Mol. Ther.* **2019**, *27*, 1718–1725. [CrossRef] [PubMed]
57. Washietl, S.; Kellis, M.; Garber, M. Evolutionary Dynamics and Tissue Specificity of Human Long Noncoding RNAs in Six Mammals. *Genome Res.* **2014**, *24*, 616–628. [CrossRef]
58. Necsulea, A.; Soumillon, M.; Warnefors, M.; Liechti, A.; Daish, T.; Zeller, U.; Baker, J.C.; Grützner, F.; Kaessmann, H. The Evolution of lncRNA Repertoires and Expression Patterns in Tetrapods. *Nature* **2014**, *505*, 635–640. [CrossRef]
59. Ingolia, N.T.; Ghaemmaghami, S.; Newman, J.R.S.; Weissman, J.S. Genome-Wide Analysis In Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **2009**, *324*, 218–223. [CrossRef]
60. Gebetsberger, J.; Żywicki, M.; Künzi, A.; Polacek, N. tRNA-Derived Fragments Target the Ribosome and Function as Regulatory Non-Coding RNA in Haloferax Volcanii. *Archaea* **2012**, *2012*, 1–11. [CrossRef] [PubMed]
61. Pircher, A.; Bakowska-Zywicka, K.; Schneider, L.; Żywicki, M.; Polacek, N. An mRNA-Derived Noncoding RNA Targets and Regulates the Ribosome. *Mol. Cell.* **2014**, *54*, 147–155. [CrossRef]
62. Van Heesch, S.; Van Iterson, M.; Jacobi, J.; Boymans, S.; Essers, P.B.; De Bruijn, E.; Hao, W.; MacInnes, A.W.; Cuppen, E.; Simonis, M. Extensive Localization of Long Noncoding RNAs to the Cytosol and Mono- and Polyribosomal Complexes. *Genome Boil.* **2014**, *15*, R6. [CrossRef]

63. Calviello, L.; Mukherjee, N.; Wyler, E.; Zauber, H.; Hirsekorn, A.; Selbach, M.; Landthaler, M.; Obermayer, B.; Ohler, U. Detecting Actively Translated Ppen Reading Frames in Ribosome Profiling Data. *Nat. Methods* **2015**, *13*, 165–170. [CrossRef]
64. Fournaise, E.; Chaurand, P. Increasing Specificity in Imaging Mass Spectrometry: High Spatial Fidelity Transfer of Proteins from Tissue Sections to Functionalized Surfaces. *Anal. Bioanal. Chem.* **2014**, *407*, 2159–2166. [CrossRef] [PubMed]
65. Natsume, T.; Yamauchi, Y.; Nakayama, H.; Shinkawa, T.; Yanagida, M.; Takahashi, N.; Isobe, T. A Direct Nanoflow Liquid Chromatography−Tandem Mass Spectrometry System for Interaction Proteomics. *Anal. Chem.* **2002**, *74*, 4725–4733. [CrossRef] [PubMed]
66. Chen, K.-Y.; Knoepfler, P. To CRISPR and Beyond: The Evolution of Genome Editing in Stem Cells. *Regen. Med.* **2016**, *11*, 801–816. [CrossRef] [PubMed]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).