Computational and Structural Biotechnology Journal 21 (2023) 5201–5211

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Large-scale ORF screening based on LC-MS to discover novel lncRNA-encoded peptides responding to ionizing radiation and microgravity

Wanshi Li [a,1], Yongduo Yu [a,1], Guangming Zhou [a], Guang Hu [b,c,d], Bingyan Li [a], Hong Ma [e], Wenying Yan [b,c,d,*], Hailong Pei [a,**]

[a] State Key Laboratory of Radiation Medicine and Protection, School of Radiation Medicine and Protection, Collaborative Innovation Center of Radiological Medicine of Jiangsu Higher Education Institutions, Soochow University, Suzhou 215123, China
[b] Department of Bioinformatics, School of Biology and Basic Medical Sciences, Suzhou Medical College of Soochow University, Suzhou 215123, China
[c] Center for Systems Biology, Soochow University, Suzhou 215123, China
[d] Jiangsu Province Engineering Research Center of Precision Diagnostics and Therapeutics Development, Suzhou 215123, China
[e] Beijing Key Laboratory for Separation and Analysis in Biomedicine and Pharmaceuticals, School of Life Science, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Keywords:
Space radiation
Microgravity
Long non-coding RNA
Peptide
Multi-omics integration

ABSTRACT

In the human genome, 98% of genes can be transcribed into non-coding RNAs (ncRNAs), among which lncRNAs and their encoded peptides play important roles in regulating various aspects of cellular processes and may serve as crucial factors in modulating the biological effects induced by ionizing radiation and microgravity. Unfortunately, there are few reports in space radiation biology on lncRNA-encoded peptides below 10kD due to limitations in detection techniques. To fill this gap, we integrated a variety of methods based on genomics and peptidomics, and discovered 22 lncRNA-encoded small peptides that are sensitive to space radiation and microgravity, which have never been reported before. We concurrently validated the transmembrane helix, subcellular localization, and biological function of these small peptides using bioinformatics and molecular biology techniques. More importantly, we found that these small peptides function independently of the lncRNAs that encode them. Our findings have uncovered a previously unknown human proteome encoded by 'non-coding' genes in response to space conditions and elucidated their involvement in biological processes, providing valuable strategies for individual protection mechanisms for astronauts who carry out deep space exploration missions in space radiation environments.

1. Introduction

The biomedical implications of space radiation pose a significant concern for astronauts engaged in deep space exploration missions. Hence, a thorough examination of the outer space environment, characterized primarily by space radiation and microgravity, is imperative to assess its impact on human physiology [1]. It is well-documented that both space radiation and microgravity induce substantial alterations in the transcriptome of various human cell lines and mouse models [2]. However, in the human genome, merely 2% of genes can be transcribed into mRNAs and translated into proteins. The remaining 98% can be transcribed into non-coding RNAs (ncRNAs), mainly including long non-coding RNAs (lncRNAs), microRNAs, circular RNAs, etc. Fu et al. [3] have noted that the sensitivity of some lncRNAs to space radiation and microgravity, potentially resulting in alterations to lncRNA expression profiles. Recently, emerging evidence has highlighted the essential roles played by certain lncRNAs and their encoded small peptides in regulating various biological processes and maintaining cellular homeostasis [4,5]. However, conventional methodologies encounter formidable challenges in detecting peptides with a molecular weight below 10 KD. Consequently, we take for granted that there are thousands of lncRNA-encoded small peptides [6–9], many of which may be

* Corresponding author at: Department of Bioinformatics, School of Biology and Basic Medical Sciences, Suzhou Medical College of Soochow University, Suzhou 215123, China.
** Corresponding author.
E-mail addresses: wyyan@suda.edu.cn (W. Yan), hlpei@suda.edu.cn (H. Pei).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2023.10.040
Received 5 August 2023; Received in revised form 12 October 2023; Accepted 18 October 2023
Available online 20 October 2023

an important factor in regulating the biological effects of ionizing radiation and microgravity [10,11]. Hence, comprehending the biological impacts of space radiation environment holds paramount importance. Unfortunately, the documentation of small peptides encoded by lncRNAs in this context remains limited.

Nevertheless, according to the records of ENCODE [12], the 35th edition of the Encyclopedia of DNA Elements, there are theoretically 16, 899 lncRNAs within the human genome. It is unpractical to verify the coding potential of such a vast number of lncRNAs through biological experiments, several online prediction tools for identifying open reading frames (ORFs) and technical approaches for detecting small peptides have been successfully developed and applied [13–15]. However, it is worth noting that the majority of genomics-based prediction methods mentioned above exhibit a high rate of false positives. To enhance the likelihood of discovering and confirming small peptides, integration of proteomics-based liquid chromatography-mass spectrometry (LC-MS) with large-scale genomics-based ORF prediction methods is imperative. Previously, our research endeavors have centered on investigating space-sensitive lncRNAs and their coding peptides through biological experimental methods [16]. Nonetheless, due to the aforementioned limitations, there remains a pressing need to employ this approach to identify additional relevant lncRNAs and their corresponding coding peptides. Such efforts would significantly contribute to the collective understanding of space radiation and advance the work of other researchers in this field. Regrettably, this method has yet to be employed in this field.

Therefore, to address the aforementioned challenges, our study initially obtained the most up-to-date and comprehensive lncRNA RefSeq IDs through lncRNA microarray sequencing on human cell lines. Subsequently, we turned our attention on investigating the impact of radiation and microgravity on lncRNA-encoded peptides within cells. In particular, we selected specific human cell lines, whose corresponding organs are more sensitive to space environment [17], and utilized LC-MS proteomics to validate ORF Finder genomics predictions. Subsequent comparisons with the Swiss-Prot protein database [18] led to the identification of 22 novel short peptides in human that had not been previously reported. Finally, we conducted biological experiments and applied bioinformatics analyses to investigate the subcellular localization, transmembrane helical structure, and biological function of these newfound lncRNA-encoded peptides affected by radiation and microgravity. More importantly, our investigation encompassed the analysis of the lncRNAs responsible for encoding these novel peptides. We determined their functions by constructing competing endogenous RNA (ceRNA) networks and conducting gene co-expression analyses [19]. This comprehensive approach unveiled that newly discovered lncRNA-encoded peptides possess autonomous functions independent of their host lncRNAs. In sum, our study has uncovered previously unreported lncRNA-encoded peptides that are responsive to the space environment, elucidating their biological functions and thus filling a critical void in this field of space radiation-related lncRNAs and their encoded peptides.

2. Methods and materials

2.1. Cell culture, cell irradiation, microgravity simulation and DNA damage induction

Variations in tissue and organ sensitivities to the space environment have been well-documented [17]. It is difficult to positively verify the conclusions of a single cell line. In order to improve the broad spectrum of this work and the credibility of the conclusions, we selected three human cell lines that are more sensitive to space radiation (HeLa, 293 T, and HL-60 from ATCC). These cells were cultured in Dulbecco's modified Eagle's medium (VIVACELL, USA), supplemented with 10% fetal bovine serum (Sorfa, New Zealand), 1% penicillin sodium, and 100 μg/mL streptomycin. The cells were incubated using an incubator (Thermo Fisher Scientific, Wilmington, DE, USA) at 37 ℃ with a 5% CO2 atmosphere. HeLa cells (1×10^6) were seeded into T25 flasks with slight shaking movement to disperse the cell evenly onto the flasks. Then the flasks were placed in a cell incubator overnight to allow the cells to adhere. Two groups of HeLa cells were exposed to X-ray irradiation at doses of 2 Gy and 8 Gy, respectively, and collected after 2 h. The X-ray irradiation was performed using an RS-2000 X-ray Biological Irradiator (Rad Source Technologies, Suwanee, GA, USA) at a dose rate of 1.12 Gy/min with an energy level of 160 kVp. A three-dimensional gyrometer (10 r/s, 24 h) was used to simulate microgravity for HL-60 cells. We treated 293 T cells with doxorubicin at a concentration of 1 μMol/L for 2 h to simulate DNA damage and the repair state.

2.2. Plasmid construction and transfection

The gene sequences of NR_125851.1-ORF4, XR_430028.4-ORF3, NR_003277.1-ORF1, NR_034009.1-ORF13 and NR_003148.3-ORF3 were constructed in pEGFP-C1 with *Bsp*EI/*Kpn*I restriction enzyme and in pEGFP-N1 with *Xho*I/*Age*I restriction enzyme (Sangon Biotech, Shanghai). The plasmids of co-transfected localization peptides are pcDNA3.1 (+) - mito mCherry COXVIIIA Signal Peptide/pcDNA3.1 (+) - mcherry Sec61β/ pcDNA3.1 (+) - mcherry Golgi (Sangon Biotech, Shanghai). All DNA constructs were produced using Escherichia coli DH5a (NCM Biotech, Suzhou) and extracted using E.Z.N.A.@ Plasmid Mini Kit l (Omega Bio-tek). For transient cell transfections, cells were plated into Confocal Petri dish to reach 40% confluency the following day and transfected with 1 μg plasmid DNA using Lipofectamine 3000 (Invitrogen, USA). Finally, the cellular images were captured utilizing a laser scanning confocal microscope (Olympus, Tokyo, Japan). DAPI was excited at 405 nm, EGFP was excited at 488 nm, and the endoplasmic reticulum, Golgi, and mitochondrial were excited at 555 nm.

2.3. Agilent LncRNA chip sequencing

Total RNA was extracted from all cell lines used in this study with TRIzol total RNA isolation reagent (Life Technologies). The Agilent Human LncRNA Microarray 2018 version (4 *180k, Design ID: 085630) was used for chip sequencing and data analysis of HeLa cell samples. In the experimental part, total RNA was quantified using a NanoDrop ND-2000 (Thermo Scientific), and RNA integrity was assessed by an Agilent Bioanalyzer 2100 (Agilent Technologies). After passing the RNA quality inspection, the chip was hybridized and eluted. The original image was obtained by scanning. In the data analysis part, feature extraction software (version 10.7.1.1, Agilent Technologies) was used to process the raw images to extract raw data. Then, quantile normalization and subsequent processing were performed. The normalized data were filtered and at least one set of 100% of the probes marked "P" in each sample used for comparison was retained for further analysis. Finally, the RefSeq IDs of all lncRNAs were extracted from the normalized data table resulting in 137,270 unique lncRNAs.

2.4. Peptide screening strategy based on genomics and peptidomics

We initiated our research by acquiring the FASTA sequence data encompassing a vast repertoire of 137,270 long non-coding RNAs (lncRNAs) from the NCBI Reference Sequence Database. Subsequently, we employed ORF Finder tool within the Sequence Manipulation Suite to systematically identify potential protein-coding regions within these lncRNA sequences [13]. ORF Finder searches for open reading frames (ORFs) within the entered DNA sequence and returns both their range and protein translation. In our study, the search criteria were established as follows: 1) ORFs must initiate with ATG; 2) ORFs were searched in reading frames 1, 2, and 3 on both forward and reverse strands. This exhaustive analysis, conducted on each lncRNA in a stepwise manner, yielded an extensive compendium of 837,717 unique ORFs.

Subsequently, we used the LC-MS method to detect the fingerprint

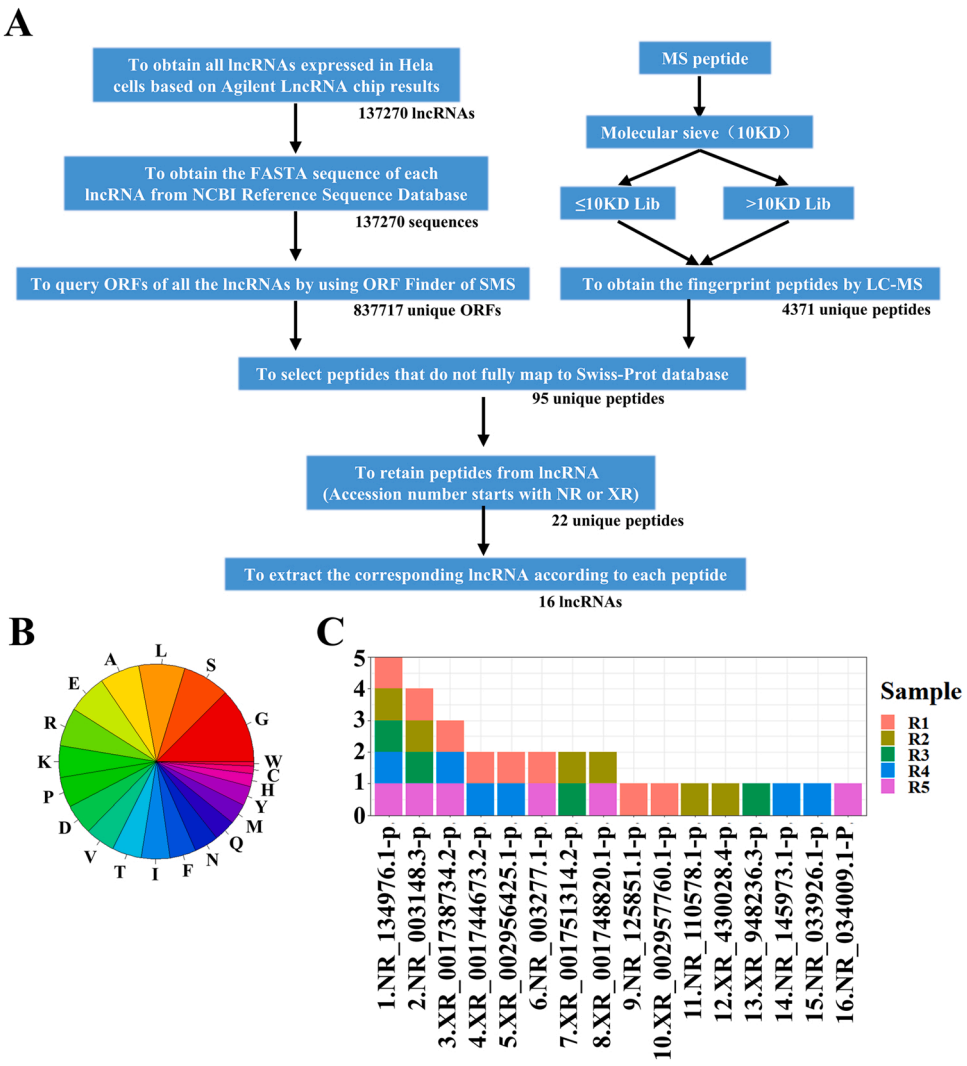

Fig. 1. Peptide and lncRNA screening methods and cell sample information. (A) Schematic diagram of screening strategies for lncRNAs and peptides. (B) Amino acid frequency of 22 distinct peptides. (C) Stacked bar graph of Hela cells in different treatment groups corresponding to the 16 lncRNA-encoded peptides, in which the serial number of each lncRNA-encoded peptide represents frequency ranking in five groups of cell samples. R1: Untreated HeLa cells; R2: 2 Gy X-rays; R3: 8 Gy X-rays; R4: DNA damage and repair group; and R5: microgravity group.

peptides expressed in five human cell lines, and finally retained 4371 different peptides. Next, we intersected these 4371 fingerprint peptides determined to be expressed in human cell lines with the 837,717 lncRNA-encoding ORFs previously found by the genomics-based ORF Finder to ensure that these fingerprint peptides are included in the sequence starting from ATG. Then, highly conserved lncRNA-encoded peptide sequences combined with genomics and peptidomics results were compared to human proteins in the Swiss-Prot Database, which is currently recognized as the most complete annotated and refined protein sequence library, using BLAST [18,20]. After conducting BLAST analysis, we found that some peptides did not exhibit 100% sequence coverage and matching rate under the Homo sapiens classification, we grouped these novel human peptides according to their LC-MS derived cell samples and uploaded them to Github (https://github.com/liwanshia/Novel-lncRNA-encoded-peptides).

The specific LC-MS experimental method is: The tryptic peptides were dissolved in solvent A, directly loaded onto a home-made reversed-phase analytical column (25-cm length, 100 μm i.d.). The mobile phase consisted of solvent A (0.1% formic acid, 2% acetonitrile/in water) and solvent B (0.1% formic acid, 90% acetonitrile/in water). Peptides were separated with the following gradient: 0–62 min, 4%– 23% B; 62–82 min, 23%– 35% B; 82–86 min, 35%– 80% B; 86–90 min, 80% B, and all at a constant flow rate of 500 nl/minon a EASY-nLC 1200 UPLC system (ThermoFisher Scientific). The separated peptides were analyzed in Q Exactive HF-X with a nano-electrospray ion source. The electrospray voltage applied was 2100 V. Precursors and fragments were analyzed at the Orbitrap detector. The full MS scan resolution was set to 120,000 for a scan range of 350–1800 *m/z*. The fragments were detected in the Orbitrap at a resolution of 15,000 and fixed first mass was set as – Up to 10 most abundant precursors were then selected for further MS/MS analyses with 30.0 s dynamic exclusion. The HCD fragmentation was performed at a normalized collision energy (NCE) of 28%. Automatic gain control (AGC) target was set at 5e4, with an intensity threshold of 2.5e5 ions/s and a maximum injection time of 40 ms. As a result, we opted to retain these novel Homo sapiens peptides encoded by lncRNA (RefSeq ID beginning with “XR” or “NR”).

Furthermore, our investigation encompassed a rigorous regimen of statistical analysis, diligently applied to the cell samples corresponding to each lncRNA-encoded peptide, ensuring the robustness and comprehensiveness of our findings.

2.5. Peptide transmembrane helix structure prediction

We utilized a support vector machine (SVM)-based TM protein

topology predictor to predict their transmembrane helix structures for the amino acid sequences of peptides [21]. Simultaneously, we generated a schematic diagram for each lncRNA-encoded peptide's ORF to exhibit its length, transmembrane helix type, and orientation.

2.6. Subcellular localization information and functional prediction of the novel lncRNA-encoded peptides

In our quest to unravel the multifaceted facets of these newly-discovered lncRNA-encoded peptides, we employed an integrated feature-based function prediction server [14] to predict subcellular localization and functional information for each peptide. The FFPred 3 is intended for assigning Gene Ontology terms to human protein chains, when homology with characterized proteins can provide little aid. Function predictions are made by scanning the input sequences against an array of Support Vector Machines (SVM). Subsequently, In order to present the location information results in a standardized and interpretable format, we undertook a normalization process for the scores obtained. This harmonization allowed us to succinctly convey the location prediction outcomes through the medium of a stacked bar graph, providing an accessible visual representation of these insights. We utilized the above method to obtain the function prediction result based on support vector machine, retained the Gene Ontology terms with SVM reliability as high and used them as the function prediction pathway result of each peptide itself [22]. These GO terms were subsequently compiled into a set of the top ten descriptors for each peptide, and these findings were artfully visualized through the generation of bubble plots, effectively encapsulating the probability scores assigned to each term. Taking a broader perspective, we embarked on an in-depth analysis of the functional attributes shared among the 16 lncRNA-encoded peptides. This entailed a systematic examination of the occurrence frequency of each GO term across this cohort. This scrutiny revealed 30 distinct terms that exhibited frequencies surpassing a threshold of 2, signifying their recurrent emergence across the lncRNA-encoded peptides. To culminate our comprehensive analysis, we proceeded to extract the genes intricately associated with these 30 enriched GO terms. By subjecting this corpus of genes to intersection analysis, we discerned those genes that were commonly shared among multiple terms. This insightful analysis shed light on the interconnectedness of these genes and their potential roles, offering valuable insights into the biological functions orchestrated by these lncRNA-encoded peptides.

2.7. Construction of the ceRNA networks of the 16 lncRNAs coding novel peptides

We employed the miRcode database to predict the target site binding to analyze the relationship between lncRNAs and miRNAs [23]. The relationship between miRNAs and their target mRNAs was analyzed using two databases: the experimentally validated miRNA-target interactions (MTIs) in the miRTarBase database [17], and the TargetScan database [18] which predicts biological target genes of miRNAs. Cytoscape software was utilized for visualizing ceRNA network location.

2.8. Co-expression analysis of lncRNAs and protein-coding genes

After randomly selecting 500 samples from the TCGA Pan-Cancer cohort, we used the cor.test function in R to calculate the Spearman correlation between each protein-coding gene and each lncRNA coding peptide in the cohort. For each lncRNA, we presented a scatterplot of its five most highly correlated expression pairs.

3. Results

3.1. Twenty-two novel lncRNA-encoded peptides were discovered by chip sequencing, large-scale ORF queries, and fingerprinting by LC-MS

Fig. 1A presents a schematic representation of the comprehensive screening process employed for the discover of novel long non-coding RNA (lncRNA)-encoded peptides. Initially, we conducted sequencing and analysis of HeLa cell line using Agilent Human lncRNA Microarray 2018 Edition, yielding RefSeq IDs for all identified human lncRNAs, resulting in the identification of 137,270 lncRNAs. Subsequently, the FASTA sequences for each lncRNA were obtained by querying their respective RefSeq IDs on the National Center for Biotechnology Information (NCBI) database. To assess the protein-coding potential of each lncRNA, specific search criteria (see Materials and methods for details) were established, and ORF searches were conducted on all lncRNAs using the ORF Finder tool from the Sequence Manipulation Suite. This analysis led to the identification of 837,717 unique open reading frames (ORFs) from the 137,270 lncRNA sequences. However, it is important to note that determine the peptide-encoding potential of a given lncRNA based solely on ORF queries may not provide entirely reliable results. Therefore, we subjected all identified ORFs to screening using liquid chromatography-mass spectrometry (LC-MS)-based peptidomics. Specifically, We have selected three commonly used human cell lines that are more sensitive to space environment from different organizations based on existing reports [17], and exposed them to varying doses of X-ray irradiation, simulated microgravity environments, and DNA damage simulations to investigate the impact of lncRNA-encoded peptides in space environment. LC-MS analysis was performed on R1 (untreated HeLa cells), R2 (HeLa cells exposed to 2 Gy X-rays), R3 (HeLa cells exposed to 8 Gy X-rays), R4 (293 T cell DNA damage and repair status induced by doxorubicin) and R5 (HL-60 cells subjected to simulated microgravity). The results indicated that only 4371 peptides out of the initial pool of 837,717 ORFs were confirmed, underscoring the necessity of validating ORF predictions with mass spectrometry proteomic data. Representative mass spectra for each sample are provided in Supplementary Figure 1. Following this, batch BLAST analysis was conducted on the results, comparing them to known human proteins in the Swiss-Prot database. Ninety-five peptides were retained, each exhibiting sequence coverage and matching rate that were not 100%. We updated the matching scores of each peptide in Table S1. Subsequently, we correlated the lncRNA information corresponding to each lncRNA-encoded peptide, ultimately identifying 22 distinct lncRNA-encoded peptides. In summary, after a series of procedures involving chip sequencing, large-scale ORF queries, and LC-MS screening, we classified these 22 peptides as newly discovered lncRNA-encoded Homo sapiens peptides associated with the space environment. Interestingly, upon closer examination of the corresponding lncRNAs for each peptide segment, we found that six of the shorter peptides were entirely encompassed by longer peptides in other segments. This means their sequences were entirely covered by the corresponding longer peptide. Additionally, these shorter and longer peptides appeared in the same lncRNA with identical ORFs. These results indicate that the abundance of lncRNA-encoded peptides we screened is high enough, proving the high confidence of the results. Consequently, in accordance with this criterion, each longer peptide corresponds precisely to one ORF and one lncRNA. The information for each peptide is presented in Table S1. The table and LC-MS results have been uploaded to GitHub (https://github.com/liwanshia/Novel-lncRNA-encoded-peptides). The sequences marked in red under the ORF column are those of lncRNA-encoded peptides. To facilitate clearer representation of this correspondence, we utilize the term 'lncRNA-p' to denote lncRNA-encoded peptides in the figure.

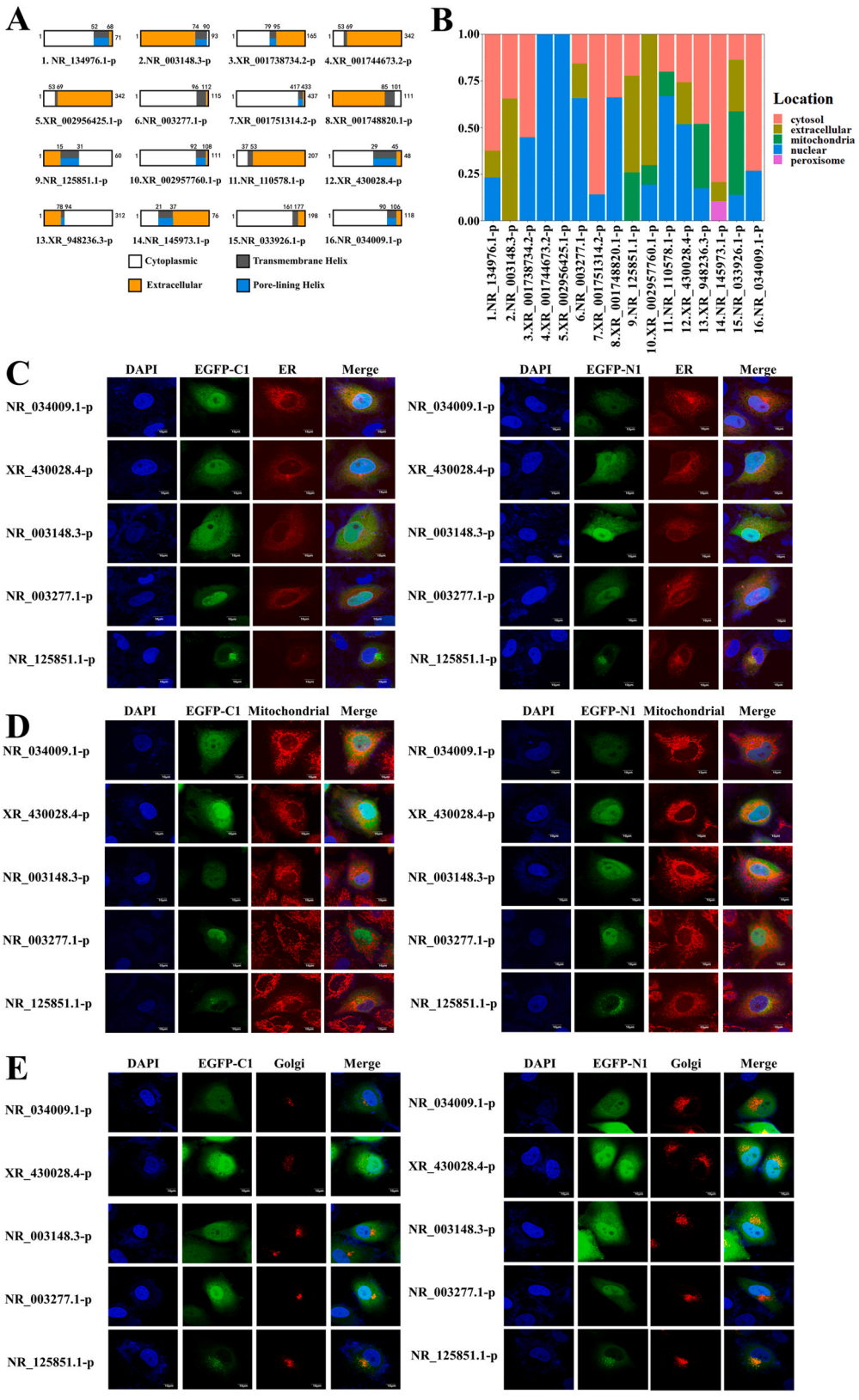

Fig. 2. Membrane helix and subcellular localization of the 16 lncRNA-encoded peptides. (A) The schematic of each lncRNA-encoded peptide. The number shows the length of ORFs and the location of the transmembrane helix structure. Each number represents an amino acid. (B) Stacked bar graph of the subcellular localization prediction of each lncRNA-encoded peptide, where the Y-axis represents the location prediction scores. (C) Subcellular localization of peptides, nucleus, and endoplasmic reticulum by immunofluorescence. Blue represents the nucleus, green represents the peptide encoded by ORF, and red represents endoplasmic reticulum. (D) Subcellular localization of peptides, nucleus, and mitochondrion by immunofluorescence. Blue represents the nucleus, green represents the peptide encoded by ORF, and red represents mitochondrion. (E) Subcellular localization of peptides, nucleus, and Golgi by immunofluorescence. Blue represents the nucleus, green represents the peptide encoded by ORF, and red represents Golgi.

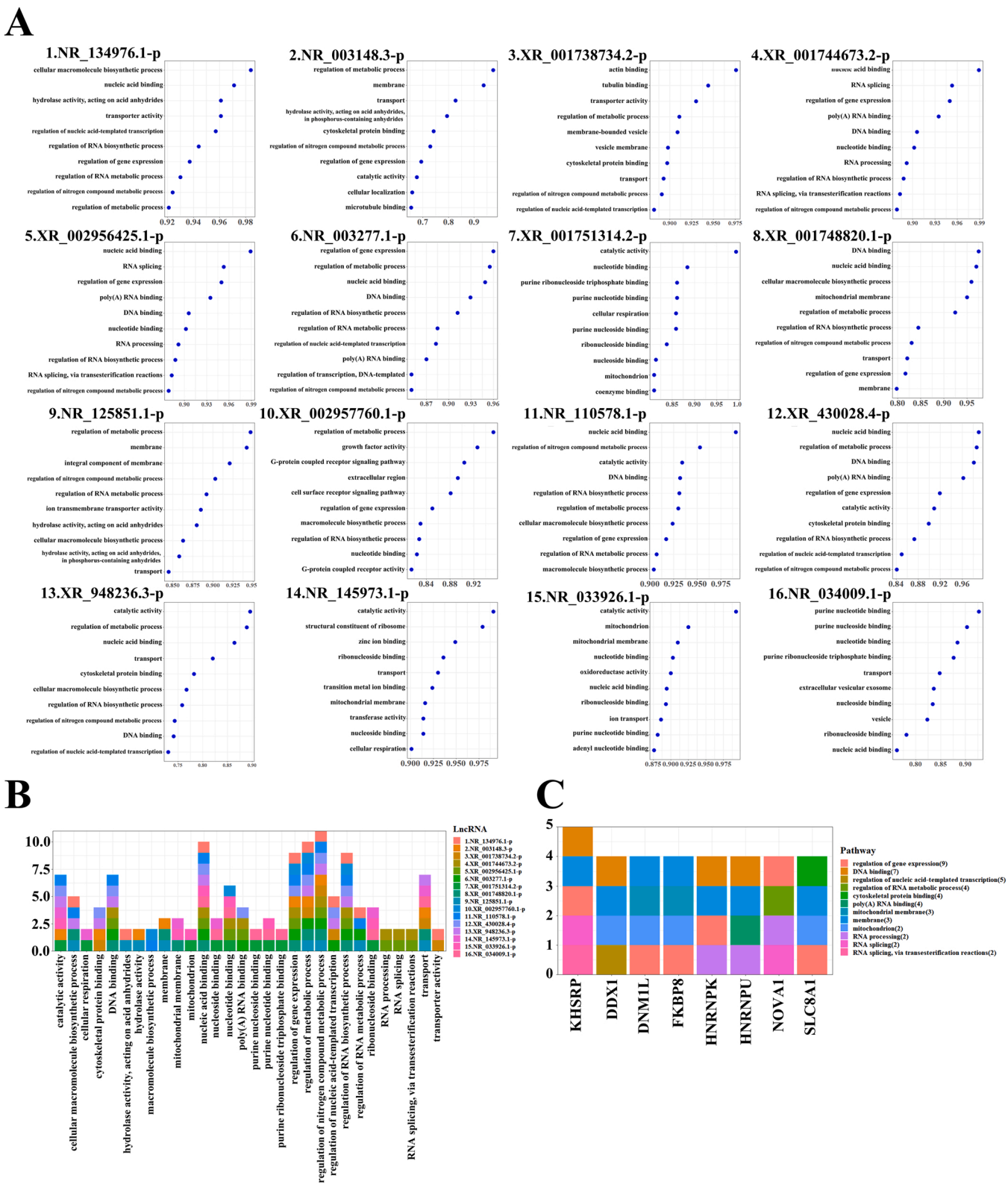

Fig. 3. Functional prediction of the 16 lncRNA-encoded peptides. (A) Bubble plots of each peptide's predicted top ten terms in GO database, where the X-axis represents the probability score. (B) Stacked bar graph shows the 30 terms that the 16 peptides were involved in (frequency≥2). (C) Stacked bar graph shows the common genes and their frequency in the 30 terms presented in (B).

3.2. Basic information on novel lncRNA-encoded peptides

Obtaining basic information on novel peptides is the first task in our research. Therefore, we initially focused on identifying which stress stimuli each peptide exhibits sensitivity and presented the results through a stacked bar graph (Fig. 1C). We ranked these lncRNA-encoded peptides based on their occurrence frequency across samples. Notably, NR_134976.1-p was consistently present in all five samples, indicating a

high level of conservation. The second and third positions were held by NR_003148.3-p and XR_001738734.2-p, respectively. Additionally, five lncRNA-encoded peptides were detected in two samples, while eight were unique to one sample.

In the field of bioinformatics, information such as amino acid sequence and occurrence frequency of occurrence plays a role in predicting protein function [24]. Proteins are constructed of 20 kinds of natural amino acids through linear combinations, And these linear sequences encapsulate significant biological information, often considered capable of elucidating and representing vital cellular processes across various organisms. Consequently, amino acid sequence and occurrence frequencies are typically utilized as information sources, often in combination with machine learning technology to predict protein function. In our study, we conducted an integration and calculation of amino acid frequencies for the 16 ORF-peptides, simultaneously, the amino acid frequencies of each individual peptide were calculated separately, as depicted in Fig. 1B and Table S1. In our previous screening strategy, all ORFs initiated ATG as the start codon, however, the methionine was not the most prevalent. Instead, the frequency of glycine (12.47%) was the highest, followed by serine (7.81%) and leucine (7.63%). Glycine, a significant components of endogenous redox proteins, plays a critical role in the synthesis of antioxidant proteins. This result may be attributed to cellular stress responses encountered under pressure conditions such as radiation and DNA damage.

In Fig. 2A, each named rectangular structure represents an identified peptide, and we can clearly see their length, transmembrane helix region and direction. For instance, taking the lncRNA-encoded peptide NR_134976.1-p as an example, it exhibits an ORF length of 70 amino acids with a transmembrane helix positioned between amino acids 52–68 and an inward-to-outward orientation. Information for each lncRNA-encoded peptide is presented in a similar format as described above. Remarkably, out of the 16 peptides studied, 10 possess a transmembrane region comprising both a transmembrane helix and pore-lining helix [21]. This finding suggests a high degree of stability in their protein secondary structure and implies crucial biological functions associated with these peptides.

3.3. Subcellular localization of the novel lncRNA-encoded peptides

Proper positioning is crucial for the correct functioning of most proteins. Therefore, we employed the PSIPRED workbench to predict the subcellular localization of the novel peptides we discovered. Our analysis, based on normalized prediction scores, revealed that these 16 lncRNA-encoded peptides exhibited distributed across various cellular compartments, including the cytosol, extracellular space, mitochondria, nucleus and peroxisome (Fig. 2B). To validate the accuracy of our predictions, we constructed the gene sequences of the micro peptides into two vectors, pEGFP-N1 and pEGFP-C1, respectively. These plasmids were co-transfected with endoplasmic reticulum, Golgi, and mitochondrial localization peptides into Hela cells. Then we observed the intracellular localizations of these peptides. The results demonstrated consistency with our predictions (Fig. 2C-E). In summary, our findings indicate that these 16 peptides are distributed across a range of subcellular structures, suggesting a potential diversity of biological functions associated with them.

3.4. Biological function prediction of the novel lncRNA-encoded peptides

To validate our hypothesis, we proceeded to conduct functional predictions for each of the lncRNA-encoded peptide. The FFPred server utilizes machine learning methods to predict protein function from amino acid sequences in the protein feature space [14]. By inputting the amino acid sequence, the FFPred server can predict subcellular localization information and Gene Ontology (GO) terms highly correlated with the amino acid sequence through Support Vector Machines (SVM). We employed FFPred to perform the GO database pathway enrichment analysis for each peptide and displayed the top ten term probability scores (Fig. 3A), offering insights into potential biological functions associated with each peptide. As shown in the Fig. 3A, all pathway scores exceeded 0.7, with the majority surpassing 0.8, signifying the reliability of our predictions. Upon statistical analysis, we observed that certain GO terms appeared in the top ten terms for different peptides simultaneously. The term regulation of nitrogen compound metabolic process was the highest among them, ranking in the top ten in pathway enrichment analysis results of 11 lncRNA-encoded peptides. Additionally, nucleic acid binding and regulation of metabolic processes were also enriched in 10 lncRNA-encoded peptides. Consequently, we summarized and compared the thirty pathways that appeared in two or more peptides in Fig. 3B to identify significantly enriched peptides. Subsequently, we shifted our focus from the pathway level to the gene level. We extracted all the genes in the 30 terms in Fig. 3B to explore common biological functions of the lncRNA-encoded peptides and identify shared functional genes. In Fig. 3C, we presented eight genes that were shared by four or more terms and their corresponding relationships with terms. Among these genes, the KH-type splicing regulator protein (KHSRP) is a single-stranded nucleic acid binding protein that widely exists in the nucleus and cytoplasm. It mainly promotes degradation by binding with unstable mRNA and regulates the maturation of microRNA to achieve post-transcriptional regulation. KHSRP was first reported as an enhancer upstream of the c-myc oncogene promoter. Further research has shown that KHSRP not only participates in cell proliferation, cell differentiation, the inflammatory response, natural immunity and lipid metabolism, but also changes in its expression level and protein structure are closely related to the occurrence and development of tumors [25,26]. SLC8A1, a member of the solute carrier family, is mainly involved in active transport of calcium and sodium ions, indicating its potential role in regulating mitochondrial stress. Overall, based on our predictions of biological functions associated with these peptides, we found shared pathways and genes among all 16 peptides. Additionally, we analyzed each novel peptide individually to reveal their functional properties.

3.5. Independence of the function of novel lncRNA encoded peptides

The protein-coding function of lncRNAs has recently attracted increasing attention. In our research, we made a remarkable discovery find that each longer novel peptide corresponds to a specific lncRNA. This led us to hypothesize: Do the biological functions of peptides relate to their corresponding lncRNA functions? To investigate the hypothesis, we conducted a comprehensive analysis involving competing endogenous RNA (ceRNA) network analysis and gene co-expression analysis on the 16 lncRNAs. Additionally, we explored their biological functions through enrichment analysis of protein-coding genes associated with their structure or expression.

In 2011, Salmena et al. introduced the concept of competing endogenous RNAs (ceRNAs) [19], which describes the competitive interactions among RNAs, including lncRNAs, for the common binding sites of target miRNAs, thereby altering the function of target miRNAs. We constructed comprehensive and independent ceRNA networks for lncRNAs (see Materials and methods for details). Using BioMart, we convert the RefSeq IDs of lncRNAs into Ensembl gene IDs to obtain the interactions between lncRNAs and miRNAs in the miRcode database [23,27]. Notably, we observed interactions for 10 out of the 16 lncRNAs under investigation. Moreover, during the gene format conversion process, we made an unexpected discovery that lncRNA XR_001744673.2 and lncRNA XR_002956425.1 share the same Ensembl gene ID, as do lncRNA NR_110578.1 and lncRNA XR_430028.4, which again demonstrates the reliability of our mass spectrometry screening method. The miRNA target genes were obtained from both the miRTarBase database and TargetScan database based on the identified miRNAs [28,29]. Accordingly, 161 nodes and 194 edges were included in our lncRNA comprehensive ceRNA network (Supplementary Figure 2). We performed GO pathway enrichment analysis for the 142 mRNAs in the

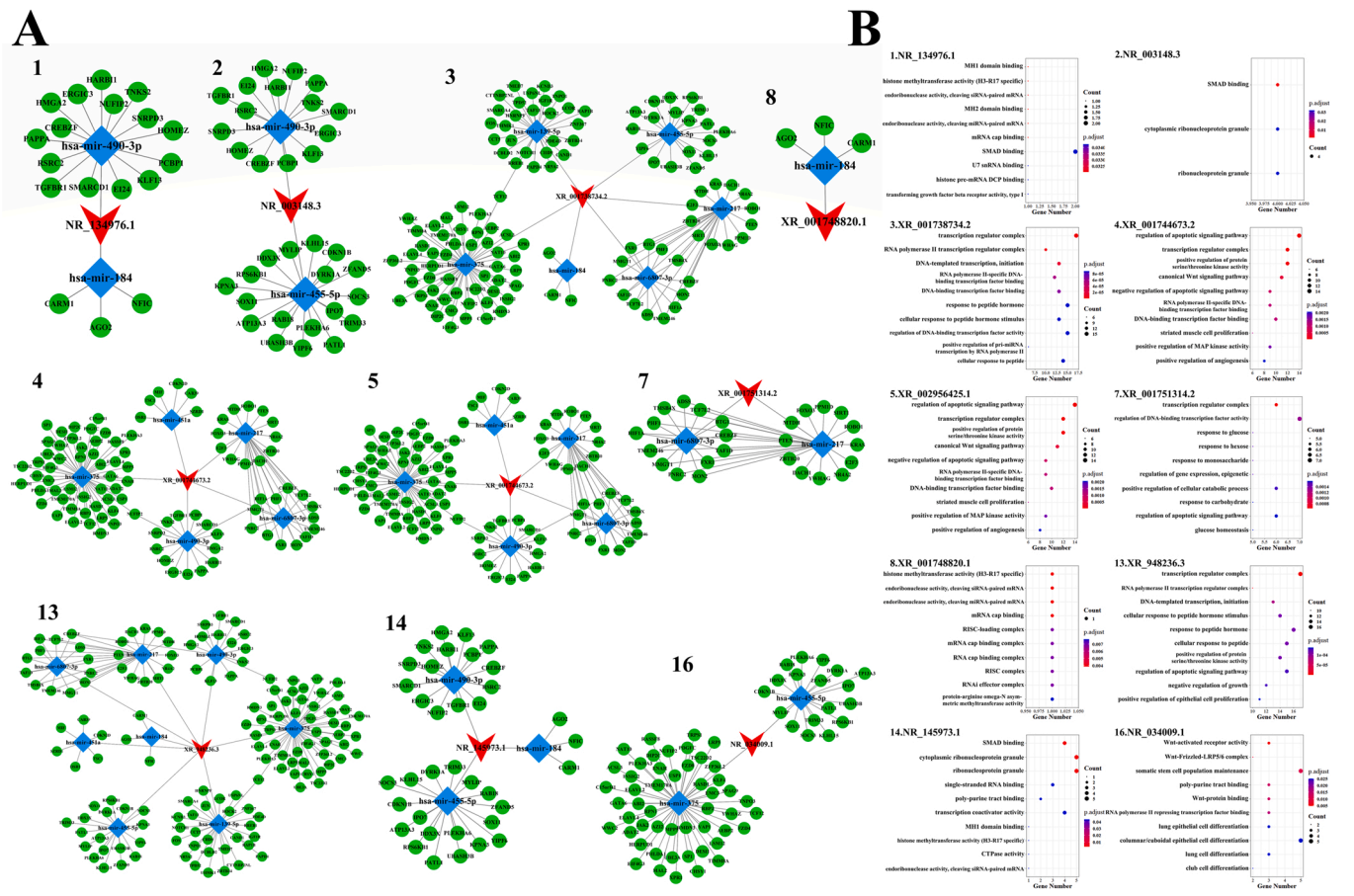

Fig. 4. Construction of ceRNA network of the lncRNAs coding novel peptides. (A) The lncRNA-miRNA-mRNA competing endogenous RNA networks. The light green ellipses indicate mRNAs, light red V shapes represent lncRNAs, and blue diamonds represent miRNAs. The number on the top left of each network represents the serial number of lncRNAs. (B) Bubble plots of GO database pathway enrichment analysis of lncRNA-related mRNAs.

Table 1
The biological functions of the peptide encoded by the 16 lncRNAs and that of the mRNAs correlated with the lncRNAs.

ID	Number of GO terms predicted by peptides	Number of GO terms analyzed by mRNA from ceRNA	Jaccard similarity index
1. NR_134976.1	101	65	0
2. NR_003148.3	38	3	0
3. XR_001738734.2	56	520	0
4. XR_001744673.2	35	422	0
5. XR_002956425.1	35	422	0
6. NR_003277.1	44	/	/
7. XR_001751314.2	34	423	0
8. XR_001748820.1	41	207	0
9. NR_125851.1	61	/	/
10. XR_002957760.1	57	/	/
11. NR_110578.1	56	/	/
12. XR_430028.4	40	/	/
13. XR_948236.3	30	647	0
14. NR_145973.1	91	10	0
15. NR_033926.1	79	/	/
16. NR_034009.1	41	51	0

network, which revealed significant enrichment in 647 terms. This suggests that the common functions of these novel lncRNA-encoded peptides may be relatively concentrated. The top 20 terms are displayed in the bubble plot (Supplementary Figure 3). Likewise, we constructed ten ceRNA networks for the lncRNAs that can be converted to Ensembl gene IDs and performed pathway enrichment analysis on the mRNAs in them (Fig. 4A, B). Importantly, we calculated the Jaccard similarity index to determine the correlation between the pathway enrichment analysis results of peptides and their corresponding lncRNAs in the ceRNA network of mRNAs. The Jaccard similarity index is used to compare similarities and differences between finite sample sets. The higher the Jaccard similarity index, the greater the sample similarity. Interestingly, Table 1 indicates that all Jaccard similarity indexes are zero, signifying that each novel lncRNA-encoded peptide we found functions independently in biological processes rather than through its corresponding ceRNA network mechanism hypothesis.

Gene co-expression analysis serves as a means to compare genes with unknown functions to those with known functions, including their

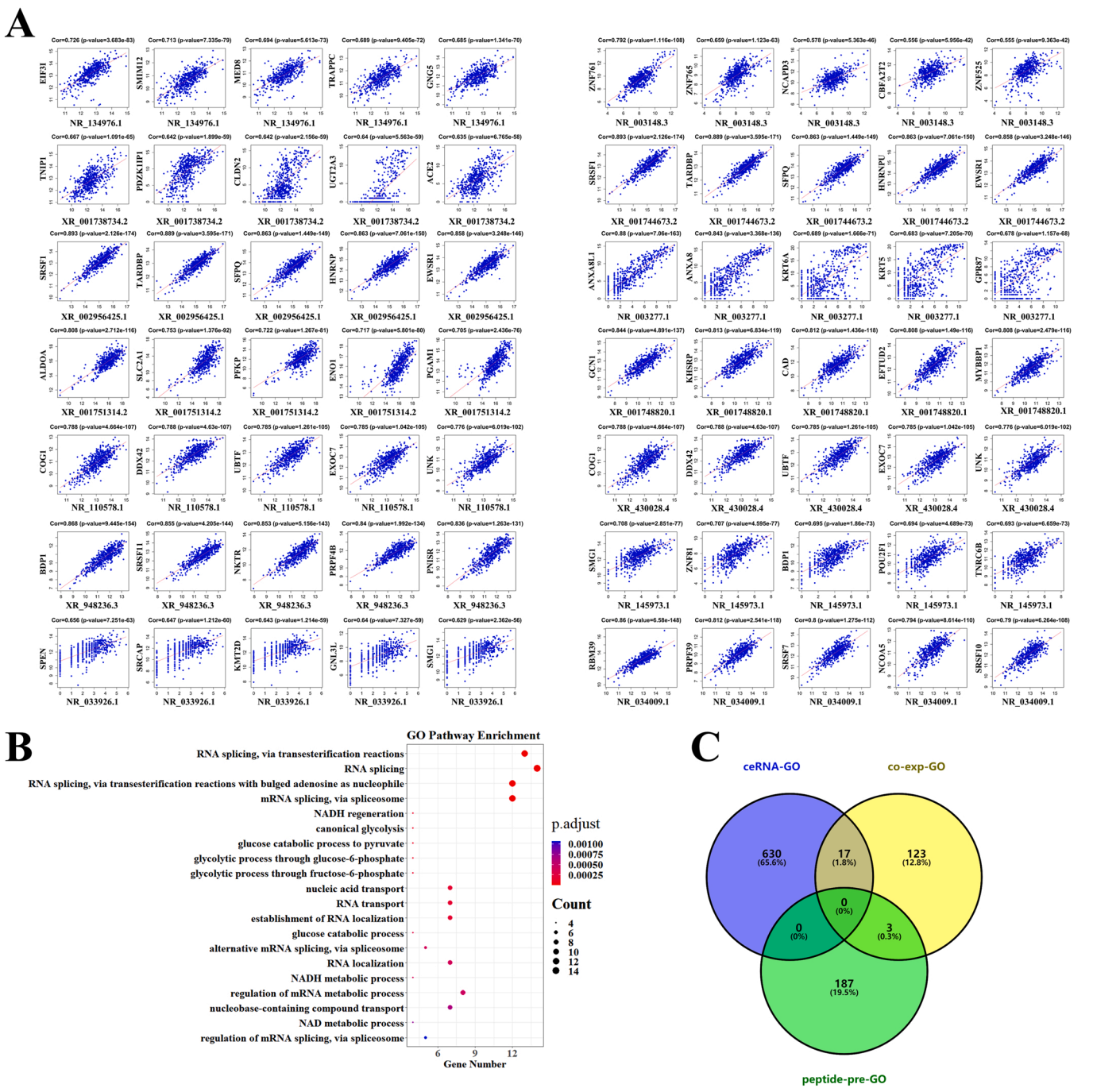

Fig. 5. Gene co-expression and functional analysis of the lncRNAs coding novel peptides. (A) Correlations between the lncRNAs and protein-coding genes in the TCGA Pan-Cancer cohort. Only the top five genes were presented for each lncRNA. The coefficients were obtained by using Spearman correlation analysis. (B) Bubble Plot of GO database pathway enrichment analysis for all the protein-coding genes in (A). (C) Venn diagram of overlapping pathways in three functional enrichment analysis results.

involvement in biological processes. This method specifically correlates gene expression profiles to determine the biological functions of novel lncRNA-encoded peptides at the gene expression level. Consequently, we randomly selected 500 case samples from the TCGA Pan-Cancer cohort, calculated the Spearman correlation between each protein-coding gene and lncRNA one by one, and we presented the top five genes with the highest correlation for each lncRNA in Fig. 5A. The ninth and tenth lncRNAs lacked corresponding Ensembl gene IDs during the gene conversion stage. Notably, the vast majority of correlations exceeded 0.6, indicating that our lncRNAs coding novel peptides are similar to protein-coding genes with known biological functions at the gene expression level. To determine the specific functions involved in these lncRNAs, we performed GO database pathway enrichment analysis on the mRNAs in Fig. 5A. The results indicated that the co-expression mRNAs of lncRNAs encoding these peptides were significantly enriched in RNA splicing and its related pathways (Fig. 5B), a crucial biological process in the gene expression of eukaryotic cells. Through RNA splicing, many functional mRNAs with coding information can be generated, which is essential for biological development and evolution [30]. Finally, we compared the results of the following three sets of functional enrichment analyses:

1. The overall biological function prediction results of the 16 novel peptides, which include all pathways significantly enriched in the prediction results for each peptide, comprise a total of 190 terms.
2. Results of functional enrichment analysis and grading of mRNAs in the comprehensive ceRNA network of the 16 lncRNAs coding novel peptides, with a total of 647 terms.
3. In the lncRNA co-expression analysis, the highly correlated protein-coding genes were subjected to corresponding functional enrichment analysis, resulting in a total of 143 terms.

The Venn diagram illustrated that there was no intersection among the three sets of pathway enrichment analysis results (Fig. 5C), particularly between the lncRNA-encoded peptide's predicted pathway enrichment analysis results and the other two sets of lncRNA-derived results, which exhibited minimal overlap. Thus, we believe that the biological functions of the novel peptides may the biological functions of the novel peptides are determined by themselves, rather than being regulated by the ceRNA hypothesis or genes with high correlation in gene co-expression analysis. This clarifies their functions under the influence of space environment and demonstrates their independence in performing these functions, providing valuable insights for the study of lncRNA-encoded peptides in the context of space radiation and microgravity.

4. Discussion and conclusions

LncRNA was initially considered the "noise" of genome transcription, a byproduct of RNA polymerase II transcription, and does not have biological functions. However, many studies have shown that lncRNA plays a vital role in numerous life processes and has begun attracting widespread attention [31]. In addition, with the development of second-generation sequencing technology, many lncRNAs and their small regulatory open reading frames (ORFs), functional peptides, and possible non-functional proteins have been discovered. Finding these hidden resources will help us understand the biological mechanisms of the impact of space radiation on the human body. Therefore, we developed an integrated genomics and peptide omics method that successfully predicted 22 novel lncRNA-encoded peptides from human sources. In addition, we determined the physical properties, subcellular localization, and unique biological functions of these peptides through bioinformatics analysis and biological experiments [32,33]. Notably, we have focused on common genes that may have similar functions to lncRNA-encoded peptides. The results showed that the KH-type splicing regulatory protein (KHSRP) appeared in five high-frequency words. It was reported that RBP (RNA binding protein) represented by KHSRP played a tumor-promoting role through post-translational modification. This space radiation-sensitive oncogene ranked first in our ranking, again proving the effectiveness of our screening method.

In addition, in the study of lncRNA encoded peptides, ribosome profiling sequencing techniques have provided evidence that many small ORFs are translated outside annotated coding sequences. Many of these ORFs have been discovered thanks to the development of ribosome profiling, a technique to sequence ribosome-protected RNA fragments. Ribosome profiling can accurately predict whether lncRNA encodes peptides based on experiments. The extended conduct of this experience has produced a series of databases containing ribo-seq (based on experimental validation and omics prediction). Therefore, based on experiments and predictions, we verified the expression of the novel lncRNA-encoded peptides we screened in the ribosome profiling database. The first method is the SmProt [34], in this database, the selected small proteins were identified from ribosome profiling data, literature, mass spectroscopy (MS), etc. The ribosome profiling data in this database are all based on experimental validation. We extracted data in which the species is of human origin, the Start Codon is AUG, and the data source is the ribosome profiling database. Then, we searched the novel lncRNA-encoded peptide obtained by screening in the modified library and saved the search results in Table S2. In addition, we also observed the same conclusion through the prediction-based ribosome profiling database named GWIPS (Supplementary Figure 4) [35]. These results based on ribosome profiling are also strong evidence for the expression of 22 novel lncRNA-encoded peptides.

Admittedly, there is still room for improvement in this research. For example, our simulation of the complex situation of space radiation needs to be more realistic [36]. Additionally, the function of lncRNA-encoded peptides has not been verified at the biological experience level, and research on each lncRNA-encoded peptide has yet to be in-depth. In general, their functions are highly consistent, but the functions of some lncRNA encoded peptides are specific and concentrated, such as XR_ 002957760.1-p, which focuses on the cell membrane and G-protein coupled receiver-related paths. It may be confirmed that under the effects of the space environment, there may be changes in the GPCR-related paths of the lncRNA-encoded peptide. Therefore, we will conduct similar studies on lncRNA-encoded peptides in subsequent work.

In summary, we found 22 novel lncRNA-encoded peptides that are sensitive to the space environment by combining lncRNA chip sequencing, large-scale ORF screening and LC-MS, verified their transmembrane helix, subcellular localization and functional analysis through bioinformatics analysis and biological experiments. Our work has discovered the hidden human protein that is encoded by 'non-coding' genes and influenced by the space environment. This study fills a gap in lncRNA-encoded peptides in space radiation biology and attributes them to protect astronauts from the effects of space radiation and microgravity environments.

Fundings

This work was supported by the National Natural Science Foundation of China (82192883, 82273578, 12275191) and Space Medical Experiment Project of China Manned Space Program (HYZHXM02003).

CRediT authorship contribution statement

Hailong Pei and Wenying Yan: Designed the research and revised all data. **Wanshi Li and Yongduo Yu:** Wrote the paper, conducted all bioinformatics analysis and experiments and contributed equally to this work. **Guangming Zhou and Guang Hu:** Revised the manuscript and proposed modification suggestions. **Bingyan Li and Hong Ma:** Provided some research directions.

Declaration of Competing Interest

The authors declare no potential conflicts of interest.

Acknowledgements

The authors wish to sincerely thank Dr. Weiwei Pei, Caiyong Ye, Wentao Hu, Ningang Liu and Jing Nie for their helpful discussion.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.10.040.

References

[1] Isasi E, Isasi ME, van Loon J. The application of artificial gravity in medicine and space. Front Physiol 2022;13:952723. https://doi.org/10.3389/fphys.2022.952723.
[2] da Silveira WA, Fazelinia H, Rosenthal SB, Laiakis EC, Kim MS, Meydan C, et al. Comprehensive multi-omics analysis reveals mitochondrial stress as a central biological hub for spaceflight impact. e20 Cell 2020;183(5):1185–201. https://doi.org/10.1016/j.cell.2020.11.002.

[3] Fu H, Su F, Zhu J, Zheng X, Ge C. Effect of simulated microgravity and ionizing radiation on expression profiles of miRNA, lncRNA, and mRNA in human lymphoblastoid cells. Life Sci Space Res (Amst) 2020;24:1–8. https://doi.org/10.1016/j.lssr.2019.10.009.
[4] Cai B, Li Z, Ma M, Wang Z, Han P, Abdalla BA, et al. LncRNA-Six1 encodes a micropeptide to activate Six1 in Cis and is involved in cell proliferation and muscle growth. Front Physiol 2017;8:230. https://doi.org/10.3389/fphys.2017.00230.
[5] Szafron LM, Balcerak A, Grzybowska EA, Pienkowska-Grela B, Felisiak-Golabek A, Podgorska A, et al. The novel gene CRNDE encodes a nuclear peptide (CRNDEP) which is overexpressed in highly proliferating tissues. PLOS One 2015;10(5):e0127475. https://doi.org/10.1371/journal.pone.0127475.
[6] Matsumoto A, Pasut A, Matsumoto M, Yamashita R, Fung J, Monteleone E, et al. mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. Nature 2016;541(7636):228–32. https://doi.org/10.1038/nature21034.
[7] Zhang M, Zhao K, Xu X, Yang Y, Yan S, Wei P, et al. A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. Nat Commun 2018;9(1). https://doi.org/10.1038/s41467-018-06862-2.
[8] Lu S, Zhang J, Lian X, Sun L, Meng K, Chen Y, et al. A hidden human proteome encoded by 'non-coding' genes. Nucleic Acids Res 2019;47(15):8111–25. https://doi.org/10.1093/nar/gkz646.
[9] Li Y, Zhang J, Sun H, Chen Y, Li W, Yu X, et al. lnc-Rps4l-encoded peptide RPS4XL regulates RPS6 phosphorylation and inhibits the proliferation of PASMCs caused by hypoxia. Mol Ther 2021;29(4):1411–24. https://doi.org/10.1016/j.ymthe.2021.01.005.
[10] Liu C, Gao X, Li Y, Sun W, Xu Y, Tan Y, et al. The mechanosensitive lncRNA Neat1 promotes osteoblast function through paraspeckle-dependent Smurf1 mRNA retention. Bone Res 2022;10(1). https://doi.org/10.1038/s41413-022-00191-3.
[11] Wang Y, Wang K, Zhang L, Tan Y, Hu Z, Dang L, et al. Targeted overexpression of the long noncoding RNA ODSM can regulate osteoblast function in vitro and in vivo. Cell Death Dis 2020;11(2). https://doi.org/10.1038/s41419-020-2325-3.
[12] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489(7414):57–74. https://doi.org/10.1038/nature11247.
[13] Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. 4 BioTechniques 2000;28(6):1102. https://doi.org/10.2144/00286ir01.
[14] Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nat Protoc 2012;7(8):1534–50. https://doi.org/10.1038/nprot.2012.086.
[15] Casimiro-Soriguer CS, Rigual MM, Brokate-Llanos AM, Muñoz MJ, Garzón A, Pérez-Pulido AJ, et al. Using AnABlast for intergenic sORF prediction in the Caenorhabditis elegans genome. Bioinformatics 2020;36(19):4827–32. https://doi.org/10.1093/bioinformatics/btaa608.
[16] Pei H, Dai Y, Yu Y, Tang J, Cao Z, Zhang Y, et al. The tumorigenic effect of lncRNA AFAP1-AS1 is mediated by translated peptide ATMLP under the control of m(6) A methylation. Adv Sci 2023;10(13):e2300314. https://doi.org/10.1002/advs.202300314.
[17] Cucinotta FA, Durante M. Cancer risk from exposure to galactic cosmic rays: implications for space exploration by human beings. Lancet Oncol 2006;7(5):431–5. https://doi.org/10.1016/S1470-2045(06)70695-7.
[18] Consortium TU. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 2020;49(D1):D480–9. https://doi.org/10.1093/nar/gkaa1100.
[19] Salmena L, Poliseno L, Tay Y, Kats L, Pandolfi PP. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell 2011;146(3):353–8. https://doi.org/10.1016/j.cell.2011.07.014.
[20] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215(3):403–10. https://doi.org/10.1016/s0022-2836(05)80360-2.
[21] Nugent T, Jones DT. Detecting pore-lining regions in transmembrane protein sequences. BMC Bioinforma 2012;13(1):169. https://doi.org/10.1186/1471-2105-13-169.
[22] Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. D419-d26 Nucleic Acids Res 2019;47(D1). https://doi.org/10.1093/nar/gky1038.
[23] Jeggari A, Marks DS, Larsson E. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. Bioinformatics 2012;28(15):2062–3. https://doi.org/10.1093/bioinformatics/bts344.
[24] Lehmann J, Libchaber A, Greenbaum BD. Fundamental amino acid mass distributions and entropy costs in proteomes. J Theor Biol 2016;410:119–24. https://doi.org/10.1016/j.jtbi.2016.08.011.
[25] Yan M, Sun L, Li J, Yu H, Lin H, Yu T, et al. RNA-binding protein KHSRP promotes tumor growth and metastasis in non-small cell lung cancer. J Exp Clin Cancer Res 2019;38(1):478. https://doi.org/10.1186/s13046-019-1479-2.
[26] Taniuchi K, Ogasawara M. KHSRP-bound small nucleolar RNAs associate with promotion of cell invasiveness and metastasis of pancreatic cancer. Oncotarget 2020;11(2):131–47. https://doi.org/10.18632/oncotarget.27413.
[27] Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart–biological queries made easy. BMC Genom 2009;10:22. https://doi.org/10.1186/1471-2164-10-22.
[28] Huang H-Y, Lin Y-C-D, Cui S, Huang Y, Tang Y, Xu J, et al. miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions. Nucleic Acids Res 2021;50(D1):D222–30. https://doi.org/10.1093/nar/gkab1079.
[29] McGeary SE, Lin KS, Shi CY, Pham TM, Bisaria N, Kelley GM, et al. The biochemical basis of microRNA targeting efficacy. Science 2019;366(6472). https://doi.org/10.1126/science.aav1741.
[30] Montes M, Sanford BL, Comiskey DF, Chandler DS. RNA splicing and disease: animal models to therapies. Trends Genet 2019;35(1):68–87. https://doi.org/10.1016/j.tig.2018.10.002.
[31] Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. Nat Chem Biol 2013;9(1):59–64. https://doi.org/10.1038/nchembio.1120.
[32] Cozzetto D, Minneci F, Currant H, Jones DT. FFPred 3: feature-based function prediction for all Gene Ontology domains. Sci Rep 2016;6:31865. https://doi.org/10.1038/srep31865.
[33] Zhu S, Wang JZ, Chen D, He YT, Meng N, Chen M, et al. An oncopeptide regulates m(6)A recognition by the m(6)A reader IGF2BP1 and tumorigenesis. Nat Commun 2020;11(1):1685. https://doi.org/10.1038/s41467-020-15403-9.
[34] Li Y, Zhou H, Chen X, Zheng Y, Kang Q, Hao D, et al. SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. Genom, Proteom Bioinforma 2021;19(4):602–10. https://doi.org/10.1016/j.gpb.2021.09.002.
[35] Michel AM, Fox G, M. Kiran A, De Bo C, O'Connor PBF, Heaphy SM, et al. GWIPS-viz: development of a ribo-seq genome browser. Nucleic Acids Res 2014;42(D1):D859–64. https://doi.org/10.1093/nar/gkt1035.
[36] Blakely EA. Biological effects of cosmic radiation: deterministic and stochastic. Health Phys 2000;79(5):495–506. https://doi.org/10.1097/00004032-200011000-00006.