# Supporting information

Deeply mining a universe of peptides Encoded by Long Noncoding RNAs

Qing Zhang[1,3,*], Erzhong Wu[2,3,*], Yiheng Tang[2,3,*], Tanxi Cai[1,3*], Lili Zhang[2,3], Jifeng Wang[1,3], Yajing Hao[2,3], Bao Zhang[2,3], Yue Zhou[1,3,4], Xiaojing Guo[1,3], Jianjun Luo[2,3#], Runsheng Chen[2,3,5#], Fuquan Yang[1,3#]

**Figure supplement 1**. Comparison of different strategies for lncRNA-encoded SEPs discovery.
**Figure supplement 2**. Representative mass spectra of several identified lncRNA-SEP peptides.
Separated file
**Figure supplement 3**. Technical replicates improve the identification of lncRNA-encoded SEPs in both cells and tissues.
**Figure supplement 4**. Bioinformatic analysis of mouse SEPs-coding lncRNAs.
**Figure supplement 5**. Peptide mass spectra information for the endogenous and synthetical peptides.
Separated file
**Figure supplement 6**. GFP and SEP-GFP fusion constructs used in current study.
**Figure supplement 7**. Full size western blotting results (Corresponding to figure 4).
**Figure supplement 8**. Characterization of identified SEPs and their corresponding lncRNA transcripts in mouse.
**Table supplement 1**. List of the primers used for plasmid construction in current study.
**Table supplement 2**. List of the primers used for RT-PCR-based analysis of the transcription of the GFP and SEP-GFP fusion coding sequences.
**Table supplement 3**. List of total MS-identified human SEPs.
Separated file
**Table supplement 4**. List of total MS-identified mouse SEPs.
Separated file
**Table supplement 5**. List of non-synthetic peptide based PRM validated SEPs.

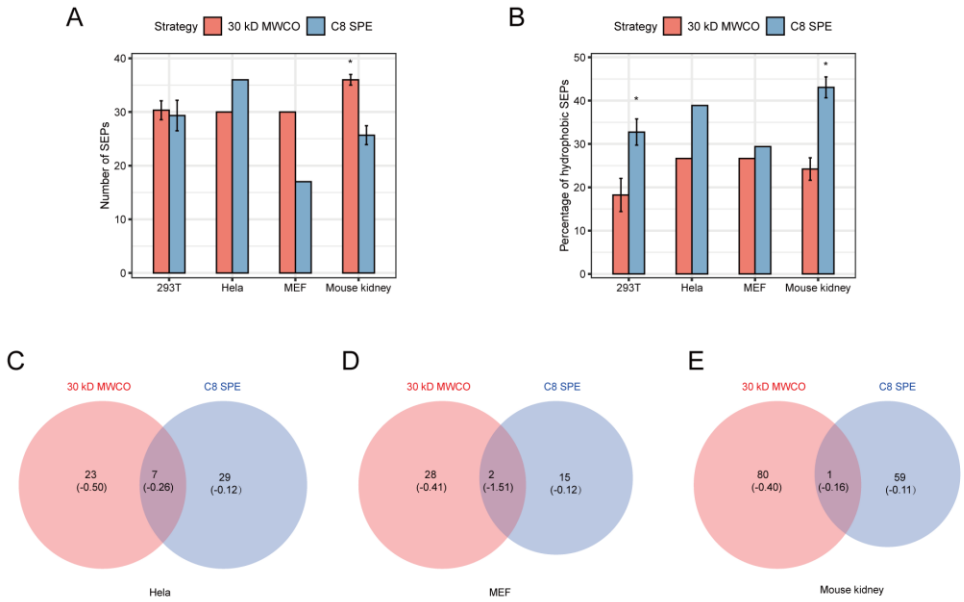


**Figure supplement 1**. Comparison of different strategies for lncRNA-encoded SEPs discovery. **(A)** The number of SEPs identified in the fractions enriched by 30-kD MWCO membrane and C8 SPE column. **(B)** Percentage of hydrophobic SEPs identified in the fractions enriched by 30-kD MWCO membrane and C8 SPE column. **(C)** Venn diagram of SEPs identified in the fractions enriched by 30-kD MWCO membrane and C8 SPE column from mouse kidney lysate. (The average Gravy values of SEPs in each region were shown in the parentheses). **(D)** Venn diagram of SEPs in the fractions enriched by 30-kD MWCO membrane and C8 SPE column from Hela cell lysates. **(E)** Venn diagram of SEPs identified in the fractions enriched by 30-kD MWCO membrane and C8 SPE column from MEF cell lysates.

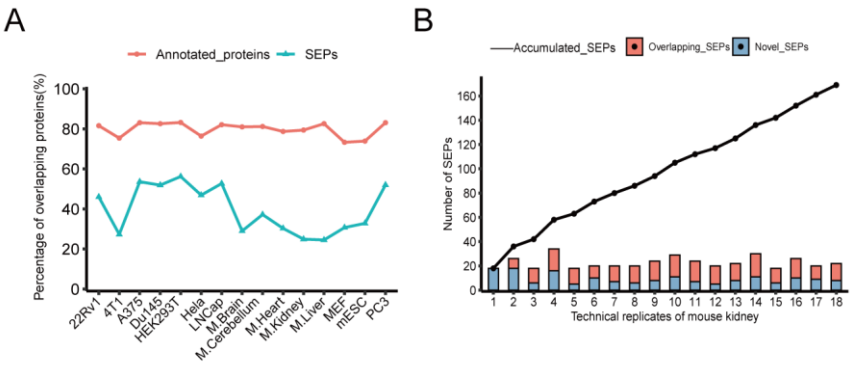


**Figure supplement 3**. Technical replicates improve the identification of lncRNA-encoded SEPs in both cells and tissues. (A) Percentage of overlapping canonical proteins and SEPs identified in three technical replicates within different cell lines and mouse tissues. (B) Number of SEPs identified in each of 18 technical replicates from mouse kidney sample. Each column represents the total number of SEPs identified in a single MS run, and the lower (orange) and upper (yellow) part of the column represent the number of newly identified SEPs in each replicate and overlapping SEPs with previous replicates, respectively. The results showed that an average of 23 SEPs was detected per run with a range between 18 and 34 SEPs in each replicate.

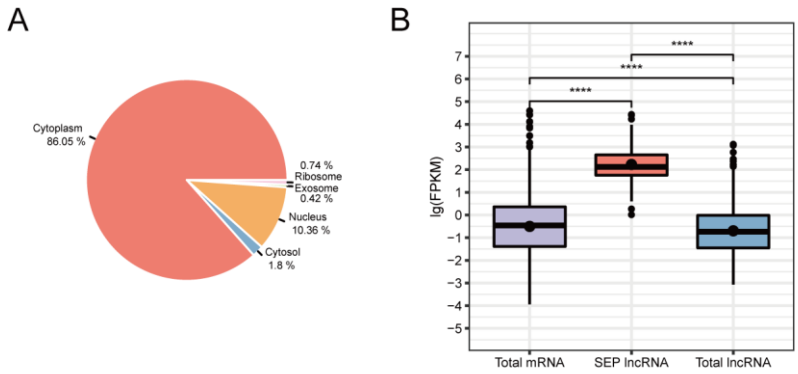


**Figure supplement 4**. Bioinformatic analysis of mouse SEPs-coding lncRNAs. (A) Prediction of the subcellular localization of mouse SEPs-coding lncRNAs. More than 87% of SEPs-coding lncRNAs transcripts were predicted to locate in the cytoplasm, while less than 11% in the nucleus. (B) Comparison of the expression levels of SEPs-coding lncRNAs, whole cell mRNAs and lncRNAs in the kidney tissue of mouse.

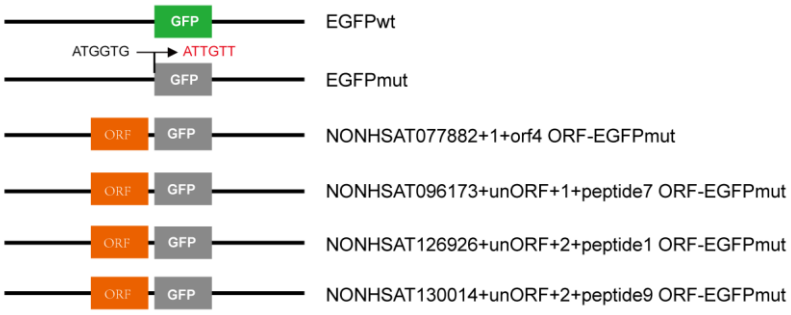


**Figure supplement 6**. GFP and SEP-GFP fusion constructs used in current study.

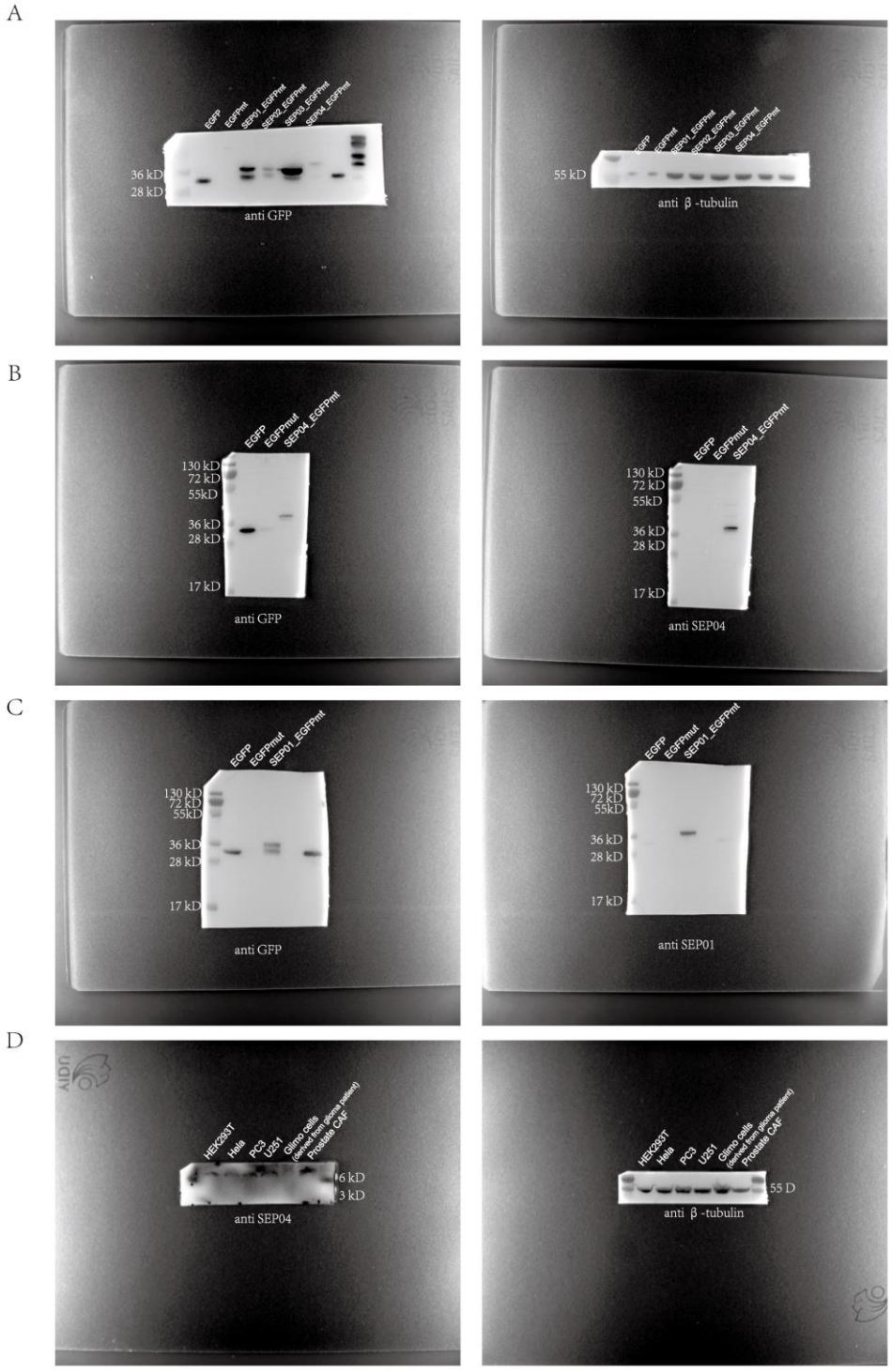


**Figure supplement 7**. Full size western blotting results. (A) Corresponding to figure 4E. (B) Corresponding to figure 4F. (C) Corresponding to figure 4G. (D) Corresponding to figure 4H.

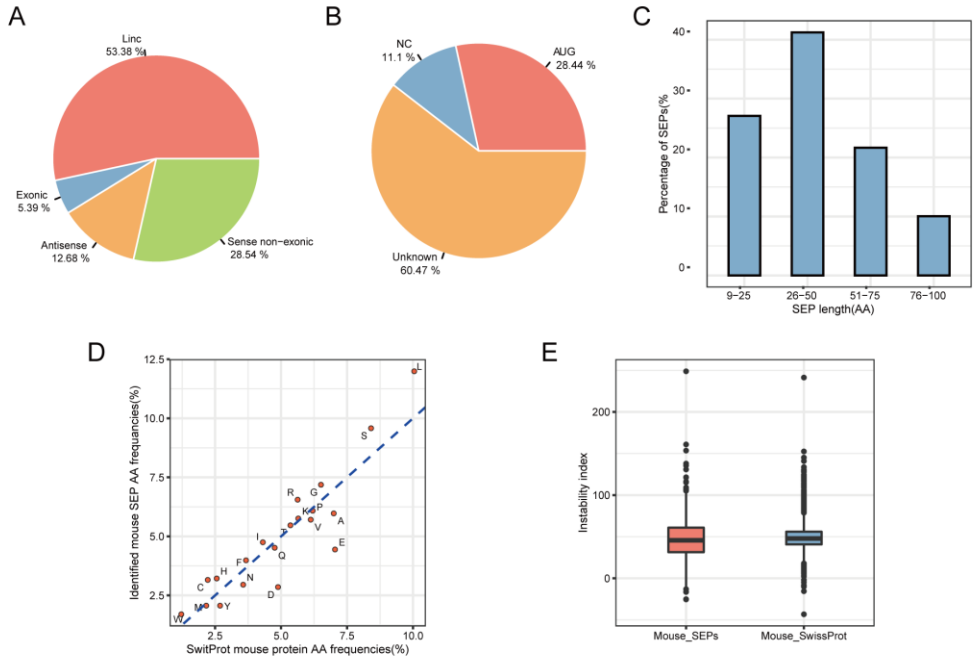


**Figure supplement 8**. Characterization of identified SEPs and their corresponding lncRNA transcripts in mouse. (A) Classification of mouse SEP-coding lncRNA transcripts based on their location on genome with respect to protein-coding genes. (B) Usage of start codon of human SEPs. 72% of the identified lncRNA-SEPs were found to be initiated with non-AUG start codons. AUG, smORF initiates with AUG; NC, near cognate start codon, there is only one base different from AUG; Unknown, smORF neither initiated with AUG nor NC. (C) Length distribution of mouse SEPs. (D) The amino acid usage of canonical proteins and identified mouse SEPs. The mouse SEPs identified in current study tended to utilize more positively charged amino acids (such as R) and less negatively charged amino acids (like D and E), while use a similar amount of uncharged amino acids as canonical proteins. (E) Stability of the total identified mouse SEPs. The instability index calculated by ProtParam was used to characterize the protein stability of mouse SEPs. There is only a minor deviation between the instability index distributions of the identified SEPs and canonical proteins.

**Table supplement 1**. List of the primers used for plasmid construction in current study.

| Primers name | | Sequence (5'-3') |
|---|---|---|
| NONHSAT077882-orf-EcoR1 | forward | CCGGAATTCATGTCTGTTTCCCCTTGGGG |
| NONHSAT077882-orf-BamH1 | reverse | CGCGGATCCGAGTGCCGCTGCCTGCTTAC |
| NONHSAT126926-orf-EcoR1 | forward | CCGGAATTCGCAAGCGGAAAAGACGGGCC |
| NONHSAT126926-orf-BamH1 | reverse | CGCGGATCCGTGGCTGACATTCTTCACCG |
| NONHSAT096173-orf-EcoR1 | forward | CCGGAATTCAGGAGAACATCCCAGCCTAT |
| NONHSAT096173-orf-BamH1 | reverse | CGCGGATCCTTTTTCCAGCCACACAGCCT |
| NONHSAT130014-orf-EcoR1 | forward | CCGGAATTCGCAAGCGGAAAAGACGGGCC |
| NONHSAT130014-orf-BamH1 | reverse | CGCGGATCCGTGGCTGACATTCTTCACCG |
| EGFPmt | forward | CACCGGTCGCCACCATTGTTAGCAAG |
| EGFPmt | reverse | CTTGCTAACAATGGTGGCGACCGGTG |
| NONHSAT077882-orf_EGFPmt | forward | CCGGAATTCATGTCTGTTTCCCCTTGGGG |
| NONHSAT077882-orf_EGFPmt | reverse | CGCGGATCCGAGTGCCGCTGCCTGCTTAC |
| NONHSAT126926-orf_EGFPmt | forward | CCGGAATTCGCAAGCGGAAAAGACGGGCC |
| NONHSAT126926-orf_EGFPmt | reverse | CGCGGATCCGTGGCTGACATTCTTCACCG |
| NONHSAT096173-orf_EGFPmt | forward | CCGGAATTCAGGAGAACATCCCAGCCTAT |
| NONHSAT096173-orf_EGFPmt | reverse | CGCGGATCCTTTTTCCAGCCACACAGCCT |
| NONHSAT130014-orf_EGFPmt | forward | CGGAATTCAAATATACTGCCCAGGTGGA |
| NONHSAT130014-orf_EGFPmt | reverse | CGCGGATCCTACGGGATACTTTTTCTTGT |

**Table supplement 2**. List of the primers used for RT-PCR-based analysis of the transcription of the GFP and SEP-GFP fusion coding sequences.

| Primers name | | Sequence (5'-3') |
|---|---|---|
| EGFP-F | forward | ATGGTGAGCAAGGGCGAGGAGCTGTT |
| EGFP-R | forward | TTACTTGTACAGCTCGTCCATGCCG |
| NONHSAT077882-orf-EGFPmt | forward | ATGTCTGTTTCCCCTTGGGG |
| NONHSAT126926-orf-EGFPmt | forward | GCAAGCGGAAAAGACGGGCC |
| NONHSAT096173-orf-EGFPmt | forward | AGGAGAACATCCCAGCCTAT |
| NONHSAT130014-orf-EGFPmt | forward | AAATATACTGCCCAGGTGGA |

**Table supplementary 5**. List of non-synthetic peptide based PRM validated SEPs. The peptides detected by shotgun proteomics were shown in red, the peptides validated by PRM were shown in green, the overlapping peptides detected in both shotgun and PRM were highlighted in blue.

| SEP ID | Detected peptide by shotgun | Detected peptides by PRM | SEP sequence |
|---|---|---|---|
| NONHSAT009704+unORF+1+peptide4 | KVIFFYPK | VIFFYPK | QAPHFLVSTAKVIFFYPKQIYISIGPQSMWDFQLCKPLHRILNK |
| NONHSAT034233+3+orf2 | GKIEVTEIITDR | IEVTEIITDR | MGKIEVTEIITDRGSGKKRGLHLFILMIITPWIRLLFRNTIL |
| NONHSAT090804+unORF+1+peptide2 | LYTLVISEK<br>YTLVISEK<br>TLVISEK | YLYTLVISEKEK | KDAESVKIKKNKDNVKFKVRCSRYLYTLVISEKEKAEKLKPVPAS |
| NONHSAT096173+unORF+1+peptide7 | KLGEMWNNTAADDKEPYEK<br>LGEMWNNTAADDKEPYEKK | LGEMWNNTAADDKEPYEK | RRTSQPIHCDVVKKLGEMWNNTAADDKEPYEKKAVWLEK |
| NONHSAT119698+unORF+2+peptide4 | AASPAAAAASAAAAAK | LARPSPAPR | ITDQFWCCPRRAASPAAAAASAAAAAKERTLPSRLARPSPAPRQPAPASATECPRRPQMRQGRGSHRPQ |
| NONHSAT123605+1+orf1 | SEAAVDTSSEITTK | MSEAAVDTSSEITTK | MSEAAVDTSSEITTKDLREKKEVVEEAENGREAPANGNAHEENRDRRLTMR |
| NONHSAT126926+unORF+2+peptide1 | TAAAAAAGTITR<br>TAAAAAAGTITRPR | EALILGSQR | ASGKDGPLPPTPERPEALILGSQRTAAAAAAGTITRPRATANCCECTGRGPGSGGGGGSRVAVKNVSH |
| NONHSAT130014+unORF+2+peptide9 | YTAQLDAEDKEDVK | YTAQLDAEDK | KYTAQLDAEDKEDVKSCAEWVSLSKAGIVEYEKQKEKMKNLIPFDQMTIEDLKKTFPETKLDKKKYP |