Title:

Discovery of microRNA-derived RNAs and proteins in human cells

Authors:

Haidong Wu[1,#], Xiaoqiang Huang[2,#,*], Weilong Zhong[3,#], Wenbo Li[4,#], Zhiyong Liu[7,#], Min Zhao[6], Xiaonan Xi[8], Bo Cao[9], Yong Pu[10], Xiangxu Kong[3], Huan Zhao[11], Ronghua Zhang[5], Keguan Lai[1], Xinze Lv[1], Yue Lv[1], Jingyu Bao[1], Meimei Wang[5], Yanan Xiong[5], Lijie Dong[4], Jifeng Zhang[2], Guangling Zhang[5,*], Yuanjing Hu[12,*], Jie Xu[2,*], Y. Eugene Chen[2,*], Shuang Chen[1,*]

[#]These authors contributed equally to this work.

Affiliations:

[1]Tianjin Key Laboratory of Early Druggability Evaluation of Innovative Drugs, Tianjin International Joint Academy of Biomedicine, Tianjin, China

[2]Center for Advanced Models for Translational Sciences and Therapeutics, Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, USA

[3]Department of Gastroenterology and Hepatology, Tianjin Medical University General Hospital

[4]Department of Ophthalmology, Eye Center of Tianjin Medical University

[5]Hebei Provincial Key Laboratory of Medical-Industrial Integration Precision Medicine, School of Clinical Medicine, North China University of Science and Technology

[6]Department of Pathology, Qinhuangdao Municipal No.1 Hospital

[7]Health Science Center, North China University of Science and Technology

[8]College of Pharmacy, Nankai University

[9]Department of Dermatology, Tianjin TEDA Hospital

[10]Department of drug development, Tianjin RuiJiYin Biotechnology Co., Ltd

[11]Department of Analytical Technology, CanSino Biologics Inc.

[12]Gynecologic Oncology Department, Tianjin Central Hospital of Gynecology Obstetrics

***Correspondence**:

xiaoqiah@umich.edu (X.H.), zhanggl@ncst.edu.cn (G.Z.), julianna_hu@163.com (Y.H.), jiex@umich.edu (J.X.), echenum@umich.edu (Y.E.C.), and shuang7332@outlook.com (S.C.)

SUMMARY

MicroRNAs (miRNAs) are a class of short noncoding RNAs that regulate gene expression through the binding of their 5'-end to mRNA. However, the biological effects of miRNA's 3'-end binding to mRNA remain unclear. Here we discover that the pairing of miRNA's 3'-end with RNA could serve as a primer to initiate the production of miRNA-derived RNAs (midRs), in the opposite direction of its originating RNA. midRs could consequently translate into miRNA-derived proteins (midPs). Starting from 2,656 *Homo Sapiens* miRNAs, we predicted 11,453 and 1,239 unique midRs and midPs for humans using a 15-nucleotide-pairing threshold. We verified the bona-fide existence of example midRs and midPs in human cells. Of clinical relevance, we demonstrate that midP0188 is highly expressed in human lung and breast cancer tissues and cells and that midP0188 and its encoding midRs represent novel anti-cancer targets. Our findings propose a miRNA→midR→midP axis that expands the central dogma and reveals thousands of novel RNAs and proteins that have immense potential for playing crucial biological and pathological roles in human cells, as well as other biological systems.

Keywords

microRNA, microRNA-derived mRNAs, microRNA-derived proteins, the central dogma

INTRODUCTION

The central dogma is the dictating principle in molecular biology. It states that the genetic information in the DNA flows to messenger RNA (mRNA) by transcription and then to protein by translation. MicroRNAs (miRNAs) are a class of short noncoding RNAs that regulate gene expression at the post-transcriptional level. miRNAs play fundamental roles in biological processes (*1-6*). It is well established that miRNAs execute their functions in transcriptional regulation through the binding of their 5'-end to the matching sequences on the target mRNAs (*7-9*). A question thus comes out naturally—will the pairing of a miRNA 3'-end with a matching RNA result in any biological effect? To our knowledge, this question has not been systematically investigated. Inspired by the knowledge of RNA-dependent RNA replication in eukaryotic cells and viruses (*10-12*), we speculate that the pairing of a miRNA 3'-end with an RNA may work as a primer to initiate the synthesis of new RNAs (miRNA-derived RNAs, midRs) that are unlikely produced through the canonical transcription process. We further speculate that midRs may encode proteins (miRNA-derived proteins, midPs) or function as noncoding RNAs.

In this work, we developed a bioinformatic workflow to predict midRs and midPs, followed by experimental verification of the bona-fide existence of example midRs and midPs in human cells. We demonstrate that an example midP, midP0188, is highly enriched in human lung and breast cancer tissues and that midP0188 and its encoding midRs represent novel anti-cancer targets. This novel miRNA→midR→midP information

flow expands the central dogma and predicts thousands of previously unknown RNAs and proteins that may have significant biological and pathological roles in human cells as well as in other biological systems. Taken together, our discovery has the potential to significantly advance our understanding of these systems and could pave the way for new discoveries and treatments.

RESULTS

Prediction of midRs and midPs

Fig. 1A illustrates the workflow to predict midRs and midPs. For each human miRNA, its 3'-end is searched against the NCBI RefSeq_RNA database (*13*) to identify RNAs with full reverse complementarity. Initially, we set a minimal matching length of 15 nucleotides (nt) for the search. If a match is identified, the matched miRNA sequence is regarded as the 5'-end of the putative midR, which is generated by extending the miRNA using the matched RNA as the template until the RNA's 5'-end is reached.

Along the extension, we look for putative internal ribosome entry sites (IRES) on the midR that would allow for translation initiation in a cap-independent manner (*14, 15*). We predict IRES by aligning each midR sequence with experimentally validated IRESs and assume that a potential IRES exists if both the coverage and aligned identity are high (see **Methods**). If an IRES is found and if there is a start codon located within 100 nt downstream of the IRES, we will predict a midP using this start codon.

It should be noted that multiple IRESs may be found in one midR and that multiple start codons may be present within the 100 nt proximity to one IRES; all the scenarios are considered for the midP prediction.

Starting from 2,656 human miRNAs from the miRBase database (*16*), we predicted 11,453 and 1,239 unique midRs and midPs, respectively, using a 15-nt matching threshold and after discarding midPs with < 10 amino acids (AAs). The unique midRs and midPs derived from 10- to 15-nt matches were predicted and summarized (**Table S1, Data S1,** and **S2**).

A nomenclature rule was developed to set identifiers for midRs, which includes the name of the miRNA, the GenInfo Identifier (and gene symbol) of the RNA, and the miRNA 3'-end matching position; the identifiers for midPs were similarly defined (see **Methods**). For ease of data presentation, the 1,239 midP proteins produced from the 15-nt match were also named sequentially (midP0001–midP1239) after ranking their sequences in ASCII (American Standard Code for Information Interchange) order. A summary of the midP names, midP identifiers, the identifiers of the encoding midR(s), and the template RNAs can be found in **Data S2**.

miRNAs, midRs, and midPs do not always have a one-to-one correspondence relationship. Multiple IRESs on one midR can give rise to various midPs. On the other hand, different miRNAs bind to the same RNA, producing different midRs but may

encode the same midP. In addition, the same miRNA may bind to the same sequences on different RNAs (i.e., different RNAs or different RNA splicing variants), leading to the different or same resultant midR and midP (**Fig. S1**).

Among the 1,239 midPs, 46 have a high-similarity sequence analog in the NCBI nr (non-redundant) protein database (*13*) with high coverage and aligned identity (see **Methods** and **Data S3**); of them, six midPs have the exact sequence matches in the nr database (**Data S4**). Of note, the 100-AA midP0532 is the same as the selenoprotein P, plasma, 1 encoded by *SEPP1* (a truncated coding sequence of *SELENOP*) through conceptual translation (*17*) (**Data S4**). According to our prediction, midP0532 is encoded by the midR (with identifier *midR.hsa-miR-619-5p.1519242743(CCDC152).2702*) that was generated by the elongation of miRNA *hsa-miR-619-5p* bound to *CCDC152* (*homo sapiens* coiled-coil domain containing 152; a 3,493-nt mRNA) at position 2702; *CCDC152* has a 1,453 bp-long reverse complementarity to *SELENOP* (**Fig. S2**). *CCDC152* was recently reported as an endogenous long non-coding RNA that inhibits selenoprotein P translation (*18*). Thus, the prediction of midP0532 and the reported SEPP1 show the feasibility of our workflow for predicting midRs and midPs.

The remaining 1,193 (96%) midPs do not have a high-similarity sequence analog in the nr database (**Data S5**), representing a brand-new cohort of proteins in human cells.

The potential existence of midPs and midRs in human cells

We next examined the existence of these midPs and midRs in human cells. Initially, SEQUEST (*19*) was used to compare the 1,193 predicted midPs with the experimental mass spectra derived from multiple tumor cell lines. The comparison suggested the potential existence of 941 midPs based on a scoring metric that includes protein sequence coverage, the number of peptide-spectrum matches, and the SEQUEST-HT score (**Data S6**).

We selected four putative midPs to examine the existence of their encoding midRs in A549 lung cancer cells (**Tables S2** and **S3**). For each midR(s), specific reverse transcript (RT) primers complementary to midR(s) and PCR primers across multi-exons were designed to avoid false positive amplification of the transcripts from coding strand genomic DNA (**Fig. 1B** and **Table S4**). Two fragments were amplified for the midP0188-encoding midRs, namely midRs0188.1 and midRs0118.2, as a dual verification (**Fig. 1C, Table S4**). For the other midP-encoding midRs, one fragment was amplified for each. RT-PCR and Sanger sequencing results confirmed the existence of these sequences and suggested the presence of midRs that encode midP0112 and midP0188 (**Fig. 1C** and **Data S7**). The midRs that are predicted to encode midP0045 and midP0323 were not detected (**Fig. 1C**).

We then examined the existence of midP0112 and midP0188. We designed a peptide immunogen to elicit rabbit polyclonal antibodies (pAbs) for each midP (**Table S2**). The anti-midP pAbs were used for immunohistochemistry (IHC) assays to detect the

expression of midP0112 and midP0188 in human cancer and normal (control) tissues. The anti-midP0188 pAb induced strong, positive staining signals in lung and breast cancer tissues (**Fig. 2**, **Fig. S3**), whereas the anti-midP0112 pAb resulted in positive signals in the tissues from lung and colon cancers (**Fig. S4**). These results indicate the two midPs are highly expressed in some human tumor tissues.

Together, these results suggested the existence of midPs and their encoding midRs in human cells.

Bona fide existence of midP0188 in human cells

We sought to further confirm the existence of midP0188, since the anti-midP0188 pAb induced extremely high IHC staining intensity in the lung and breast cancer tissues (**Fig. 2**), from the tissues collected from different patients (**Fig. S3**). To do so, the anti-midP0188 pAb was incubated with A549 cell lysates, followed by midP0188 capture with the protein A/G magnetic beads. SDS-PAGE was run to separate the lysate, and an 18 kDa band that matches the predicted molecular weight of midP0188 was observed (**Fig. 1D**). This band was harvested and digested using trypsin, chymotrypsin, or Glu-C (staphylococcal V8 protease) separately, followed by mass spectrometric analysis. A total of thirteen peptide fragments appearing in midP0188 were confirmed, of which three were from the trypsin digestion, three were from the Glu-C digestion, and the remaining seven were from the chymotrypsin digestion (**Fig. 1E** and **Fig. S5**). The full

mass spectrometric analysis report is shown in **Data S8**. These 13 fragments covered 55% (94/171) AAs of midP0188 (**Fig. 1E**). A search of the four longer consecutive peptide fragments against the NCBI nr protein database returned no hits of identical or high-similarity (**Table S5**), suggesting that these peptides are previously unknown biological entities.

We next tested if downregulating the midP0188-deriving miRNAs would indeed knock down this protein. midP0188 was predicted by using a stringent threshold of the 15-nt match. Specifically, *hsa-miR-619-5p* is the miRNA that was predicted to bind mRNA *AP1M1* variant 1 or 2, which both give rise to midP0188-encoding midRs (**Table S3**, **Fig. S6**). We note however if we relax the matching threshold (e.g. from 15-nt to 14-nt or fewer), there may be additional miRNAs that could derive midP0188-encoding midRs (**Fig. S6**). Therefore, to knock down midP0188, we should consider targeting most if not all potential originating miRNAs.

We relaxed the matching length to 12 nt and identified four miRNAs that may eventually give rise to midP0188, i.e., *hsa-miR-619-5p, hsa-miR-3135b*, *hsa-miR-4728-3p, and hsa-miR-5585-3p*, all making a ≥ 12-nt match with *AP1M1* (**Fig. S6**). We previously reported the development of circular single-stranded DNA (CSSD) as an efficient multiplex knockdown method by adsorbing multiple miRNAs (*20*). We thus designed a CSSD, namely CSSD4, to knock down the four miRNAs simultaneously (**Fig. 1F**, **Fig. S7**). CSSD that contains a scrambled sequence was used as a control. Transfection of

CSSD4 to A549 cells effectively reduces the levels of all four miRNAs (**Fig. 1G**), compared with the control. Importantly, the levels of midP0188 in the CSSD4-treated MDA-MB-231 and A549 cells are markedly reduced (**Fig. 1H**).

We then asked if targeting the relevant midRs would also similarly reduce the miP0188 protein levels. To achieve this, we generated stable cell lines expressing short hairpin RNA (shRNA) to target the midP0188-encoding midRs (**Fig. S8A**). We designed three interfering lentiviruses targeting the midP0188-encoding midRs and a control lentivirus containing a scrambled sequence. Western blot showed that shRNA-2 had the highest interfering efficacy (**Fig. S8A**). As expected, a remarkable reduction of the midP0188 levels was observed in shRNA-2-treated MDA-MB-231 and A549 cells (**Fig. S8B**).

Together, our work established the bona fide existence of midP0188 in human cells, and that midP0188 is produced through the miRNA→midR→midP axis.

Targeting midP0188 in lung and breast cancer cells *in vitro*

We speculate that some midPs and midRs could serve as biomarkers and/or therapeutic targets for human diseases such as cancer. As a proof of concept, we worked to explore if midP0188, its originating miRNAs, and its encoding midRs represent targets for cancer drug development.

In the transwell and wound healing assays, we demonstrated that CSSD4-treated cells had compromised invasion (**Fig. 3A**) and migration (**Fig. 3B**) capacities, compared to

tumor cells that received the control CSSD. Scanning electron microscopy imaging revealed that CSSD4 treatment adversely altered the morphology of these tumor cells (**Fig. 3C**). Similar results were observed when we used shRNA-expressing inference lentivirus to target the midP0188-encoding midRs in MDA-MB-231 and A549 cells (**Fig. S8C-E**). Thus, targeting the midP0188-originating miRNAs or the midP0188-encoding midRs compromised tumor cells' invasion and migration *in vitro*.

Conversely, when we overexpressed midP0188 in A549 cells by transfecting the PiggyBac Dual promoter plasmid with the midP0188-encoding sequence (**Fig. S9A**), transwell and wound healing assays indicated that such overexpression promoted cell migration (**Fig. S9B**) and invasion (**Fig. S9C**), compared to the control that transfected with vehicle vector. Scanning electron microscopy imaging also indicated a remarkable morphology change after midP0188 overexpression in A549 cells (**Fig. S9D**).

Targeting midP0188 in lung and breast cancer cells *in vivo*

We proceeded to evaluate the potential benefits of targeting midP0188 in animal models. Specifically, to examine the effects of targeting midP0188 on (i) solid tumor growth and (ii) metastasis *in vivo*, we established (i) the subcutaneous solid tumor xenograft model and (ii) the tail vein injection metastasis mouse models. The A549 and MDA-MB-231 cells were used to represent lung and breast cancer cells, respectively.

In the subcutaneous solid tumor model, 24 five-week-old BALB/c nude mice were

randomly divided into four groups. A549 or MDA-MB-231 cells were inoculated subcutaneously at a concentration of 5×10^6 cells/mouse. When the tumor size reached about 4 mm in diameter, CSSD4 or Ctrl-CSSD was injected intratumorally into the tumor mass every seven days for a total of three times. Tumor sizes were measured every three days (**Fig. 3D**). After three administrations of CSSD4 or Ctrl-CSSD, the mice were euthanized, and solid tumor tissues were collected for size measurement and IHC analysis. The results showed that tumor sizes in the CSSD4 groups were significantly smaller than those in Ctrl-CSSD groups for both tumor cell types (**Fig. 3E**). At the same time, IHC staining revealed that the protein levels of midP0188 were effectively reduced in the CSSD4 treated group (**Fig. 3F**).

In the tail vein metastasis model, 12 five-week-old BALB/c nude mice were randomly divided into four groups. MDA-MB-231 and A549 cells that stably express luciferase were established and injected into nude mice through the tail vein at a concentration of 5×10^6 cells/mouse. Ten days later, CSSD4 or Ctrl-CSSD (30 μg/mouse) was administered intravenously. The metastasis of tumor cells in nude mice was evaluated after three administrations (day 10, day 17, and day 24). The results indicated that the CSSD4-treated animals suffered much less lung metastasis than those in the control group for both tumor cell types (**Fig. 3G**).

These results showed that targeting miRNAs to knockdown midP0188 inhibits tumor cell growth and metastasis *in vivo*.

As a proof of concept, these *in vitro* and *in vivo* experiment results reveal that midP0188 plays an important role in promoting lung and breast tumor genesis and metastasis and that downregulating midP0188 through CSSD4 represents a novel strategy to treat lung and breast cancers.

DISCUSSION

The central dogma dictates the DNA→mRNA→protein rule as the basic principle in biology. More and more evidence shows that translation is pervasive, occurring in not only canonical coding regions, but also canonical untranslated regions, introns, and noncoding RNAs (*21-23*). Besides their roles in regulating gene expression at the post-transcriptional level, pri-miRNAs containing open reading frames were also found to encode polypeptides (*24-26*). In plants, miRNA-encoded polypeptides are even more ubiquitous (*27, 28*). Nevertheless, most existing studies are focused on the post-transcriptional regulation of miRNA, which is based on the reverse pairing between miRNA 5'-end and target mRNA. In this study, we have discovered that the reverse pairing between a miRNA's 3'-end with an RNA can give rise to new biological molecules. In this process, miRNA and the paired RNA act like "primer" and "template", leading to the production of midRs. We demonstrate that midRs can consequently lead to the production of midPs. This discovery will serve as an important addition to the central dogma, enhancing our understanding of fundamental biological processes.

Even under the stringent 15-nt threshold, more than 1,000 new midPs have been predicted. The number of midRs is even more enormous. Relaxing the matching threshold to 12-nt will increase the possible number of midPs to about 4000, and midRs to > 2 million. These novel biological molecules add new complexity to the biological system and represent brand-new research entities.

In our pilot proof of concept study, we have confirmed the presence and role of midP0188 in cancer progression and demonstrated that its downregulation can reduce the tumor genesis and metastasis of lung and breast cancers. While we were not able to experimentally verify all predicted midRs and midPs or study their biology due to practical constraints, we believe that these molecules have the potential to play important physiological and pathophysiological roles. This exciting discovery warrants further investigation in future research.

The molecular basis for the synthesis of midRs in human cells is to be determined. It likely utilizes enzymes that have similar functions to RNA-dependent RNA polymerases (RdRp), which are primarily found in prokaryotes (*11, 29, 30*). Though, proteins with demonstrated or putative RdRp functions have also been reported in eukaryotes (*10, 31-33*). Our discovery of midRs and midPs strongly argues for the presence of a yet uncharacterized endogenous biological pathway in human (more broadly, eukaryotic) cells that can copy RNA, which is also supported by a previous study (*34*). We reason that our discovery could provide indirect yet reasonable evidence of the existence of

RdRp-like proteins in human cells. However, the identity of such proteins is to be further investigated. We also ask a few intriguing questions. When a host cell is infected by RNA viruses, will the viral RdRp be employed to derive midRs from the host cell's miRNAs paired with the viral RNA, and/or will midRs be derived from the pairing of the viral miRNAs and host RNAs? If so, what are their biological functions on host-pathogen interaction? Further studies are required to answer these questions.

Our bioinformatic pipeline works for the prediction of midRs and midPs, yet it has some limitations. For example, in our workflow, IRES was required to translate midRs, but it remains unclear if IRES is a must to initiate translation. Besides, we predicted IRES by searching IRES-like sequences in midRs through sequence alignment, which may not be rigorous. In addition, it is purely empirical to require start codons located within 100 nt downstream of the IRES. With these, we speculate that a set of midRs and midPs may not be predicted by our current bioinformatic pipeline and related algorithms, which need further improvement in follow-up studies.

In sum, we report the discovery of miRNA-derived RNAs and proteins in human cells. Our work proposes a miRNA→midR→midP axis that expands the central dogma and predicts thousands of previously unknown RNAs and proteins that may have significant biological and pathological roles in human cells as well as in other biological systems.

MATERIALS AND METHODS

Bioinformatic prediction of midRs and midPs

2,656 human miRNAs were obtained from the miRbase website (https://www.mirbase.org/). The 3'-end (e.g. 15-nt fragment) of each miRNA (U→T) was searched against the NCBI RefSeq_RNA database (*13*) using BLASTN (*35*), to identify paired RNA templates to generate midRs:

blastn -task blastn-short -db db_refseq_rna -query query_15nt.fa -out out_15nt.xml -outfmt 5 -evalue 1 -gapopen 5 -gapextend 2 -reward 1 -penalty -3 -ungapped -max_target_seqs 10000 -strand minus -taxids 9606

Since the queries are short, the option "*-task blastn-short*" was used. Default parameters "*-gapopen 5 -gapextend 2 -reward 1 -penalty -3*" were applied for "*blastn-short*" tasks. The option "*-ungapped*" ensures no gap in alignments. The option "*-strand minus*" restricts the search to reverse matches. The option "*-taxids 9606*" restricts the search to human RNAs. For 15-nt queries, the calculated e-value for an exact, ungapped 15-nt match is 0.941241, and hence a slightly higher e-value cutoff "*-evalue 1*" was used. Similarly, the e-value cutoffs were set to 5, 20, 100, 500, and 1,000 for 14-, 13-, 12-, 11-, and 10-nt queries, respectively. With these parameters, a sufficiently high cutoff ("*-max_target_seqs 10000*") was set to output RNA hits; this threshold was not reached for the 10- to 15-nt queries. For X-nt queries, the searched RNA hits were further filtered to

retain those with exact X-nt matches. For the sake of description, below we use the 15-nt match as an example. The information for a hit such as the GenInfo Identifier (GI) number, NCBI Reference Sequence number, start and end matching positions, and gene symbol was recorded.

The RNA sequence of a hit was retrieved using its GI:

blastdbcmd -entry gi -db db_refseq_rna

The midR sequence was generated by reverse transcribing the hit RNA from the end matching position till its 5'-end. The number of midRs derived by each human miRNA varies substantially. For instance, some miRNAs (e.g. *hsa-miR-99b-5p*, *hsa-miR-9985*, *hsa-miR-9903*, etc) do not yield any midR while *has-miR-619-5p* alone derives 3,422 midRs. A total of 13,310 midRs were derived from all 2,656 human miRNAs (**Table S1**). Among them, 11,453 unique midRs were used to create a database:

mkblastdb -in db_RNA_uniq.fa -out db_RNA_uniq -dbtype nucl

The unique midRs were translated into proteins following two criteria: (1) a midR must have a predicted IRES (*14*) and (2) a start codon is available within 100 nt downstream of the IRES. The first criterion was checked by comparing midRs with the experimental IRES sequences taken from the IRESbase database (*15*), which contains 554 viral IRESs, 691 human IRESs, and 83 IRESs from other eukaryotes. Because the length of IRESs varies substantially, they were searched against the midR database with different

BLASTN parameters.

if an IRES has ≤ 15 nt:

blastn -db db_ uniq_midr -query ires_le15.fa -out ires_le15.xml -outfmt 5 -evalue 1e5 -task blastn-short -max_target_seqs 2500000 -reward 1 -penalty -3 -gapopen 5 -gapextend 2 -ungapped -strand plus

else if an IRES has < 50 nt:

blastn -db db_ uniq_midr -query ires_lt50.fa -out ires_lt50.xml -outfmt 5 -evalue 1e5 -task blastn-short -max_target_seqs 2500000 -reward 1 -penalty -2 -gapopen 0 -gapextend 2 -strand plus

otherwise:

blastn -db db_ uniq_midr -query ires_ge50.fa -out ires_ge50.xml -outfmt 5 -evalue 1e5 -task blastn -max_target_seqs 2500000 -reward 2 -penalty -3 -gapopen 5 -gapextend 2 -strand plus

The option "*-strand plus*" restricts the search to forward matches. For very short IRES queries (≤ 15 nt), a high gap opening penalty "*-gapopen 5*" and "*-ungapped*" were used to disallow gaps; "*-penalty -3*" was used to disfavor mismatches. For IRES queries in (15, 50) nt, the penalty scores were reduced to allow more mismatches and gap openings. For > 50-nt IRES queries, default parameters for a conventional "*blastn*" task were used.

In all cases, high cutoffs for e-value ("*-evalue 1e5*") and the number of targets ("*-max_target_seqs 2500000*", larger than the number of midRs) were used.

An IRES site was predicted to be existing if coverage ≥ 1 and aligned identity ≥ 1 for ≤ 15-nt IRES queries, or coverage ≥ 0.7 and aligned identity ≥ 0.75 for > 15-nt IRES queries. Here, the coverage and aligned identity were defined as $coverage = \frac{aligned\ length}{query\ length}$ and $aigned\ identity = \frac{identity}{aligned\ length}$, and the aligned length, query length, and identity were extracted from the BLASTN output.

The midRs were then translated following the second criterion. A midR may contain multiple IRESs and following an IRES may exist multiple start codons. All possibilities were considered when translating a midR into midP(s). A total of 5,369 midPs were obtained with > 10 AAs, of which 1,239 are unique. We searched the midP sequences against the NCBI nr protein database using BLASTP (*35*) to check if they have an existing, high-similarity human protein analog:

if a midP has ≤ 15 AAs:

blastp -db nr -query midp_le15.fa -out midp_le15.xml -outfmt 5 -evalue 1 -task blastp-short -taxids 9606

else if a midP has < 30 AAs:

blastp -db nr -query midp_lt30.fa -out midp_lt30.xml -outfmt 5 -evalue 1e-4 -task blastp-

short -taxids 9606

otherwise:

blastp -db nr -query midp_ge30.fa -out midp_ge30.xml -outfmt 5 -evalue 1e-4 -task blastp -taxids 9606

Here, “high-similarity” was defined as coverage ≥ 1 and aligned identity ≥ 0.9 for midPs with ≤ 15 AAs, or coverage ≥ 0.9 and aligned identity ≥ 0.9 otherwise (coverage and aligned identity are parameters defined as above).

Similar procedures were performed for 10- to 14-nt matches, and the numbers of RNA hits, midRs, and midPs are summarized in **Table S1**.

Nomenclature of midR and midP identifiers and names

We developed a nomenclature to set identifiers for midRs and midPs that can reflect the nature that midRs and midPs are derived from the elongation of miRNA with RNA as a template. The midR identifier is defined as “midR.{miRNA name}.{GenInfo Identifier number of the RNA (gene symbol of the RNA)}.{the end matching position on RNA}”, in which “midR” is the prefix. In principle, we can define midP identifiers similarly by appending {start codon position on midR} to the midR identifier. However, we noticed that a midP can be completely defined by template RNA together with the position on RNA that corresponds to the start-codon position on midR. Thus, a midP identifier is

defined as “midP.{GenInfo Identifier number of the RNA (gene symbol of the RNA)}.{the midR start codon-paired position on RNA}”. The midP identifiers generated in this way largely prevent using multiple midR-based midP identifiers for the same midP encoded by midRs derived from different miRNAs bound to one RNA.

Though meaningful, it should be noted that the above-defined midR and midP identifier may still not be unique for the same midR or midP as a result of RNA variants caused by RNA splicing. Since the identifiers are verbose for data presentation, sequential names were also defined and used to describe specific midRs and midPs (**Tables S2** and **S3**, **Data S2**).

Cell line and transfection

Lung cancer cell line A549 and breast cancer cell line MDA-MB-231 were purchased from the Cell Resource Center of the Chinese Academy of Medical and authenticated by short tandem repeat analysis. All cell lines tested negative for mycoplasma. A549 cells were cultured in minimum Eagle’s medium with 10% heat-inactivated fetal calf serum (Gibco, USA) and penicillin-streptomycin (Beyotime, China). MDA-MB-231 cells were cultured in Dulbecco’s modified Eagle’s medium with 10% heat-inactivated fetal calf serum (Gibco, USA) and penicillin-streptomycin (Beyotime, China). Cells were maintained at 37 °C with 5% CO_2.

For midP0188 knockdown, shRNA lentivirus was synthesized by Obio Technology

(Shanghai, China). RNA interference oligonucleotides were annealed and inserted into the pSLenti-U6- shRNA-CMV-EGFP-F2A-Puro-WPRE lentiviral vector. The shRNA sequences are: shRNA-1, GGAGTCATCTCCACTGCATTT; shRNA-2, GCGTCTACTTCCTTCGTCTTG; shRNA-3, GCTAGGTGCGTCTGCTTCATT; ctrl-shRNA, CCTAAGGTTAAGTCGCCCTCG. For midP0188 overexpression, the midP0188 encoding sequence was cloned into PiggyBac Dual promoter to generate midP0188-PiggyBac Dual promoter (GFP-Puro) Plasmid by GENEWIZ (Suzhou). According to the instructions of Neofect (Neofect Biotech Co., Ltd), midP0188-PiggyBac Dual promoter (GFP-Puro) Plasmid and PiggyBac Transposase(GFP-Puro) was transfected into A549 cells. After 48h, puromycin (1 μg/ml) was added to the culture for positive clone selection in the next 7 days. The midP0188-overexpressed A549 cell lines were confirmed by Western blot.

Transwell assay

Transwell chambers were placed into a 24-well plate and 300 μl pre-warmed serum-free medium was added to the upper chamber. The culture was discarded after 15 min at room temperature. The treated cells were trypsinized into single cells and resuspended in a serum-free medium. After counting, the cells were seeded into the upper chamber of 10^5 cells/well. 500 μl 10% fetal bovine serum medium was added to the lower chamber. All cells were cultured in an incubator with CO_2 at 37 °C for 24 h. After wiping cells in the upper chamber, 4% formaldehyde was used to fix cells for 30 min. After washing with

phosphate-buffered saline (PBS) and staining with 0.1% crystal violet, invaded cells were observed under a microscope (200× magnification).

Wound healing assay

Cells were digested and seeded in 24-well plates at a density of 2×10^5 cells/well. After cell density reached 90%, we used a 100 μl pipette tip to make straight scratches. The culture medium was discarded, and the cells were washed three times with PBS to remove cell debris and cultured with a serum-free medium in an incubator with 5% CO_2 at 37°C. After 48 h, photos were taken and the cell migration rates were calculated.

Scanning electron microscope imaging

Cells were first fixed in pre-chilled 2.5% glutaraldehyde at 4 °C for 2 h and then dehydrated with a gradually increased concentration of ethanol (30%, 50%, 70%, 80%, 90%, 95%, and 100% v/v) for 10 min at each concentration. Following this, cells were treated with a gradually increased concentration of tert-butanol (30%, 50%, 70%, 80%, 90%, 95%, and 100% v/v) for 15 min at each concentration. Finally, cell phenotypes were observed under a scanning electron microscope (JEOL, Japan) after dried and gold spray.

RT-PCR

Total cell RNAs were extracted using trizol (Beyotime, China) and quantified. 3 μg RNA

was used to carry out reverse transcription (RT) reaction to obtain cDNA following the instruction of PrimeScript RT Reagent Kit (Takara, Japan). RT primers are specific to midRs. Next, we designed primers that span at least two exons to run PCR with the cDNA as a template. The PCR products were identified by agarose gel electrophoresis and DNA sequencing. All primer sequences are provided in **Table S4**.

Western blot

Cell proteins were extracted using radioimmunoprecipitation assay lysis buffer that contains protease inhibitors (Beyotime, China). After quantification using BCA Protein Assay Kit (Beyotime, China), proteins were separated by 10% SDS-PAGE and transferred to the PVDF membrane (Millipore, USA). The membrane was placed in 5% bovine serum albumin and incubated with the anti-midP0188 pAb (1:500 dilution) or the anti-actin antibody (1:2000, Affinity, USA) for 2 h at room temperature, and then incubated with horseradish peroxidase-labeled secondary antibody (1:3000, Affinity, USA) at room temperature for 1 h. After washing the membrane with phosphate-buffered saline with Tween20 three times, we used the ECL Chemiluminescence Kit (Beyotime, China) to detect proteins.

Silver staining

After the cells were washed twice with 1× PBS, proteins were extracted using 500 μl cell lysate (P1048, Beyotime, China). 50 μl protein A/G magnetic beads (HY-K0202, MCE,

USA) were incubated with the anti-midP0188 pAb at room temperature for 30 min, and then the harvested beads were incubated with protein at 4 °C for 2 h. The protein complex eluted using 1× SDS-PAGE loading buffer was then separated by SDS-PAGE electrophoresis. The silver staining was performed according to instructions (P0017S, Beyotime, China). The gel was fixed with 100 ml fixative solution at room temperature for 20 min, then rinsed with 30% ethanol and pure water for 10 min. The silver solution was then treated for 10 min at room temperature and rinsed with pure water. Then, the reaction was terminated after being treated in the color-developing solution for 10 min.

Mass spectrometry

Acuity UPLC (Thermo Scientific) connected to a QExactive orbitrap (Thermo Scientific) was used for protein identification using the SDS-PAGE bands by peptide mapping. The bands of interest were cut out of the gel, de-stained, and subjected to in-situ trypsin digestion for 16 h at 37 °C. The digested peptides were extracted by solution (acetonitrile–formic acid–water 7:1:2, v/v/v) and directly dried by vacuum concentrator. The peptides were separated using a UPLC system (Thermo Scientific) equipped with a reversed-phase CSH C18 column (150 mm, 1.7 μm, Waters) and analyzed using a QExactive orbitrap (Thermo Scientific) run in high-sensitivity mode. The flow rate was 200 μl/min and the column temperature was kept at 40 °C. Solvent A consisted of 0.1% formic acid (FA) in water. Solvent B consisted of 0.1% FA in acetonitrile. During the 120-min gradient, the solvent B level was increased from 2% to 29% in 60 min, from 29% to

55 % in 5 mins, from 55% to 75% in 2 min, and remained at 75% for 5 min, and then returned to 2% over 2 min and stayed at 2% for the remaining 6 min. The masses of the digested peptides were then entered into the NCBI nr database (*Homo sapiens*) using the Sequest search engine (Thermo) to identify the proteins. The digestion products by Glu-C and chymotrypsin were similarly analyzed.

CSSD synthesis

We designed CSSD4 that can absorb four miRNAs and a control CSSD (Ctrl-CSSD) with scramble DNA. For Ctrl-CSSD, the single-stranded DNA (5'-AATTCAAAGAATTAACCTTAATTGAA GGGGAGGGTTCAGTACTTTTGTGTAGTACAAATATCAGTACTTTTGTGTAGTACAAAA GGGAGGGCTTCAATTAAGGTTAATTCTTTG-3') was treated with T4 DNA ligase (Takara, Beijing, China) after annealing. For CSSD4, two single-stranded DNAs were synthesized: one contains *miR-619-5p* and *miR-3135b* binding sequences (5'-GAATTCAAAGAATTAACCTTAATTGAAGGGGAGGGTTCACCACTGCACTCGCTCCAG CCATATGGCTCATGCCTGTAATCCCAGCAAGGGAGGGCTTCAATTAAGGTTAATTCTT TG -3'); the other contains a *miR-4728-3p* and *miR-5585-3p* binding sequence (5'-AATTCAAAGAATTAACCTTAATTGAAGGGGAGGGTTCTGGGGCAGGAGGGAGGTCA GCATGATATACCTGTAGTCCCAGCTATTCAGAAGGGAGGGCTTCAATTAAGGTTAATT CTTT-3'). After annealing, the two DNA strands were treated with T4 DNA ligase to form a completely closed circular DNA (**Fig. S7**). All single-stranded DNAs were synthesized

by Genewiz (Beijing, China).

Polyclonal Antibody (pAb)

For chosen midPs, specific peptide immunogens were designed and synthesized (Apeptide, Shanghai, China). After mass spectrometric verification, the synthetic peptides were cross-linked to the Keyhole limpet hemocyanin protein to prepare immunogenic antigens. For the first immunization, 350 μg antigen and 1000 μl Freund's complete adjuvant (Sigma, F5881) were mixed and injected into the rabbits' skin. For 3 follow-up immunizations, 200 μg antigen mixed with 1000 μl Freund's incomplete adjuvant (Sigma, F5506) was injected each time. After the rabbits were euthanized, blood was collected and centrifuged to obtain serum for subsequent immunoaffinity purification. After cross-linking peptides and epoxy-modified agarose (Wechsler Chromatography, Beijing), the serum was combined with peptides and eluted to obtain affinity-purified specific antibodies, which were then used for western blot and immunohistochemical staining assays. All immunogenic peptide sequences are provided in **Table S3**.

Clinical specimens

Paraffin-embedded tumor samples from Qinhuangdao Municipal No. 1 Hospital were reviewed by a pathologist. The study was performed by the guidelines from the Declaration of Helsinki and its amendments or comparable ethical standards.

Immunohistochemistry (IHC) assay

Tumor tissues were fixed with 10% formalin and embedded in paraffin. For IHC, the embedded tissues were cut into 4 μm-thick sections. The sodium citrate solution (Beyotime, China) was used for antigen retrieval, followed by blocking with 3% H_2O_2. Then the sections were incubated with the anti-midP0188 pAb (1:200) overnight at 4 °C in the dark. After washing with PBS and incubated with horseradish peroxidase-conjugated secondary antibody (1:300, Affinity, USA), the DAB kit (Beyotime, China) and hematoxylin were used for staining development and counterstaining, respectively.

Xenograft tumor in nude mice

BALB/C nude mice were used for evaluating tumor growth and metastasis. For the tumor growth experiments, 24 mice were randomly divided into four groups. Two groups were subcutaneously inoculated with A549 cells at a concentration of 5×10^6 cells/mouse and were then treated with CSSD4 or Ctrl-CSSD through intratumoral injection at 20 μg/mouse once every seven days for a total of three times. The same procedures were conducted on the other two groups with MDA-MB-231 cells used. All mice were euthanized on day 34 by CO_2 asphyxiation. For the tumor metastasis experiments, 12 mice were randomly divided into four groups. Two groups of mice were inoculated with A549 cells into the tail veins at 5×10^6 cells/mouse, and ten days later CSSD4 or Ctrl-CSSD was injected into the tail veins for treatment at 30 μg/mouse once every seven

days three times. All animal experiments were permitted and carried out according to the practices prescribed by the Institutional Animal Care and Use Committee (IACUC) of the Tianjin International Joint Academy of Biomedicine.

Statistical analysis

Statistical analyses were performed using GraphPad Prism software 9.0 (La Jolla, CA, USA), and the results were expressed as the mean ± standard deviation. Comparisons between the two groups were performed using the unpaired, two-tailed student's *t*-test. Comparisons between multiple groups were performed using one-way ANOVA and Tukey's posthoc test. Significance is shown as: **, $P < 0.01$; *, $P < 0.05$.

Data and code availability

The mature miRNA sequences were downloaded at https://www.mirbase.org/ftp.shtml (accessed on December 20, 2020). The experimentally determined IRES sequences were downloaded at http://reprod.njmu.edu.cn/cgi-bin/iresbase/index.php (accessed on January 2, 2021). The RefSeq_RNA database was downloaded at https://ftp.ncbi.nlm.nih.gov/blast/db (accessed on December 20, 2020). The NCBI nr protein database was downloaded at https://ftp.ncbi.nlm.nih.gov/blast/db (accessed on January 2, 2021). The BLAST suite (version 2.11.0) was downloaded at https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+. In case these databases or software have been updated, the versions used in this study can be provided upon request. All

Perl scripts for the prediction and analysis of midRs and midPs, Data S1–S8, and other related source data are publicly available at https://doi.org/10.5281/zenodo.7826531.

ACKNOWLEDGMENTS

This work was financially supported by the National Natural Science Fundation of China (Grant No. 81972746 to S.C.), the Government Funded Clinical Medicine Talents Training Project of Hebei Province (Grant No. [2020] 397 to G.Z.), the Natural Science Foundation of Hebei Province (Grant No. H2021209026 to G.Z.), and the Returned Overseas Students of Hebei Province (Grant No. C20210340 to G.Z.). The computational research in this publication was supported in part by computational resources and services provided by Advanced Research Computing at the University of Michigan.

AUTHOR CONTRIBUTIONS

H.W., X.H., J.X., Y.E.C., and S.C. conceived the project. Y.E.C. and S.C. supervised and administrated the project. X.H. developed the bioinformatic workflow to predict midRs and midPs and analyzed computational data. H.W., W.L., M.Z., X.K., R.Z., K.L., X.L., Y.L., M.W., L.D., and S.C. performed the cellular experiment. X.X. and H.Z. performed the mass spectrometry experiment. Z.L., W.Z., B.C., and S.C. performed the animal experiment. W.Z. performed the IHC experiment. H.W., W.Z., W.L., R.Z., G.Z. and S.C. performed the molecular experiment. H.W., X.H., W.Z., Z.L., Y.H., Y.P.,and S.C.

performed formal analysis. M.Z., M.W., G.Z., Y.H., J.X., Y.E.C., and S.C. provided resources. H.W., X.H., W.Z., W.L., Z.L., Y.H., and S.C. wrote the initial manuscript. J.X. and Y.E.C. edited the manuscript. All authors discussed the results and contributed to the final manuscript.

COMPETING INTERESTS

H.W., X.H., J.X., Y.E.C., and S.C. are co-inventors of a patent application covering the predicted midRs and midPs described in this work. The remaining authors declare no competing interests.

REFERENCES

1. T. X. Lu, M. E. Rothenberg, MicroRNA. *J Allergy Clin Immunol* **141**, 1202-1207 (2018).
2. M. Chalfie, H. R. Horvitz, J. E. Sulston, Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell* **24**, 59-69 (1981).
3. V. Ambros, H. R. Horvitz, Heterochronic mutants of the nematode Caenorhabditis elegans. *Science* **226**, 409-416 (1984).
4. G. P. Brennan, D. C. Henshall, MicroRNAs as regulators of brain function and targets for treatment of epilepsy. *Nat Rev Neurol* **16**, 506-519 (2020).
5. T. Kim, C. M. Croce, MicroRNA and ER stress in cancer. *Semin Cancer Biol* **75**, 3-14 (2021).
6. T. J. Lee *et al.*, Strategies to Modulate MicroRNA Functions for the Treatment of Cancer or Organ Injury. *Pharmacol Rev* **72**, 639-667 (2020).
7. Y. Tay, J. Zhang, A. M. Thomson, B. Lim, I. Rigoutsos, MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature* **455**, 1124-1128 (2008).
8. S. Vasudevan, Y. Tong, J. A. Steitz, Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318**, 1931-1934 (2007).
9. U. A. Orom, F. C. Nielsen, A. H. Lund, MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Mol Cell* **30**, 460-471 (2008).

10. Y. Maida *et al.*, An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA. *Nature* **461**, 230-235 (2009).
11. T. Krischuns, M. Lukarska, N. Naffakh, S. Cusack, Influenza Virus RNA-Dependent RNA Polymerase and the Host Transcriptional Apparatus. *Annu Rev Biochem* **90**, 321-348 (2021).
12. Y. Wang, V. Anirudhan, R. Du, Q. Cui, L. Rong, RNA-dependent RNA polymerase of SARS-CoV-2 as a therapeutic target. *J Med Virol* **93**, 300-310 (2021).
13. N. A. O'Leary *et al.*, Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).
14. P. Weber *et al.*, Therapy-related transcriptional subtypes in matched primary and recurrent head and neck cancer. *Clin Cancer Res*, (2021).
15. J. Zhao *et al.*, IRESbase: A Comprehensive Database of Experimentally Validated Internal Ribosome Entry Sites. *Genomics Proteomics Bioinformatics* **18**, 129-139 (2020).
16. A. Kozomara, M. Birgaoanu, S. Griffiths-Jones, miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155-D162 (2019).
17. J. C. Venter *et al.*, The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
18. Y. Mita *et al.*, Identification of a novel endogenous long non-coding RNA that inhibits selenoprotein P translation. *Nucleic Acids Res.* **49**, 6893-6907 (2021).
19. J. K. Eng, A. L. McCormack, J. R. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976-989 (1994).
20. J. Meng *et al.*, Derepression of co-silenced tumor suppressor genes by nanoparticle-loaded circular ssDNA reduces tumor malignancy. *Sci Transl Med* **10**, (2018).
21. J. S. Kesner *et al.*, Noncoding translation mitigation. *Nature*, (2023).
22. Z. Ji, R. Song, A. Regev, K. Struhl, Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* **4**, e08890 (2015).
23. N. T. Ingolia *et al.*, Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**, 1365-1379 (2014).
24. D. Lauressergues *et al.*, Primary transcripts of microRNAs encode regulatory peptides. *Nature* **520**, 90-93 (2015).
25. H. Zhou *et al.*, A peptide encoded by pri-miRNA-31 represses autoimmunity by promoting T(reg) differentiation. *EMBO Rep* **23**, e53475 (2022).
26. A. Montigny *et al.*, Drosophila primary microRNA-8 encodes a microRNA-encoded peptide acting in parallel of miR-8. *Genome Biol* **22**, 118 (2021).
27. A. Sharma, P. K. Badola, C. Bhatia, D. Sharma, P. K. Trivedi, Primary transcript of miR858 encodes regulatory peptide and controls flavonoid biosynthesis and

development in Arabidopsis. *Nat Plants* **6**, 1262-1274 (2020).
28. J. M. Couzigou, D. Lauressergues, G. Becard, J. P. Combier, miRNA-encoded peptides (miPEPs): A new tool to analyze the roles of miRNAs in plant biology. *RNA Biol* **12**, 1178-1180 (2015).
29. Y. Gao *et al.*, Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **368**, 779-782 (2020).
30. B. Malone, N. Urakova, E. J. Snijder, E. A. Campbell, Structures and functions of coronavirus replication-transcription complexes and their relevance for SARS-CoV-2 drug design. *Nat Rev Mol Cell Biol* **23**, 21-39 (2022).
31. Y. Maida, M. Yasukawa, K. Masutomi, De Novo RNA Synthesis by RNA-Dependent RNA Polymerase Activity of Telomerase Reverse Transcriptase. *Mol Cell Biol* **36**, 1248-1259 (2016).
32. P. Ahlquist, RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science* **296**, 1270-1273 (2002).
33. L. M. Iyer, E. V. Koonin, L. Aravind, Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct. Biol.* **3**, 1 (2003).
34. P. Kapranov *et al.*, New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism. *Nature* **466**, 642-646 (2010).
35. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

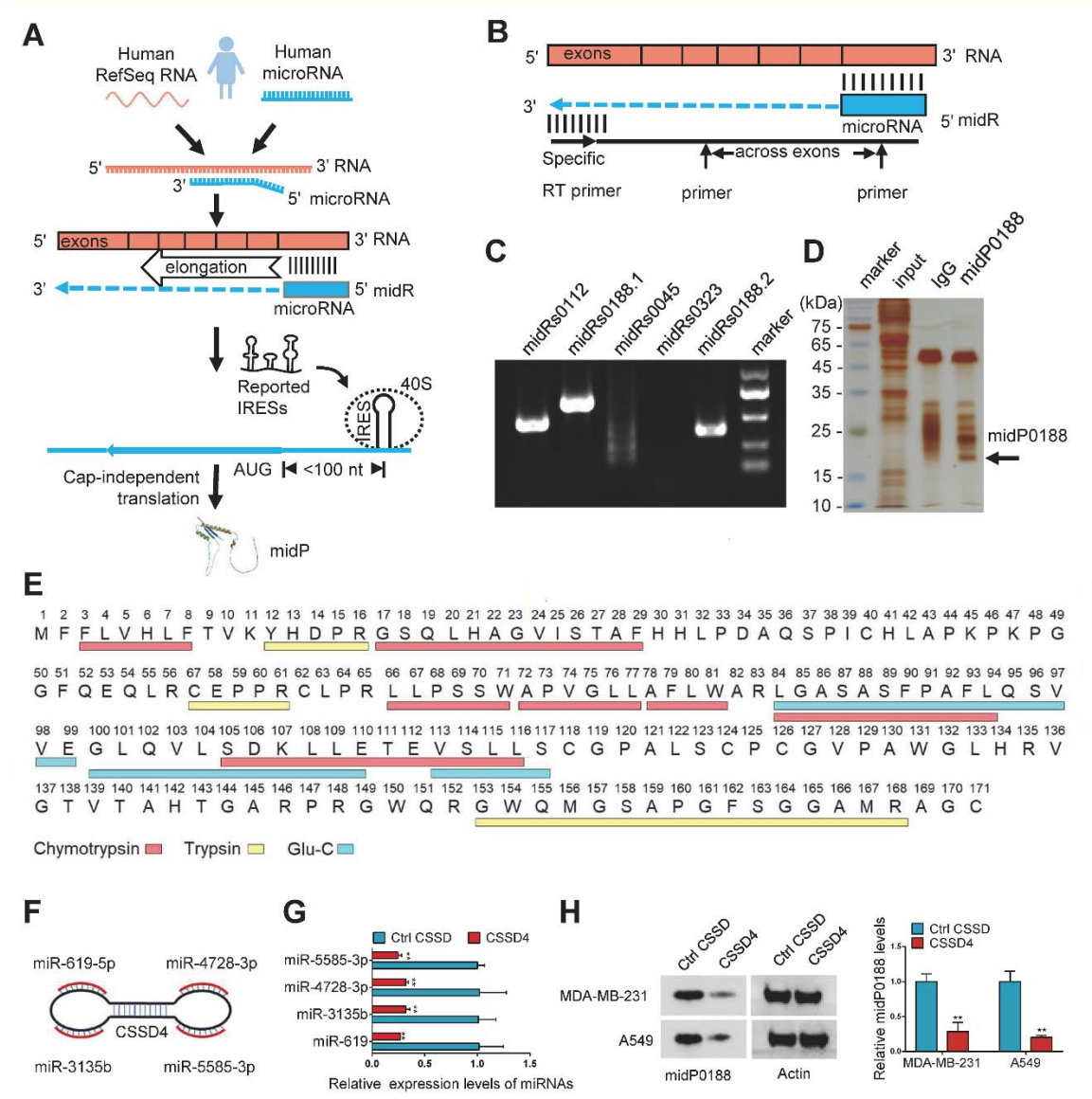

Fig. 1. Prediction and validation of miRNA-derived RNAs (midRs) and proteins (midPs). (**A**) Bioinformatic workflow to predict midRs and midPs. (**B**) Scheme for designing specific reverse transcription (RT) primer and amplification exon-crossing PCR primers for midR. (**C**) Agarose gel electrophoresis analysis of midR fragments amplified by specific primers crossing exons. (**D**) Silver staining of proteins captured by the anti-midP0188 polyclonal antibody after SDS-PAGE electrophoresis separation. The

arrow indicates the band of midP0188. (**E**) Mass spectrometry analysis to confirm midP0188 after digestion using trypsin, chymotrypsin, and Clu-C. (**F**) Schematic diagram of CSSD4 that can simultaneously adsorb four miRNAs (*miR-619-5p*, *miR-5585-3p*, *miR-4728-3p*, and *miR-3135b*) bound to the *AP1M1* mRNA with ≥ 12 nt. (**G**) PCR-determined expression levels of *miR-619-5p*, *miR-5585-3p*, *miR-4728-3p*, and *miR-3135b* in A549 cells after CSSD transfection. (**H**) midP0188 expression levels in A549 and MDA-MB-231 cells after CSSD transfection.

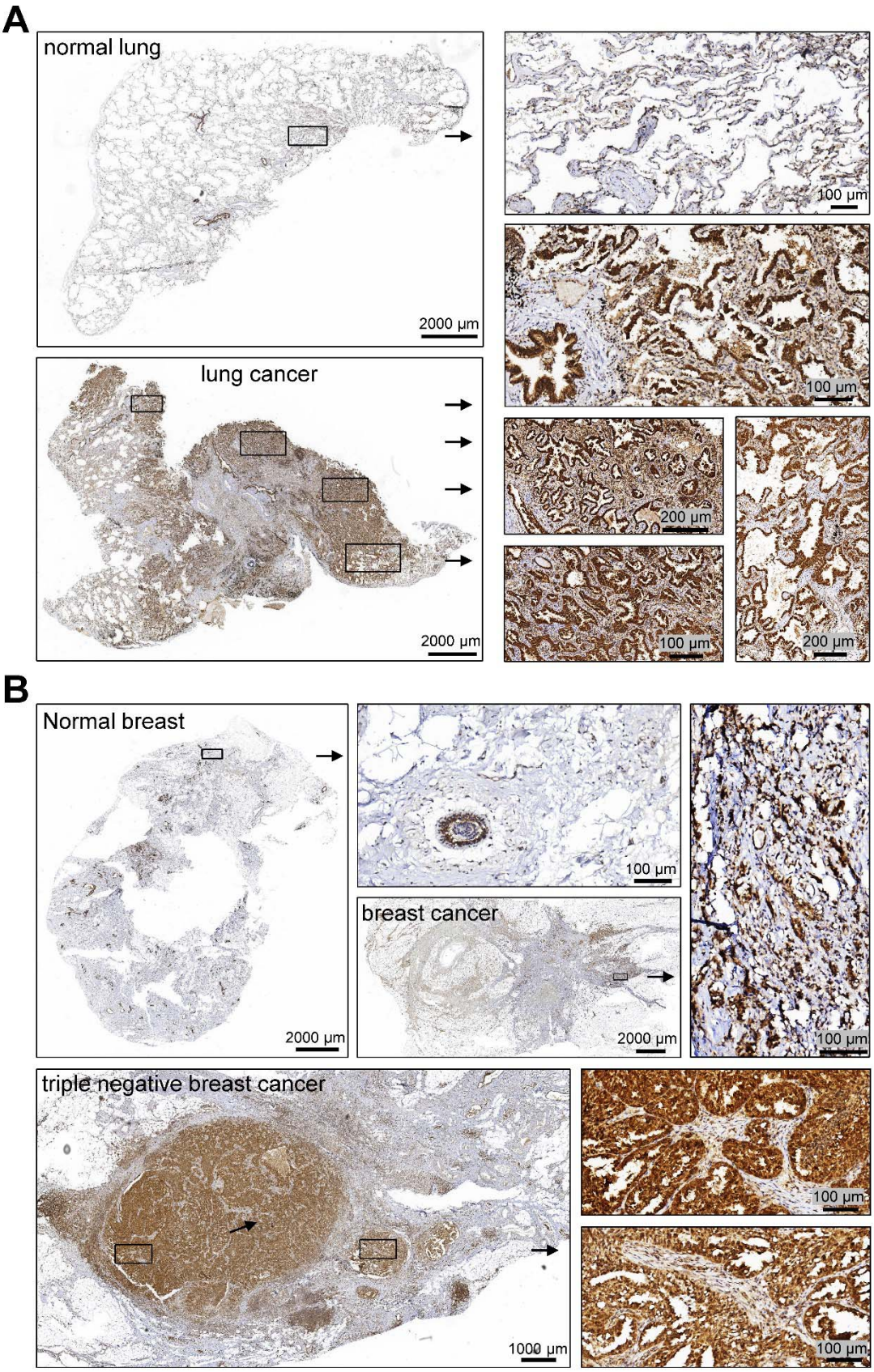
A
normal lung
2000 µm
100 µm
100 µm
lung cancer
2000 µm
200 µm
100 µm
200 µm
B
Normal breast
2000 µm
100 µm
breast cancer
2000 µm
100 µm
triple negative breast cancer
1000 µm
100 µm
100 µm

Fig. 2. High expression of midP0188 in lung cancer and breast cancer. (**A**) Representative immunohistochemical staining of midP0188 in normal lung and lung cancer tissues using the anti-midP0188 polyclonal antibody (1:200 dilution). The left side shows the unselected full immunohistochemical section; the enlarged typical staining areas are shown on the right side. (**B**) midP0188 staining intensity in normal breast, general breast cancer, and triple-negative breast cancer with the anti-midP0188 polyclonal antibody (1:200 dilution). Typical staining areas in the full section were zoomed in.

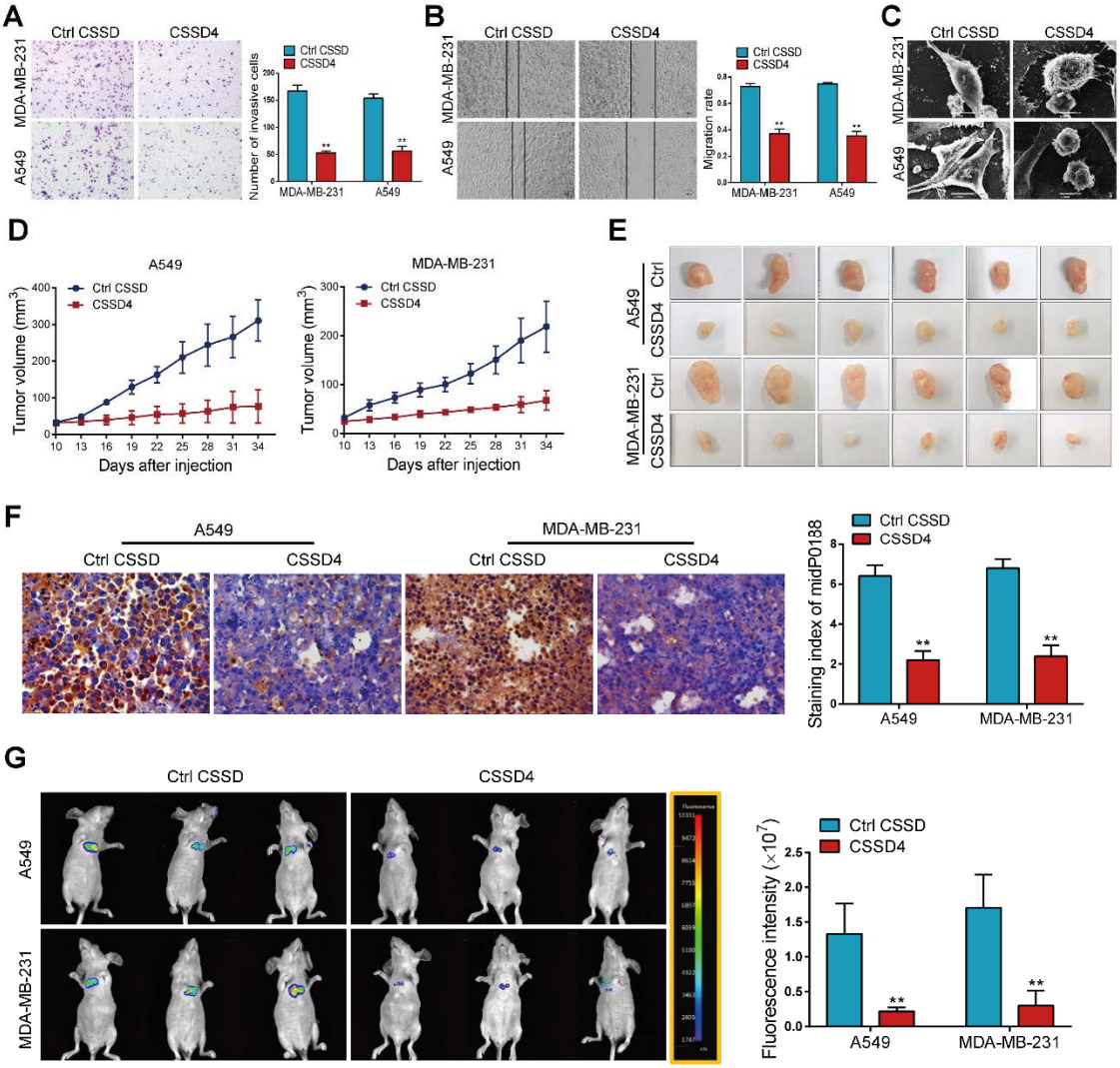

Fig. 3. CSSD4 inhibits the progression of lung and breast cancer cells *in vitro* and *in vivo*. (**A**) Effects of CSSD on the invasion of A549 and MDA-MB-231 cells. (**B**) Effects of CSSD on the migration of A549 and MDA-MB-231 cells. (**C**) Cell phenotype change detected by scanning electron microscopy after CSSD transfection. (**D**) Solid tumors of A549 and MDA-MB-231 cells in nude mice after CSSD treatment. (**E**) The volume of the tumors in Fig. 3D. (**F**) Immunohistochemical staining intensity of midP0188 in solid tumors. (**G**) Lung metastasis of A549 and MDA-MB-231 cells after CSSD treatment.

A549, lung cancer cell line. MDA-MB-231, breast cancer cell line.

Supplementary Materials

Discovery of microRNA-derived RNAs and proteins in human cells

Haidong Wu[1,#], Xiaoqiang Huang[2,#,*], Weilong Zhong[3,#], Wenbo Li[4,#], Zhiyong Liu[7,#], Min Zhao[6], Xiaonan Xi[8], Bo Cao[9], Yong Pu[10], Xiangxu Kong[3], Huan Zhao[11], Ronghua Zhang[5], Keguan Lai[1], Xinze Lv[1], Yue Lv[1], Jingyu Bao[1], Meimei Wang[5], Yanan Xiong[5], Lijie Dong[4], Jifeng Zhang[2], Guangling Zhang[5,*], Yuanjing Hu[12,*], Jie Xu[2,*], Y. Eugene Chen[2,*], Shuang Chen[1,*]

[#]These authors contributed equally to this work.

[1]Tianjin Key Laboratory of Early Druggability Evaluation of Innovative Drugs, Tianjin International Joint Academy of Biomedicine, Tianjin, China

[2]Center for Advanced Models for Translational Sciences and Therapeutics, Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, USA

[3]Department of Gastroenterology and Hepatology, Tianjin Medical University General Hospital

[4]Department of Ophthalmology, Eye Center of Tianjin Medical University

[5]Hebei Provincial Key Laboratory of Medical-Industrial Integration Precision Medicine, School of Clinical Medicine, North China University of Science and Technology

[6]Department of Pathology, Qinhuangdao Municipal No.1 Hospital

[7]Health Science Center, North China University of Science and Technology

[8]College of Pharmacy, Nankai University

[9]Department of Dermatology, Tianjin TEDA Hospital

[10]Department of drug development, Tianjin RuiJiYin Biotechnology Co., Ltd

[11]Department of Analytical Technology, CanSino Biologics Inc.

[12]Gynecologic Oncology Department, Tianjin Central Hospital of Gynecology Obstetrics

*Correspondence: xiaoqiah@umich.edu (X.H.), zhanggl@ncst.edu.cn (G.Z.), julianna_hu@163.com (Y.H.), jiex@umich.edu (J.X.), echenum@umich.edu (Y.E.C.), and shuang7332@outlook.com (S.C.)

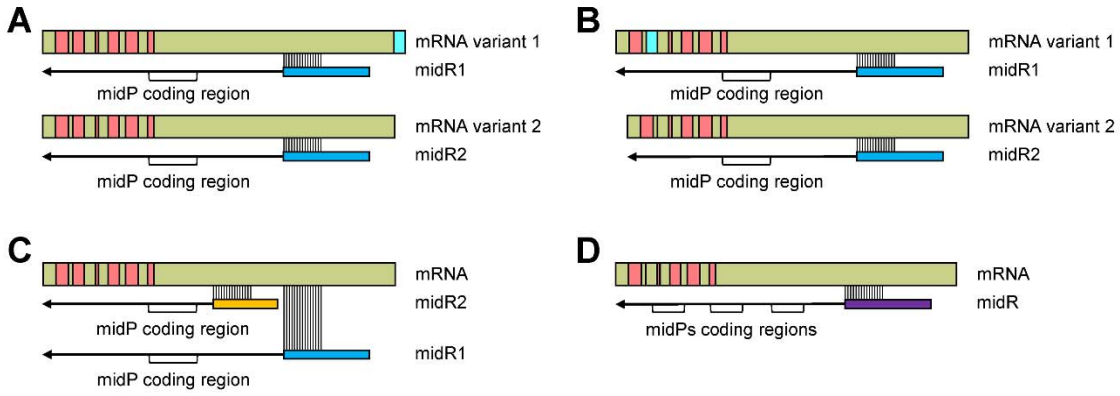

Fig. S1. Scenarios for midR and midP derivation. (**A**) RNA splicing beyond the miRNA elongation region yields identical midRs (midR1 = midR2). The excised exon in mRNA variant 1 is shown in cyan. (**B**) RNA splicing within the miRNA elongation region yields different midRs (midR1 ≠ midR2) which can encode identical midPs. (**C**) Different miRNAs bound to the same RNA to derive different midRs (midR1 ≠ midR2) which can encode identical midPs. (**D**) One midR may encode multiple midPs.

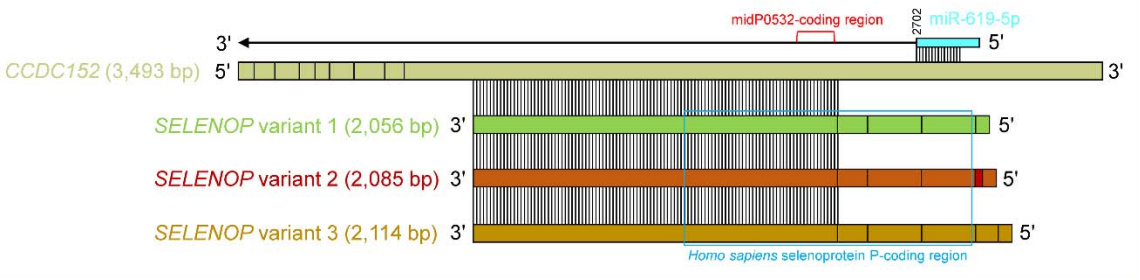

Fig. S2. Derivation of midP0532 and related midR based on *miR-619-5p* and mRNA *CCDC152*. *CCDC152* has a 1,453 bp-long reverse complementarity to all three SELENOP variants that can encode the protein, selenoprotein P. midP0532 is the same as the selenoprotein P, plasma, 1 encoded by *SEPP1* (a truncated coding region of *SELENOP*) through conceptual translation. The exons on each mRNA are shown as rectangle blocks.

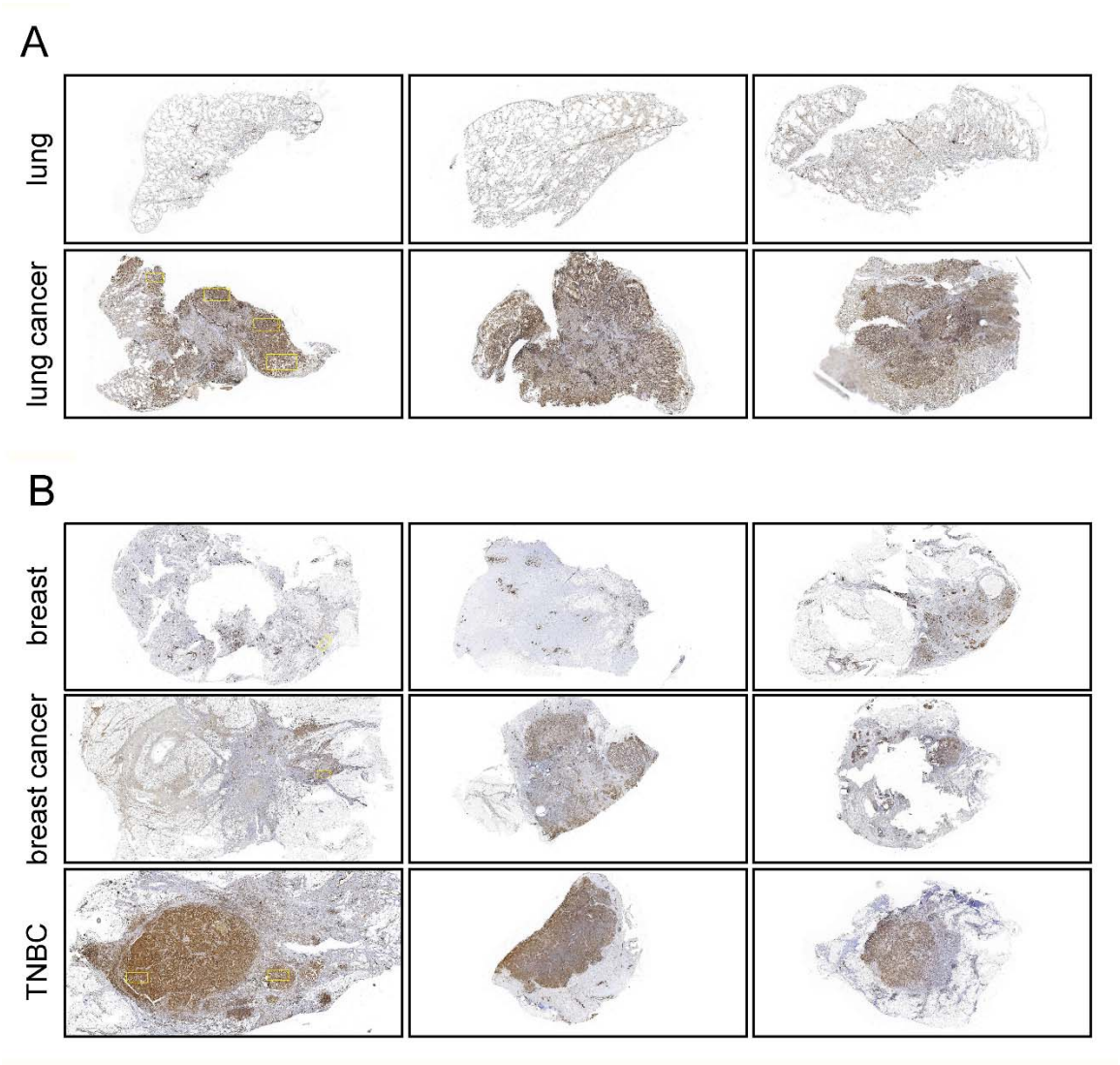

Fig. S3. Immunohistochemistry staining for tissues from different patients using the anti-midP0188 polyclonal antibody. (**A**) midP0188 staining full section images for three normal lungs and three lung cancer tissues. (**B**) midP0188 staining full section Images for three normal breast tissues, three breast cancer tissues, and three triple-negative breast cancer (TNBC) tissues. Each image is from a different patient.

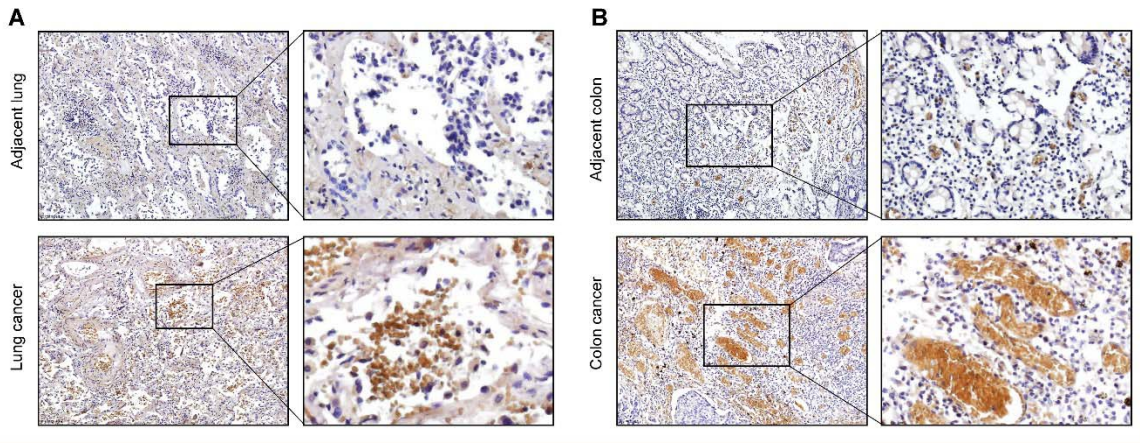

Fig. S4. Expression of midP0112 in lung and colon cancer tissues. (**A**) Immunohistochemical staining intensity with the anti-midP0112 polyclonal antibody in lung cancer and adjacent lung tissues (left). Sections bounded by black rectangles are magnified to 100x (right) for each core section. (**B**) The staining intensity with the anti-midP0112 polyclonal antibody in colon cancer and adjacent colon specimens (left). Sections bounded by black rectangles are zoomed in by 100x (right) for each core section.

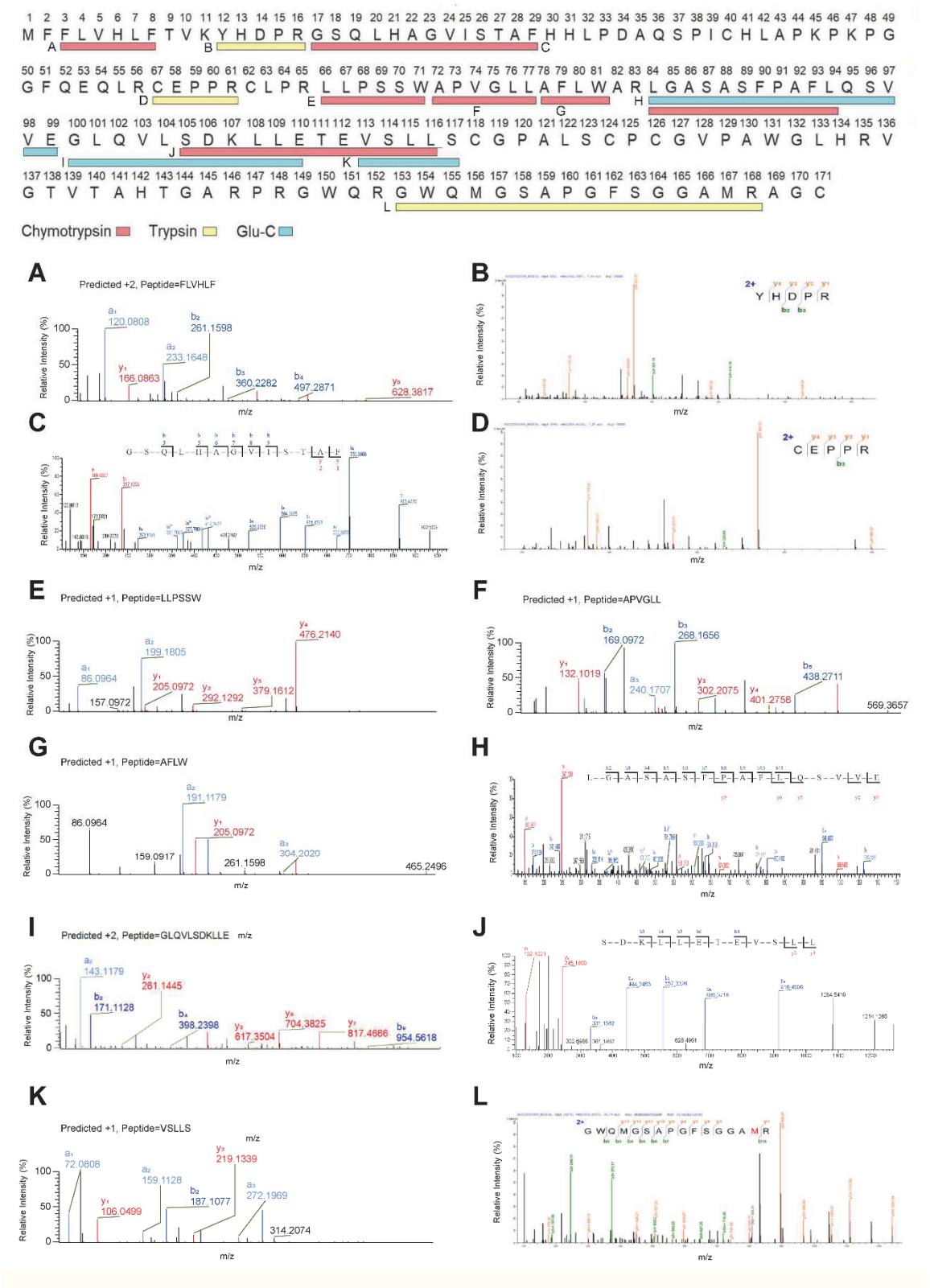

Fig. S5. Mass spectrometry assay to verify midP0188 peptide fragments after

protease digestion. The protein binding to the anti-midP0188 polyclonal antibody in immunoprecipitation was separated by SDS-PAGE and silver-stained to obtain a potential target band, which was harvested and digested using different proteases. The digested samples were analyzed by mass spectrometry. Thirteen peptide fragments appearing in midP0188 were confirmed, of which three were from trypsin digestion (yellow), three were from Glu-C digestion (cyan), and seven were from chymotrypsin digestion (red).

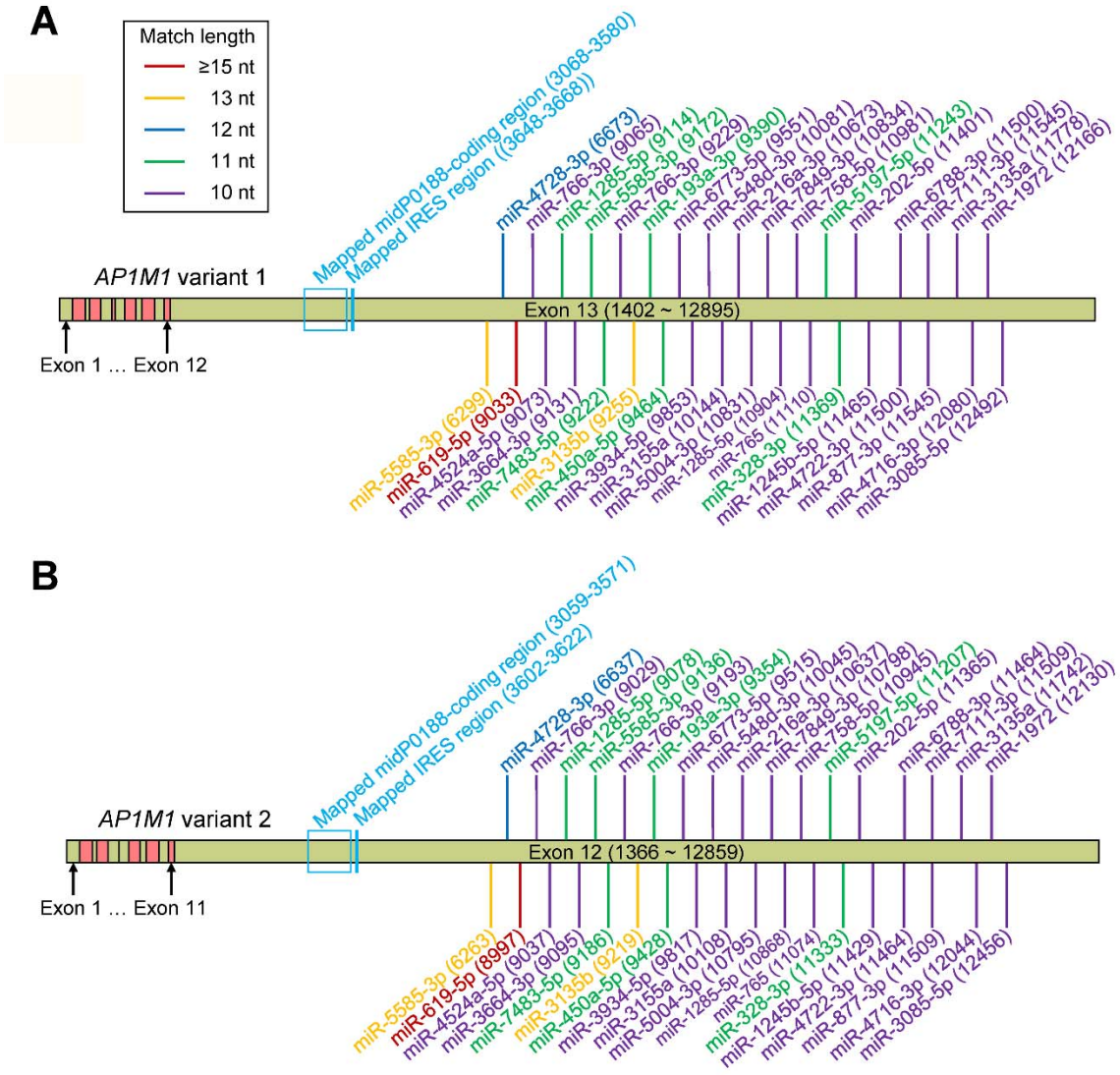

Fig. S6. Example miRNAs bound to the *AP1M1* mRNA with 10- to 15-nt match. (**A**) For variant 1, all ≥ 10-nt matches are shown after position 9033 and only ≥ 12-nt matches are shown before this position. (**B**) For variant 2, all ≥ 10-nt matches are shown after position 8997 and only ≥ 12-nt matches are shown before this position.

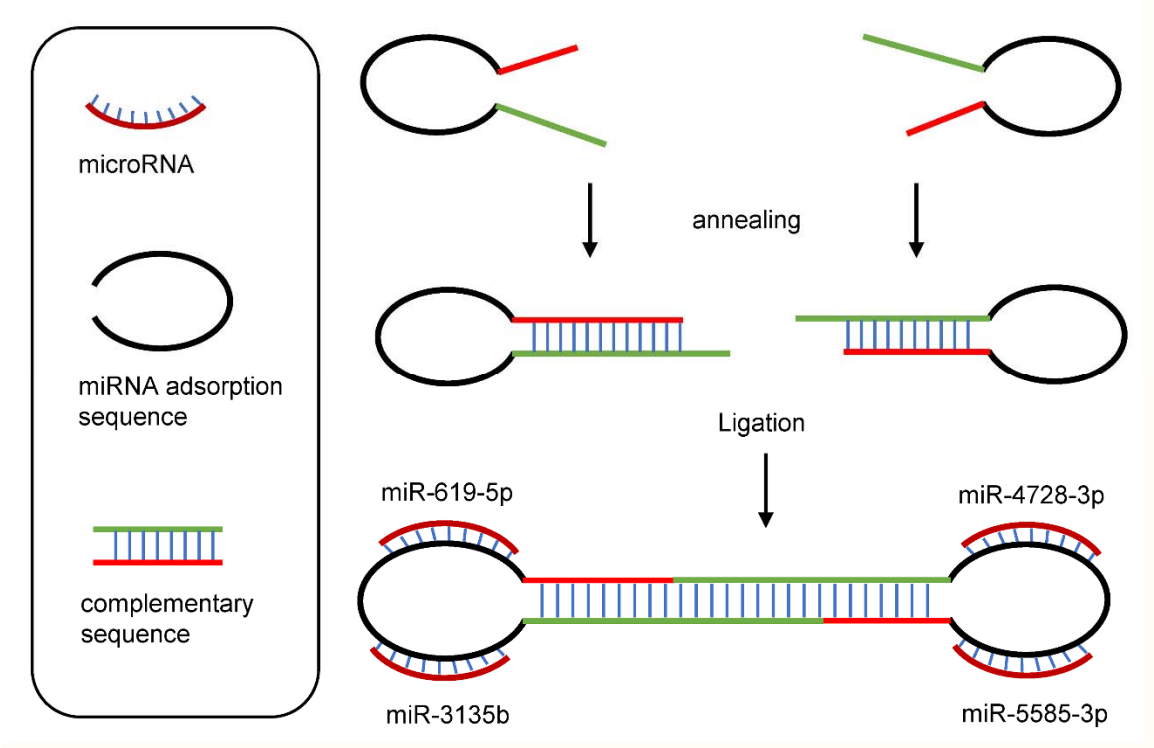

Fig. S7. Diagram of the route to design CSSD4. CSSD4 is designed to bind *midR-619-5p*, *miR-3135b*, *miR-4728-3p*, and *miR-5585-3p* simultaneously. Both CSSD4 (which contains multiple miRNAs complementary binding sites) and Ctrl-CSSD (which contains a scrambled DNA sequence) are composed of two DNA half-rings. Single-stranded DNAs were first synthesized and formed half-ring after annealing. These half-rings were then treated with T4 DNA ligase to form a completely closed circular DNA.

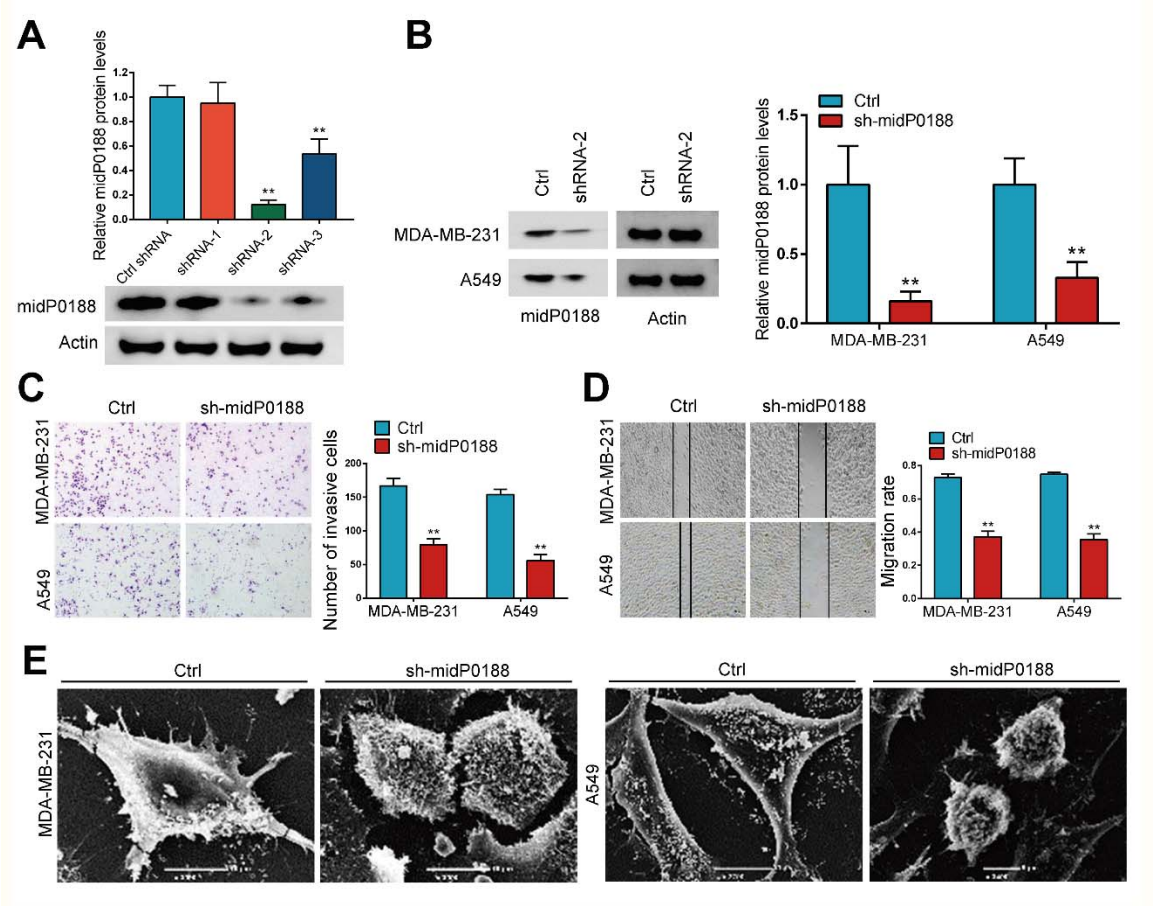

Fig. S8. Downregulation of midP0188 by transfection of shRNAs inhibits the progression of lung and breast cancer cells *in vitro*. (**A**) Three interfering lentiviruses targeting midR0188 and a control lentivirus containing a scramble sequence were transfected into A549 cells. midP0188 protein level was detected by western blot to verify the Efficacy comparison of three shRNAs. (**B**) Downregulation of midP0188 levels in A549 and MDA-MB-231 cells. (**C**) Effects of sh-midP0188 on the invasion of A549 and MDA-MB-231 cells. (**D**) Effects of sh-midP0188 on the migration of A549 and MDA-MB-231 cells. (**E**) Observation of cell phenotype by scanning electron microscopy after sh-midP0188 transfection. A549, lung cancer cell line. MDA-MB-231, breast cancer cell line.

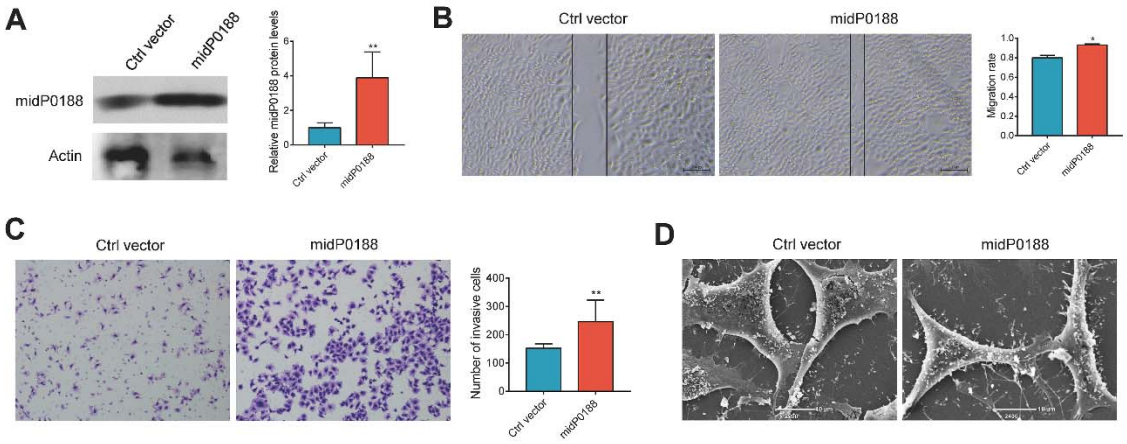

Fig. S9. Overexpression of midP0188 promotes A549 cells' migration and invasion. A549 cells were transfected with an empty vector (Ctrl) or the PiggyBac Dual promoter plasmid with midP0188 coding sequence. (**A**) Western blot to detect the protein level of midP0188. (**B**) Effects of midP0188 overexpression on the migration of A549 cells. (**C**) Effects of midP0188 overexpression on the invasion of A549 cells. (**D**) Effects of midP0188 overexpression on the phenotype of the A549 cells visualized by scanning electron microscopy.

Table S1. Summary of the number of matched RNAs, midRs, and midPs derived from 15- to 10-nt miRNA–RNA reverse matches.

Match length	Number of matched RNAs	Number of unique midRs[1]	Number of unique (and redundant) midPs[2]
15 nt	13,310	11,453	1,239 (5,369)
14 nt	29,079	24,492	2,135 (9,629)
13 nt	81,633	67,350	4,014 (22,829)
12 nt	246,696	201,067	4,252 (34,225)
11 nt	835,868	680,634	4,397 (71,208)
10 nt	2,953,351	2,401,247	3,784 (162,694)

[1] The number of redundant midRs is identical to the number of matched RNAs.

[2] Theoretically, it is expected that unique midPs derived from shorter matches will always cover those derived from longer matches, but this may not be true in practice because NCBI BLAST is a heuristic algorithm which means it may not report all alignments it finds, and hence there is a chance that previous findings disappear as the database (of midRs) grows.

Table S2. Summary of four selected midPs for immunohistochemistry assay.

midP name	Length (AA)	Sequence	Immunogen sequence
midP0045	15	MANIFPVFAPVTFRT	CANIFPVFAPVTFRT
midP0112	25	MDFSTKHPVFSLKHKCHHVTGSCLT	PVFSLKHKCHHVTGSC
midP0188	171	MFFLVHLFTVKYHDPRGSQLHAGVISTAFHHLPDAQSPICHLAPKPKPGGFQEQLRCEPPRCLPRLLPSSWAPVGLLAFLWARLGASASFPAFLQSVVEGLQVLSDKLLETEVSLLSCGPALSCPCGVPAWGLHRVGTVTAHTGARPRGWQRGWQMGSAPGFSGGAMRAGC	LAPKPKPGGFQEQLRC
midP0323	68	MGFYHVGQAGLELLNSSDPPATASQSAGITGMGHRAQPLRSDFKQLYICELIYSGTHSVFIIIIVCIF	CGMGHRAQPLRSDFKQ

The predicted immunogenic peptide fragments in midPs are underlined, and they were further designed into peptide immunogens by adding a cysteine at the N terminus when the N-terminal or C-terminal amino acid is not cysteine.

Table S3. Summary of the identifiers and names for selected midPs and corresponding midRs.

midP name	midP and midR Identifier(s)[1]	midR name[2]
midP0045	midP.1370471854(RPS6KB1).414, midR.hsa-miR-1972.1370471854(RPS6KB1).2511	midRs0045
midP0112	midP.1370451977(NAXE).1372, midR.hsa-miR-5585-3p.1370451977(NAXE).3203	midRs0112
midP0188	midP.1677479768(AP1M1).3571, midR.hsa-miR-619-5p.1677479768(AP1M1).8997; midP.1677502062(AP1M1).3607, midR.hsa-miR-619-5p.1677502062(AP1M1).9033	midRs0188
midP0323	midP.1819229406(C9orf129).1643, midR.hsa-miR-3686.1819229406(C9orf129).4888	midRs0323

[1] The nomenclature of midP and midR identifiers can be referred to in the **Methods** section. Different midR sequences may encode the same midP sequence; such midRs can be derived from the elongation of the same miRNA bound to mRNA variants, or the elongation of different miRNAs bound to the same RNA. For instance, *AP1M1* variant 1 and variant 2 have 12,895 and 12,859 nt, respectively; their only difference is that the 36-nt exon 6 in variant 1 is sliced in variant 2, resulting in a 36-nt shift of the follow-up exons in variant 2 (**Fig. S5**). For each miRNA able to bind to variant 1 at a position after exon 6, it can also bind to variant 2 at a corresponding position; thus, two *AP1M1*

variants can yield similar midRs which may encode the same protein with different identifiers by our nomenclature (e.g. both midP.1677479768(AP1M1).3571 and midP.1677502062(AP1M1).3607 are identifiers for midP0188).

[2] The number in midRs name was arbitrarily set identical to that in midP for convenience of data presentation.

Table S4. Summary of the PCR-amplified regions of selected midRs.

midR name[1]	Amplified region[2]	RT primer[3]	Forward Primer[3]	Reverse Primer[3]
midRs0045	442–831 (exon 4–8)	ATGGCGGCAG CGGCTGTGGT	AAACTCCACCA ATCCACAGC	AGTAAAGCATC CCTTCATCG
midRs0112	186–675 (exon 1–5)	ATGGCGAGCA CGGTGGTGAA G	TGGCAATGGG CACAGTGAGT	GCGAGCACGG TGGTGAAGTA
midRs0188.1	602–1318 (exon 5–11)	GGTTGCCACAT CCAAGAAGA	CCGGAGGTAG TGAAGTAAGG	AGGCATCAAGT ATCGGAAGA
midRs0188.2	594–969 (exon 5–9)	AGACCACCGA CAGCAAGAT	CGTGGGTGTT GAGACGGTAG G	CGGTCCGAAG GCATCAAGTA
midRs0323	156–577 (exon 1–3)	GCATCTTCCAG ATCCCAAT	CGATGTAGAAG AGCCAGGGAA GTA	TGGGATGTAAC AGAGCAAGCG

[1] Two regions were amplified for midRs0188 (midRs0188.1 and midRs0188.2).

[2] Amplified regions are indexed on template RNAs.

[3] All primers are from 5’ to 3’.

Table S5. Results of BLASTP search to identify sequence analogs to four longer consecutive peptides in midP0188.

Peptide	Sequence	Result 1	Result 2
midP0188_pep1	YHDPRGSQLHAGVISTAF	N.S.S.F.	N.H.F.
midP0188_pep2	LLPSSWAPVGLLAFLW	N.S.S.F.	N.H.F.
midP0188_pep3	LGASASFPAFLQSVVEGLQVLSDKLLET EVSLL	N.S.S.F.	N.H.F.
midP0188_pep4	GWQMGSAPGFSGGAMR	N.S.S.F.	N.H.F.

Result 1 was obtained using the BLASTP server with default parameters for short queries. N.S.S.F., no significant similarity found.

Result 2 was obtained using the standalone BLASTP program with the local nr database. N.H.F., no hit found.