Discovery of microsized RNA-derived RNAs and proteins in human cells

Xiaoqiang Huang[1,#,*], Haidong Wu[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[2], Xinze Lv[2], Yue Lv[2], Jingyu Bao[2], Meimei Wang[5], Yanan Xiong[5], Lijie Dong[4], Jifeng Zhang[1], Guangling Zhang[5,*], Yuanjing Hu[12,*], Jie Xu[1,*], Y. Eugene Chen[1,*], Shuang Chen[1,2,*]

[1]Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48109, USA

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

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

[4]Tianjin Key Laboratory of Retinal Functions and Diseases, Tianjin Medical University Eye Hospital

[5]North China University of Science and Technology Affiliated Hospital, North China University of Science and Technology

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

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

[8]State Key Laboratory of Medicinal Chemical Biology, 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

[#]These authors contributed equally

*Correspondence to: shuanche@umich.edu (S.C., lead contact), xiaoqiah@umich.edu (X.H.), haidong123abc@sina.com (H.W.), zhanggl@ncst.edu.cn (G.Z.), julianna_hu@163.com (Y.H.), jiex@umich.edu (J.X.), echenum@umich.edu (Y.E.C.)

Abstract

MicroRNAs (miRNAs) are a class of short noncoding RNAs that can 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 exact reverse pairing of the 3'-ends of miRNAs or miRNA-like RNAs, collectively referred to as microsized RNAs (msRNAs), with template RNAs could initiate the production of msRNA-derived RNAs (msdRs), which consequently translate into msRNA-derived proteins (msdPs). Starting with 2,632 human miRNAs or miRNA-like msRNAs taken from the miRBase database, we predicted 11,121 and 1,239 unique msdRs and msdPs respectively using a 15-nt pairing threshold. We verified the authentic presence of example msdRs and msdPs in human cells. Of clinical relevance, we demonstrated that msdP0188 is highly expressed in human lung and breast cancer tissues and cells, and its corresponding msdRs represent novel anti-cancer targets. Intriguingly, inhibiting the human telomerase reverse transcriptase, a putative RNA-dependent RNA polymerase suggested by bioinformatic screening, led to reduced levels of msdP0188 in human cells. Our findings propose a novel "msRNA → msdR → msdP" axis that not only represents an augmentation to the central dogma of molecular biology but also 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.

One-Sentence Summary: This work reports the discovery of msRNA-derived RNAs and proteins that may have significant roles in biology and medicine.

Keywords: small RNA, microRNA (miRNA), microsized RNA (msRNA), msRNA-derived RNA (msdR), msRNA-derived protein (msdP)

Introduction

MicroRNAs (miRNAs) are a class of short noncoding RNAs pivotal in regulating gene expression post-transcriptionally. They play crucial roles in various biological processes (*1-6*). It is well established that miRNAs exert their regulatory functions by binding their 5'-ends to complementary sequences on target mRNAs (*7-9*). A question thus comes naturally: could the pairing of a miRNA's 3'-end with a complementary RNA induce any biological effects? To our knowledge, this question remains largely unexplored. Drawing inspiration from the RNA replication mechanism facilitated by RNA-dependent RNA polymerase (RdRP) in viruses (*10, 11*), we hypothesize that a miRNA or a miRNA-like RNA, collectively referred to as microsized RNA (msRNA), upon its 3'-end pairing with a template RNA, might act as a primer, initiating the synthesis of novel RNAs (msRNA-derived RNAs, msdRs) that are unlikely to be produced via conventional transcription processes. Furthermore, we postulate that msdRs may encode proteins (msRNA-derived proteins, msdPs).

In this study, we developed a bioinformatic workflow aimed at predicting msdRs and msdPs, followed by experimental validation of the bona-fide existence of example msdRs and msdPs in human cells. We demonstrate that an example msdP, namely msdP0188, is highly enriched in human lung and breast cancer tissues and that msdP0188-encoding msdRs represent novel anti-cancer targets. Our work also implicates telomerase reverse transcriptase (TERT) as a potential RdRP involved in the production of msdRs and msdPs. This unconventional "msRNA → msdR → msdP" information pathway presents an augmentation to the central dogma and potentially predicts thousands of previously unknown RNAs and proteins that may have significant biological and pathological functions in human cells and other biological contexts.

Results

Prediction of msdRs and msdPs

The workflow for predicting msdRs and msdPs is depicted in Fig. 1A. The 3'-end of a msRNA sequence, e.g., the last 15 nt, is queried against the NCBI RefSeq RNA database (*17*) using BLASTN (*18*) to search for RNAs that show full complementarity to the msRNA. Upon finding a match, the msRNA sequence is designated as the 5'-end of the putative msdR, and the full msdR sequence is then generated by extending the msRNA along the template RNA until the RNA's 5'-end is reached. We then search for potential internal ribosome entry sites (IRESs) within the msdR, enabling translation initiation in a cap-independent manner (*19, 20*). The prediction of IRES involves aligning each msdR sequence with experimentally validated IRESs. We assume the existence of a potential IRES if both *coverage* and *aligned identity* are sufficiently high (see Methods). Upon identifying a putative IRES and confirming the presence of a start codon within 100 nt downstream of the IRES, we proceed to predict a msdP starting from the start codon. One msdR may contain multiple IRESs, and within 100 nt of each IRES, there may exist multiple start codons; all such scenarios are taken into account to predict msdPs.

Initially, we were interested in employing miRNAs as example msRNAs because they are tiny (~22 nt) and are known to bind RNAs. Starting with 2,632 unique human mature miRNAs from miRBase (*21*) (table S1, data S1) and a matching threshold of 15 nt, we predicted 11,121 unique msdRs (miRNA templates were

excluded) and 1,239 unique msdPs after discarding msdPs with <10 amino acids (AAs). The unique msdRs and msdPs arising from matching thresholds of 10 to 14 nt were also predicted (table S2, data S2, data S3).

We then estimated the potential false discovery rate (FDR) for msdR prediction by comparing the number of msdRs predicted with the 2,632 miRBase miRNAs (i.e., predicted positive, PP) and the mean number of msdRs predicted with nine sets of the same number of shuffled msRNA artifacts (i.e., false positive, FP) for various matching lengths (see Methods, fig. S1), using a shuffling method adopted from previous studies on miRNA target prediction (*22, 23*). For the 15-nt match, the PP and simulated FP were 11,121 and 1,695 ± 245, respectively, leading to an FDR of 15.2 ± 2.2%; the estimated FDRs increased substantially for shorter matches (table S3). We further estimated the potential range of FDRs using different ways to simulate msRNA artifacts (fig. S1, Methods). We found that using completely randomly generated msRNA artifacts further reduced FDRs and yielded an FDR of 10.9 ± 0.9% for the 15-nt match (table S4). In contrast, using the much less-random msRNA artifacts generated by shifting the 3'-ends of miRBase miRNAs resulted in increased FDRs and yielded an FDR of 35.2 ± 2.7% or 39.4 ± 7.6% for the 15-nt match by right or left shifting, respectively (table S5, table S6). Together, these results indicated that our pipeline showed much stronger signals for prediction with miRBase miRNAs compared to those with simulated msRNAs. We also wonder if our method could distinguish the database-documented miRNAs from their reverse complements or passenger miRNAs (fig. S1, Methods). For the former, the miRBase human miRNAs were used to generate reverse complements, while for the latter, we took human mature and passenger miRNAs from mirGeneDB (*24-26*), a database of manually curated miRNA genes that have been validated and annotated. Interestingly, both the reverse complements and passenger miRNAs produced more predicted msdRs than the parental miRNAs (table S7, table S8). We reason that these control miRNAs may be as meaningful as the mature counterparts as they are parts of the human genome rather than unrelated entities. The simulated msRNA sequences and total numbers of msdRs predicted with each set of either experimental or control msRNAs are presented in data S4 and data S5, respectively.

A nomenclature rule has been established to assign identifiers for msdRs. The rule comprises the msRNA name, the template RNA GenInfo Identifier (and gene symbol), and the msRNA 3'-end matching position on the template RNA; the identifiers for msdPs were similarly defined (see Methods). To facilitate data presentation, the 1,239 msdP proteins were also sequentially named (msdP0001–msdP1239) following the sequences' ASCII (American Standard Code for Information Interchange) order. The msdP names, msdP identifiers, msdR(s) identifiers, and template RNAs are provided (data S2, data S3). msRNAs, msdRs, and msdPs do not always exhibit a one-to-one correspondence. For instance, one msdR may produce multiple msdPs (fig. S2A), distinct msRNAs bound to one RNA may yield different msdRs but encode identical msdPs (fig. S2B), and one msRNA may bind to multiple RNAs (e.g., different RNAs or splicing variants), leading to either different or identical msdR/msdP outcomes (fig. S2, C and D).

Among the 1,239 msdPs, 46 exhibit high sequence similarity with entries in the NCBI nr (non-redundant) protein database (*17*) according to BLASTP (*18*) search, demonstrating high *coverage* and *aligned identity* (data S6; see Methods). Six msdPs possess exact sequence matches within the nr database; of note, the 100-AA msdP0532 corresponds to selenoprotein P, plasma, 1, encoded by *SEPP1* (a truncated coding sequence of *SELENOP*) through conceptual translation (*27*) (data S7). Our prediction suggests that msdP0532 is

encoded by *msdR.hsa-miR-619-5p.1519242743(CCDC152).2702*, generated through the elongation of msRNA *hsa-miR-619-5p* bound to *CCDC152* (coiled-coil domain containing 152; a 3,493-nt mRNA) at position 2702. *CCDC152* exhibits a 1,453 bp-long reverse complementarity to *SELENOP* (fig. S3). *CCDC152* was recently reported to function as an endogenous long noncoding RNA that suppresses selenoprotein P translation (*28*). Hence, the prediction of msdP0532 and the documented *SEPP1* exemplify the efficacy of our workflow in predicting msdRs and msdPs.

The remaining 1,193 msdPs (96%) lack a high-similarity sequence analog in the NCBI nr database (data S8), thus representing a fresh cohort of proteins within human cells.

The potential presence of msdRs and msdPs in human cells

The 1,193 msdPs were then compared to experimental mass spectrometry (MS) spectra obtained from four tumor cell lines using the Sequest-HT (*29*) search engine (fig. S4, Methods). It was revealed that 941 msdPs (79%) have high matching scores that included considerations of protein sequence coverage, the number of peptide-spectrum matches (PSMs), and the Sequest-HT score, suggesting their potential existence in human tumor cells (data S9).

We also utilized PepQuery2 (*30*) to interrogate the same 1,193 msdPs against PepQueryDB (*30*) (fig. S4), a database that constitutes a comprehensively indexed over a billion (1,012,696,284) MS/MS spectra. Through this pipeline, we identified PSMs with high confidence suggested by PepQuery2 for 1,095 peptide fragments from 555 msdPs (data S10). Between the 941 hits identified by Sequest-HT and the 555 hits by PepQuery2, 459 are common; this comparison suggested the potential presence of many msdPs in human cells.

We randomly selected four msdPs (msdP0045, msdP0112, msdP0188, and msdP0323) to investigate the presence of their corresponding msdRs in the A549 lung cancer cells (table S9, table S10). For each putative msdR(s), primers were designed to amplify the sequences that only exist within the complementary DNA (cDNA) sequences of the msdR but not within the human genomic DNA sequences (Fig. 1B). Two fragments were amplified for msdP0188-encoding msdRs for dual verification, and one for each remaining msdR(s) (Fig. 1C, table S11). Reverse transcription (RT)-PCR and Sanger sequencing confirmed the presence of RNA fragments of msdP0112- and msdP0188-encoding msdRs (Fig. 1C, data S11), whereas RNA fragments were not detected for msdRs predicted to encode msdP0045 and msdP0323 (Fig. 1C).

Next, we investigated further the presence of msdPs. We developed msdP-specific peptide immunogens to generate rabbit polyclonal antibodies (pAbs) against chosen msdPs (table S9) and demonstrated the high sensitivity and specificity of the pAbs (see Methods, fig. S5, fig. S6, fig. S7). The pAbs were employed to detect the expression of msdP0112 and msdP0188 in human cancer tissues and normal tissues (control) using IHC assays. Compared to the control, the anti-msdP0188 pAb led to strong, positive staining in lung and breast cancer tissues collected from various patients (Fig. 2, fig. S8). Similarly, the anti-msdP0112 pAb resulted in robust positive staining in lung and colon cancer tissues (fig. S9). These results suggest the potential expression of the two msdPs in human cancer tissues.

Collectively, these findings imply the presence of msdPs and their corresponding msdRs endogenously

encoded within human tumor cells, hinting at their potential involvement in cancer biology.

Bona fide presence of msdP0188 in human cells

We next worked to further confirm the existence of msdP0188. Following SDS-PAGE and silver staining, a ~18 kDa protein band that matches the molecular weight of msdP0188 was identified in the A549 cell lysates (Fig. 1D). This band was harvested, desilverized, reduced and alkylated, and then digested using trypsin, chymotrypsin, or Glu-C separately, followed by MS analysis. Thirteen peptide fragments within msdP0188 were confirmed: three from trypsin digestion, three from Glu-C, and the remaining seven from chymotrypsin (Fig. 1E, fig. S10, data S12). These fragments cover 55% (94/171) AAs of msdP0188. Querying four longer consecutive fragments against the NCBI nr protein database with BLASTP (*18*) returned no hits of high similarity (table S12), suggesting that these peptides are previously unknown biological entities.

We proceeded to assess whether knocking down the msdP0188-deriving msRNAs would effectively reduce the expression of msdP0188. msdP0188 was predicted by using a stringent match threshold of 15 nt between a msRNA's 3'-end and RNA. Specifically, *hsa-miR-619-5p* is the msRNA that was predicted to bind mRNA *AP1M1* variants 1 or 2, both giving rise to msdRs that may encode msdP0188. We note however, that if we relax the threshold (e.g., from 15 to 14 nt, or fewer), there may be additional msRNAs that could give rise to msdP0188-encoding msdRs (fig. S11). Therefore, if we want to knock down msdP0188, we should consider targeting at least the most potential originating msRNAs. We relaxed the threshold to 12 nt and identified four msRNAs that may eventually give rise to msdP0188, i.e., *hsa-miR-619-5p*, *hsa-miR-3135b*, *hsa-miR-4728-3p*, and *hsa-miR-5585-3p* (fig. S11).

We then designed a circular single-stranded DNA (CSSD), namely CSSD4, to knock down the four msRNAs simultaneously (Fig. 1F, fig. S12), whereas the control CSSD (Ctrl CSSD) harbors four scrambled sequences. CSSD, as demonstrated in a previous report, is a multiplex knockdown tool that works like an absorbing sponge of target msRNAs (*31, 32*). The specificity of CSSD4 was verified by checking the expression of three randomly chosen, msdP0188-unrelated msRNAs (*hsa-miR-21-5p*, *hsa-miR-15a-5p*, and *hsa-miR-34a-5p*) following transfection of CSSD4 or Ctrl CSSD into A549 cells. The results indicated that the CSSD4 transfection had no discernible impact on the expression levels of the non-target msRNAs (fig. S13). In contrast, transfection of CSSD4 to A549 cells significantly diminished the levels of all four target msRNAs (Fig. 1G), and as expected, the protein levels of msdP0188 were decreased, compared to those treated with Ctrl CSSD (Fig. 1H).

We also investigated whether targeting the msdRs directly would similarly decrease the msdP0188 protein level. We generated stable cell lines expressing short hairpin RNAs (shRNAs) to target msdP0188-coding msdRs in MDA-MB-231 and A549 cells. Western blot analysis showed that shRNA-2 exhibited the highest interfering efficacy (fig. S14A), leading to a significant decrease in the msdP0188 levels in both cell lines (fig. S14B).

Together, our efforts conclusively demonstrate the presence of msdP0188 in human cells and that msdP0188 is produced through the "msRNA → msdR → msdP" axis.

Targeting msdP0188 in lung and breast cancer cells *in vitro*

We hypothesize that certain msdPs and msdRs could serve as potential biomarkers or therapeutic targets for human diseases such as cancer. To validate this idea, we investigated whether msdP0188, its originating msRNAs, and its encoding msdRs could be viable targets for anti-cancer drug development.

Transwell and wound healing assays indicated that certain tumor cells, such as MDA-MB-231 and A549 cells, treated with CSSD4 exhibited compromised invasion (Fig. 3A) and migration (Fig. 3B) capacities compared to those with Ctrl CSSD. Scanning electron microscopy (SEM) imaging revealed that CSSD4 treatment adversely affected the morphology of these tumor cells (Fig. 3C). Similar outcomes were observed when shRNA-expressing lentiviruses were utilized to target msdP0188-encoding msdRs in these cells (fig. S14, C–E).

Conversely, when msdP0188 was overexpressed in A549 cells (fig. S15A), wound healing and transwell assays revealed enhanced cell migration (fig. S15B) and invasion (fig. S15C), compared to the controls transfected with vehicle vector, and SEM imaging also illustrated a morphological alteration (fig. S15D).

Together, these results demonstrate that targeting msdP0188-deriving msRNAs or msdP0188-encoding msdRs compromised certain tumor cells' invasion and migration *in vitro*.

Targeting msdP0188 in lung and breast cancer cells *in vivo*

We proceeded to evaluate the potential anti-cancer benefits by targeting msdP0188 in animal models. Specifically, we aimed to examine the effects of targeting msdP0188 on tumor growth and metastasis *in vivo*. To achieve this, we established a subcutaneous solid tumor xenograft model and a tail vein injection metastasis mouse model.

In the subcutaneous solid tumor model, 24 five-week-old BALB/c nude mice were randomly divided into four groups, and A549 or MDA-MB-231 cells were subcutaneously inoculated at a concentration of 5×10^6 cells/mouse. Once the tumors reached approximately 4 mm in diameter, CSSD4 or Ctrl CSSD (20 μg/mouse) was intratumorally injected into the tumor every seven days for a total of three doses. 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 indicated that tumor sizes were significantly smaller in CSSD4 groups compared to control groups for both tumor cell types (Fig. 3E). Additionally, IHC staining revealed a reduction of msdP0188 levels in CSSD4 groups (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 cell lines stably expressing luciferase were established and injected into nude mice via 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 (on day 10, day 17, and day 24). The results indicated that the animals treated with CSSD4 experienced fewer lung metastases compared to those with Ctrl CSSD for both tumor cell

types (Fig. 3G).

As a proof of concept, the *in vitro* and *in vivo* experiment results demonstrate that msdP0188 plays an important role in promoting the genesis and metastasis of lung and breast tumors, and downregulating msdP0188 via CSSD4 represents a potential strategy for treating lung and breast cancers.

TERT is a putative RNA-dependent RNA polymerase (RdRP) involved in msdR/msdP generation

The presence of msdRs and msdPs implies the existence of RdRPs within human cells. We hypothesized that such proteins would exhibit conserved characteristics similar to those observed in extensively studied eukaryotic RdRPs, viral RdRPs, and/or DNA-dependent RNA polymerases (DdRPs). To explore this, we conducted a bioinformatic screening to identify potential human RdRPs (Fig. 4A).

Phylogenetic analysis indicated that eukaryotic RdRPs and the β' subunit of DdRPs bear a conserved, catalysis-contributing D*x*DGD motif, where *x* typically represents a bulky residue (*33*). We hence examined human proteins that harbor a D*x*DGD motif. Using a size cutoff of >500 AAs, we obtained 358 proteins from the human proteome that contain a D*x*DGD motif (Fig. 4A); among them, many are DdRPs or their subunits (data S13).

On the other hand, all known viral RdRPs lack a D*x*DGD motif; but they possess a conserved polymerase domain comprising three subdomains: *fingers*, *palm*, and *thumb*, with the *palm* subdomain being structurally conserved. We assessed the structural similarity of the *palm* subdomains from three viruses (SARS-CoV-2, human picobirnavirus strain Hy005102, and *Thosea asigna* virus) using TM-align (*34*). All TM-score (*35*) values are >0.5 (fig. S16), indicating a fundamental similarity at the fold level among the *palm* subdomains. Thus, it may be feasible to identify through structure comparison the human proteins that lack a D*x*DGD motif but bear a *palm*-like subdomain. Subsequently, we searched for human protein structural analogs to the three *palm* subdomains within the Protein Data Bank (PDB) (*36*) and the AlphaFold Protein Structure Database (AFDB) (*37, 38*), employing a TM-score cutoff of >0.5, and obtained 33 hits from PDB and 84 hits from AFDB (Fig. 4A, data S14, data S15).

One prominent hit from the *palm* search is the telomerase reverse transcriptase (TERT), which has been demonstrated to possess RdRP activity (*39-42*). Besides TERT, we also found a few endogenous retroviral proteins exhibiting a significant level of structural similarity to the three viral *palm* domains at the fold level (i.e., TM-score > 0.5) (data S14, data S15).

Considering the high telomerase activity in many tumor cells (*43, 44*) and the high msdP0188 levels in lung and breast cancer cells revealed in this work, we postulated that TERT may play a role in msdR genesis, and worked to assess the involvement of TERT in msdR/msdP production.

We attempted to knock out the TERT gene in A549 cells using the CRISPR/Cas9 technology, which however led to significant death of A549 cells, preventing us from conducting subsequent experiments. Alternatively, we employed siRNAs to knock down the TERT gene in A549 cells. Of the five siRNAs tested (fig. S17), siRNA-2 showed the best knockdown efficacy. As expected, the knockdown with siRNA-2

significantly reduced the protein levels of TERT and msdP0188 simultaneously (Fig. 4B), suggesting a role of TERT in msdP0188 production. We also treated A549 cells with TERT inhibitors BIBR1532 or zidovudine. BIBR1532 is a noncompetitive inhibitor of TERT (*45*), whereas zidovudine is a competitive retroviral inhibitor by binding to TERT to halt DNA extension (*46*). Following treatment with either inhibitor, Western blot assays were conducted to quantify the expression levels of msdP0188 and TERT; both inhibitors reduced the protein levels of msdP0188 without affecting TERT expression (Fig. 4C).

Collectively, these findings suggest TERT as a putative RdRP involved in the "msRNA → msdR → msdP" axis.

Discussion

An increasing body of evidence indicates the pervasive nature of translation, extending beyond canonical coding regions to include canonical untranslated regions, introns, and noncoding RNAs (*47-49*). Besides their roles in post-transcriptional gene expression regulation, pri-miRNAs containing open reading frames have been discovered to encode polypeptides (*50-52*). Nevertheless, existing studies on msRNAs are mainly focused on their regulatory functions like RNA silencing. Here, we discover a novel mechanism wherein the pairing between the msRNA 3'-end and RNA can give rise to new biological molecules. In this process, the msRNA and the paired RNA act as "primer" and "template," leading to the production of msdRs, which may consequently translate into msdPs. This discovery represents an important augmentation to the central dogma of molecular biology, enriching our comprehension of fundamental biological processes.

Even when adhering to the stringent 15-nt threshold, >1,000 new msdPs were predicted, with the number of msdRs being more substantial. These novel biological molecules introduce additional complexity to the biological system and represent new avenues for research. In our proof-of-concept study, we validated the presence and function of msdP0188 in cancer progression. We demonstrated that reducing its expression can mitigate tumor genesis and metastasis in lung and breast cancers. While practical constraints prevent us from experimentally verifying all predicted msdRs and msdPs or delving into all their biology, we believe these molecules hold significant potential for playing pivotal physiological and pathophysiological roles. This exciting discovery warrants further investigation in future research endeavors.

The molecular basis for msdR synthesis in human cells needs to be further determined. The synthesis likely utilizes enzymes that possess functions like RdRPs that are mainly found in prokaryotes (*15, 16, 53*). However, proteins with RdRP functions have also been reported in eukaryotes (*33, 39, 40, 54*). Our discovery of msdRs and msdPs provides indirect evidence of the presence of an endogenous biological pathway in eukaryotic cells capable of RNA-templated RNA synthesis, in line with previous findings (*39, 55*). Our experimental results support the involvement of TERT in this process, e.g., knockdown or inhibition of TERT led to decreased msdP0188 levels. Whether and how TERT functions as an RdRP at the molecular level remains to be elucidated. Further, we cannot rule out the possibility that other proteins may carry out similar functions and facilitate msdR generation, given the presence of numerous candidate proteins containing the conserved D*x*DGD motif or *palm*-like substructure such as the endogenous retroviral proteins suggested by bioinformatic search; it is also possible that different msdPs engage

different RdRPs. We also want to point out another possibility that human cells acquire RdRP capability after viral infection. Will viral RdRPs be utilized to derive msdRs? Will host cell msRNAs pair with viral RNA(s), or viral msRNAs pair with host RNAs to produce msdRs? If so, what are their biological implications for host-pathogen interactions? Further investigation is warranted to address these questions.

We would like to point out several scientific questions that have immediately arisen with our discovery. For instance, it is unclear whether msdRs possess a 5'-cap or 3'-ploy(A) tail and how they are stabilized if either feature is missing. It is also unknown whether the msdR/RNA double strands trigger innate immune responses and what forces separate them. Elucidating the molecular mechanisms underlying these questions is critical and warrants further research. While we do not yet have answers to these interesting questions, we give some discussions based on our understanding of relevant studies.

First, mRNAs are known to contain a 5'-cap and 3'-poly(A) tail that contribute to transcript stability, nuclear export, translation initiation, and mRNA stability. Other RNA modifications, such as N6-methyladenosine (m^6A), N1-methyladenosine (m^1A), and 5-hydroxymethylcytosine(hm^5C), are also important factors in determining mRNA half-life (*56, 57*). In the cytoplasm, ribosomal transit rate, RNA binding protein, and codon optimality are major determinants of mRNA stability. The decrease in ribosomal transit rate enhances the shortening of poly(A) and cap removal, ultimately shortening the mRNA half-life (*58, 59*). Moreover, secondary structures of 5' UTR or 3' UTR often contribute to mRNA stability, and RNA located in the cytoplasm sometimes appears to be more stable than RNA in the nucleus (*60, 61*). In tumor cells, the regulatory mechanism for RNA stability is more complex. For example, c-myc RNA that lacks poly(A) is more stable than poly$(A)^+$ c-myc RNA in HL-60 and HeLa cells (*62*). Perhaps due to the influence of these factors, some RNAs (e.g., pri-miRNAs) that do not have a 5'-cap or 3'-poly(A) tail can exist and translate into proteins. We thus speculate that msdRs likely do not have a polyA tail because they are more likely cytoplasmic, but some msdRs are stable enough to allow being translated into proteins. Again, these are only bold speculations that need solid follow-up studies to test.

Second, in eukaryotic cells, double-stranded RNA (dsRNA) is usually considered to be non-self infection and causes immune responses. For example, protein kinase R (PKR) induces immune responses by sensing viral dsRNAs upon infection. Growing evidence suggests that PKR can also be activated by endogenously expressed dsRNAs (*63, 64*), even without infection. As such, whether the endogenous dsRNAs formed by msdR/mRNA would be sensed by PKR and induce immune responses remains an important question to be answered.

Third, an intriguing question on the msdR translation is how the dsRNAs of msdR/mRNA separate to allow msdRs to get translated. It likely engages one or more RNA helicases. Some candidates to be investigated include MOV10, RNAhlicase A, and DDX3X. MOV10 is a 5'-to-3' ATP-dependent RNA helicase that unwinds duplex RNA (*65*). RNA helicase A as a DExH-box helicase can bind dsRNA N-terminal and unwind small regions within the dsRNA repeatedly until complete separation of the RNA duplex (*66*). DDX3X as a conserved DEAD-box RNA helicase can unwind RNA/RNA and RNA/DNA duplex, and compared to the *S. cerevisiae* ortholog Ded1p, DDX3X separates long RNA duplex faster (*67, 68*).

We also want to note that our bioinformatic pipeline does have some limitations, although it effectively predicts msdRs and msdPs. For example, IRES is used for msdR translation in the pipeline, yet it remains uncertain whether or not IRES is essential or sufficient for msdP translation. The method to determine potential IRESs and start codons on msdRs is somewhat empirical. Therefore, certain msdPs may not be accessible to our pipeline, highlighting the need for improvement in the future.

In sum, we report the central finding of msRNA-derived RNAs and proteins in human cells. Our work proposes a novel "msRNA → msdR → msdP" axis, representing an augmentation to the central dogma, and predicts many previously unknown RNAs and proteins that may have significant biological and pathological roles in human cells as well as in other biological systems.

References

1. V. N. Kim, J. Han, M. C. Siomi, Biogenesis of small RNAs in animals. *Nature reviews Molecular cell biology* **10**, 126-139 (2009).
2. L. He, G. J. Hannon, MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews genetics* **5**, 522-531 (2004).
3. M. C. Siomi, K. Sato, D. Pezic, A. A. Aravin, PIWI-interacting small RNAs: the vanguard of genome defence. *Nature reviews Molecular cell biology* **12**, 246-258 (2011).
4. Z. Su, B. Wilson, P. Kumar, A. Dutta, Noncanonical roles of tRNAs: tRNA fragments and beyond. *Annu. Rev. Genet.* **54**, 47-69 (2020).
5. M. Billmeier *et al.*, Mechanistic insights into non-coding Y RNA processing. *RNA Biol.* **19**, 468-480 (2022).
6. T. X. Lu, M. E. Rothenberg, MicroRNA. *J Allergy Clin Immunol* **141**, 1202-1207 (2018).
7. 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).
8. V. Ambros, H. R. Horvitz, Heterochronic mutants of the nematode Caenorhabditis elegans. *Science* **226**, 409-416 (1984).
9. 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).
10. T. Kim, C. M. Croce, MicroRNA and ER stress in cancer. *Semin Cancer Biol* **75**, 3-14 (2021).
11. T. J. Lee *et al.*, Strategies to Modulate MicroRNA Functions for the Treatment of Cancer or Organ Injury. *Pharmacol Rev* **72**, 639-667 (2020).
12. 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).
13. S. Vasudevan, Y. Tong, J. A. Steitz, Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318**, 1931-1934 (2007).
14. 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).
15. 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).
16. Y. Gao *et al.*, Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science*

368, 779-782 (2020).
17. 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).
18. 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).
19. P. Weber *et al.*, Therapy-related transcriptional subtypes in matched primary and recurrent head and neck cancer. *Clin Cancer Res*, (2021).
20. J. Zhao *et al.*, IRESbase: A Comprehensive Database of Experimentally Validated Internal Ribosome Entry Sites. *Genomics Proteomics Bioinformatics* **18**, 129-139 (2020).
21. A. Kozomara, M. Birgaoanu, S. Griffiths-Jones, miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155-D162 (2019).
22. B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, C. B. Burge, Prediction of mammalian microRNA targets. *Cell* **115**, 787-798 (2003).
23. J. Brennecke, A. Stark, R. B. Russell, S. M. Cohen, Principles of microRNA-target recognition. *PLoS Biol.* **3**, e85 (2005).
24. B. Fromm *et al.*, A Uniform System for the Annotation of Vertebrate microRNA Genes and the Evolution of the Human microRNAome. *Annu. Rev. Genet.* **49**, 213-242 (2015).
25. B. Fromm *et al.*, MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.* **48**, D132-D141 (2020).
26. B. Fromm *et al.*, MirGeneDB 2.1: toward a complete sampling of all major animal phyla. *Nucleic Acids Res.* **50**, D204-D210 (2022).
27. J. C. Venter *et al.*, The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
28. 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).
29. 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).
30. B. Wen, B. Zhang, PepQuery2 democratizes public MS proteomics data for rapid peptide searching. *Nat Commun* **14**, 2213 (2023).
31. J. Meng *et al.*, Derepression of co-silenced tumor suppressor genes by nanoparticle-loaded circular ssDNA reduces tumor malignancy. *Sci Transl Med* **10**, (2018).
32. H. Wu *et al.*, G-quadruplex-enhanced circular single-stranded DNA (G4-CSSD) adsorption of miRNA to inhibit colon cancer progression. *Cancer Med* **12**, 9774-9787 (2023).
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. Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302-2309 (2005).
35. Y. Zhang, J. Skolnick, Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710 (2004).
36. H. M. Berman *et al.*, The Protein Data Bank. *Acta Crystallographica Section D Biological*

Crystallography **58**, 899-907 (2002).
37. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
38. K. Tunyasuvunakool *et al.*, Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590-596 (2021).
39. Y. Maida *et al.*, An RNA-dependent RNA polymerase formed by TERT and the RMRP RNA. *Nature* **461**, 230-235 (2009).
40. 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).
41. M. Takahashi *et al.*, Eribulin penetrates brain tumor tissue and prolongs survival of mice harboring intracerebral glioblastoma xenografts. *Cancer Sci.* **110**, 2247-2257 (2019).
42. M. Machitani *et al.*, Maintenance of R-loop structures by phosphorylated hTERT preserves genome integrity. *Nat. Cell Biol.* **26**, 932-945 (2024).
43. X. Zhang, V. Mar, W. Zhou, L. Harrington, M. O. Robinson, Telomere shortening and apoptosis in telomerase-inhibited human tumor cells. *Genes Dev* **13**, 2388-2399 (1999).
44. S. Mocellin *et al.*, Telomerase reverse transcriptase locus polymorphisms and cancer risk: a field synopsis and meta-analysis. *J. Natl. Cancer Inst.* **104**, 840-854 (2012).
45. E. Pascolo *et al.*, Mechanism of human telomerase inhibition by BIBR1532, a synthetic, non-nucleosidic drug candidate. *J. Biol. Chem.* **277**, 15566-15572 (2002).
46. Y. Quan, L. Rong, C. Liang, M. A. Wainberg, Reverse transcriptase inhibitors can selectively block the synthesis of differently sized viral DNA transcripts in cells acutely infected with human immunodeficiency virus type 1. *J. Virol.* **73**, 6700-6707 (1999).
47. J. S. Kesner *et al.*, Noncoding translation mitigation. *Nature* **617**, 395-402 (2023).
48. 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).
49. N. T. Ingolia *et al.*, Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**, 1365-1379 (2014).
50. D. Lauressergues *et al.*, Primary transcripts of microRNAs encode regulatory peptides. *Nature* **520**, 90-93 (2015).
51. H. Zhou *et al.*, A peptide encoded by pri-miRNA-31 represses autoimmunity by promoting T(reg) differentiation. *EMBO Rep* **23**, e53475 (2022).
52. A. Montigny *et al.*, Drosophila primary microRNA-8 encodes a microRNA-encoded peptide acting in parallel of miR-8. *Genome Biol* **22**, 118 (2021).
53. 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).
54. P. Ahlquist, RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science* **296**, 1270-1273 (2002).
55. P. Kapranov *et al.*, New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism. *Nature* **466**, 642-646 (2010).
56. I. A. Roundtree, M. E. Evans, T. Pan, C. He, Dynamic RNA Modifications in Gene Expression

Regulation. *Cell* **169**, 1187-1200 (2017).
57. A. Radhakrishnan *et al.*, The DEAD-Box Protein Dhh1p Couples mRNA Decay and Translation by Monitoring Codon Optimality. *Cell* **167**, 122-132 e129 (2016).
58. V. Presnyak *et al.*, Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111-1124 (2015).
59. H. Bae, J. Coller, Codon optimality-mediated mRNA degradation: Linking translational elongation to mRNA stability. *Mol Cell* **82**, 1467-1476 (2022).
60. K. Shi, T. Liu, H. Fu, W. Li, X. Zheng, Genome-wide analysis of lncRNA stability in human. *PLoS Comput Biol* **17**, e1008918 (2021).
61. M. B. Clark *et al.*, Genome-wide analysis of long noncoding RNA stability. *Genome Res* **22**, 885-898 (2012).
62. S. G. Swartwout, H. Preisler, W. D. Guan, A. J. Kinniburgh, Relatively stable population of c-myc RNA that lacks long poly(A). *Mol Cell Biol* **7**, 2052-2058 (1987).
63. Y. Kim *et al.*, PKR Senses Nuclear and Mitochondrial Signals by Interacting with Endogenous Double-Stranded RNAs. *Mol Cell* **71**, 1051-1063 e1056 (2018).
64. Y. Kim *et al.*, PKR is activated by cellular dsRNAs during mitosis and acts as a mitotic regulator. *Genes Dev* **28**, 1310-1322 (2014).
65. L. H. Gregersen *et al.*, MOV10 Is a 5' to 3' RNA helicase contributing to UPF1 mRNA target degradation by translocation along 3' UTRs. *Mol Cell* **54**, 573-585 (2014).
66. H. R. Koh, L. Xing, L. Kleiman, S. Myong, Repetitive RNA unwinding by RNA helicase A facilitates RNA annealing. *Nucleic Acids Res* **42**, 8556-8564 (2014).
67. D. Sharma, A. A. Putnam, E. Jankowsky, Biochemical Differences and Similarities between the DEAD-Box Helicase Orthologs DDX3X and Ded1p. *J Mol Biol* **429**, 3730-3742 (2017).
68. H. Song, X. Ji, The mechanism of RNA duplex recognition and unwinding by DEAD-box helicase DDX3X. *Nat Commun* **10**, 3085 (2019).
69. Y. Maida, M. Yasukawa, K. Masutomi, De novo RNA synthesis by RNA-dependent RNA polymerase activity of telomerase reverse transcriptase. *Molecular and Cellular Biology*, (2016).
70. N. Okamoto *et al.*, Maintenance of tumor initiating cells of defined genetic composition by nucleostemin. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 20388-20393 (2011).

Acknowledgments:

We thank the Advanced Research Computing at the University of Michigan for providing computational resources and services to support the computational and bioinformatic research in this publication.

Funding:

National Natural Science Foundation of China grant 81972746 (S.C.); Government Funded Clinical Medicine Talents Training Project of Hebei Province grant [2020] 397 (G.Z.); Natural Science Foundation of Hebei Province grant H2021209026 (G.Z.); Returned Overseas Students of Hebei Province grant C20210340 (G.Z.)

Author contributions:

X.H., H.W., J.X., Y.E.C., and S.C. conceived the project. Y.E.C. and S.C. supervised and administrated the project. X.H. conducted all computational and bioinformatic research. H.W., W.L., M.Z., X.K., R.Z., K.L., X.L., Y.L., M.W., L.D., and S.C. conducted transwell, wound healing, and SEM experiments. X.X. and H.Z. performed the MS experiment. Z.L., W.Z., B.C., and S.C. conducted the animal experiment. W.Z. and Z.L. conducted the IHC experiment. H.W., W.Z., W.L., R.Z., and S.C. carried out PCR and Western blot experiments. H.W., G.Z., and S.C. conducted the silver staining experiment. H.W., W.L., and Z.L. designed msdP-specific antigens and produced antibodies. X.H., H.W., W.Z., Z.L., Y.H., and Y.P., and S.C. performed formal analysis. X.H., M.Z., M.W., G.Z., Y.H., J.X., Y.E.C., and S.C. provided resources. X.H. and S.C. wrote the initial manuscript. X.H., J.Z., J.X., and Y.E.C., and S.C. edited the manuscript. All authors participated in result discussions and contributed to the final manuscript.

Competing interests:

X.H., H.W., J.X., Y.E.C., and S.C. are co-inventors of a patent application that encompasses the predicted msdRs and msdPs described in this study. The other authors declare no competing interests.

Data and code availability

The miRBase miRNA sequences were retrieved from https://www.mirbase.org/ftp.shtml (accessed on December 20, 2020). Experimentally determined IRES sequences were obtained from http://reprod.njmu.edu.cn/cgi-bin/iresbase/index.php (accessed on January 2, 2021). The NCBI RefSeq RNA database was acquired from https://ftp.ncbi.nlm.nih.gov/blast/db (accessed on December 20, 2020). The NCBI nr protein database was obtained from the same source (accessed on January 2, 2021). The BLAST suite (version 2.11.0) was downloaded from https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+. The human proteome was retrieved from https://www.uniprot.org/proteomes/UP000005640 (accessed on May 19, 2023). The Protein Data Bank (PDB) was accessed from https://files.rcsb.org/pub/pdb/data/structures/divided/mmCIF (accessed on May 19, 2023). The AlphaFold Protein Structure Database (AFDB) for *Homo sapiens* was accessed from https://alphafold.ebi.ac.uk/download (accessed on May 19, 2023). The PepQuery2 program (version 2.0.2)

was downloaded from http://pepquery.org/download.html (accessed on December 7, 2023). The PepQueryDB database is available at http://pepquery.org/index.html (accessed by PepQuery2 with internet connection). The mirGeneDB database is available at https://mirgenedb.org (v2.1, accessed on August 7, 2024). In case these databases or software have been updated, the specific versions employed in this study can be provided upon request. All scripts for the prediction and analysis of msdRs and msdPs, along with supplementary data and other related source data, are made publicly available at https://doi.org/10.5281/zenodo.7826531.

Supplementary Materials:

Methods

Figs. S1 to S17

Tables S1 to S13

Data S1 to S15

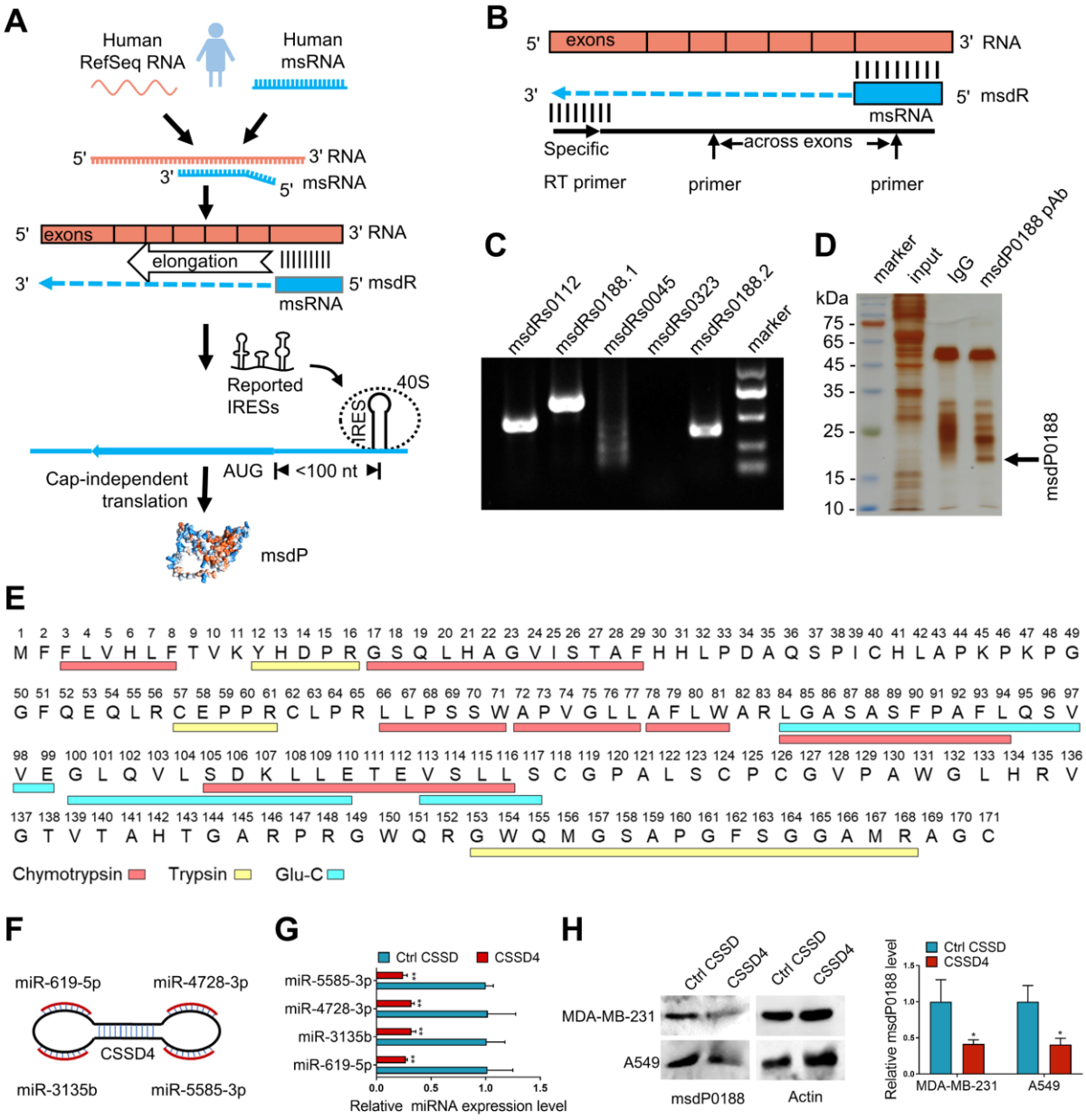

Fig. 1. Prediction and validation of microsized RNA (msRNA)-derived RNAs (msdRs) and proteins (msdPs). (**A**) Bioinformatic pipeline for msdR/msdP prediction. (**B**) Schematic representation for designing specific reverse transcription (RT) primers and exon-crossing PCR primers for msdR. (**C**) Analysis of msdR fragments amplified by specific primers spanning exons through agarose gel electrophoresis. (**D**) Silver staining of proteins captured by the anti-msdP0188 polyclonal antibody (pAb) following SDS-PAGE electrophoresis separation. The msdP0188 band was marked. (**E**) Confirmation of msdP0188 through mass spectrometry analysis after digestion with trypsin, chymotrypsin, and Clu-C. (**F**) Schematic illustration of CSSD4, capable of adsorbing four msRNAs (*miR-619-5p*, *miR-5585-3p*, *miR-4728-3p*, and *miR-3135b*) simultaneously, bound to *AP1M1* mRNA with a length of ≥12 nt. (**G**) Expression levels of *miR-619-5p*, *miR-5585-3p*, *miR-4728-3p*, and *miR-3135b* in A549 cells after CSSD transfection. (**H**) Expression levels

of msdP0188 in A549 and MDA-MB-231 cells after CSSD transfection. In all figures, error bars represent the standard deviation of n = 3 independent biological replicates. *$P < 0.05$; **$P < 0.01$; Student's *t*-test.

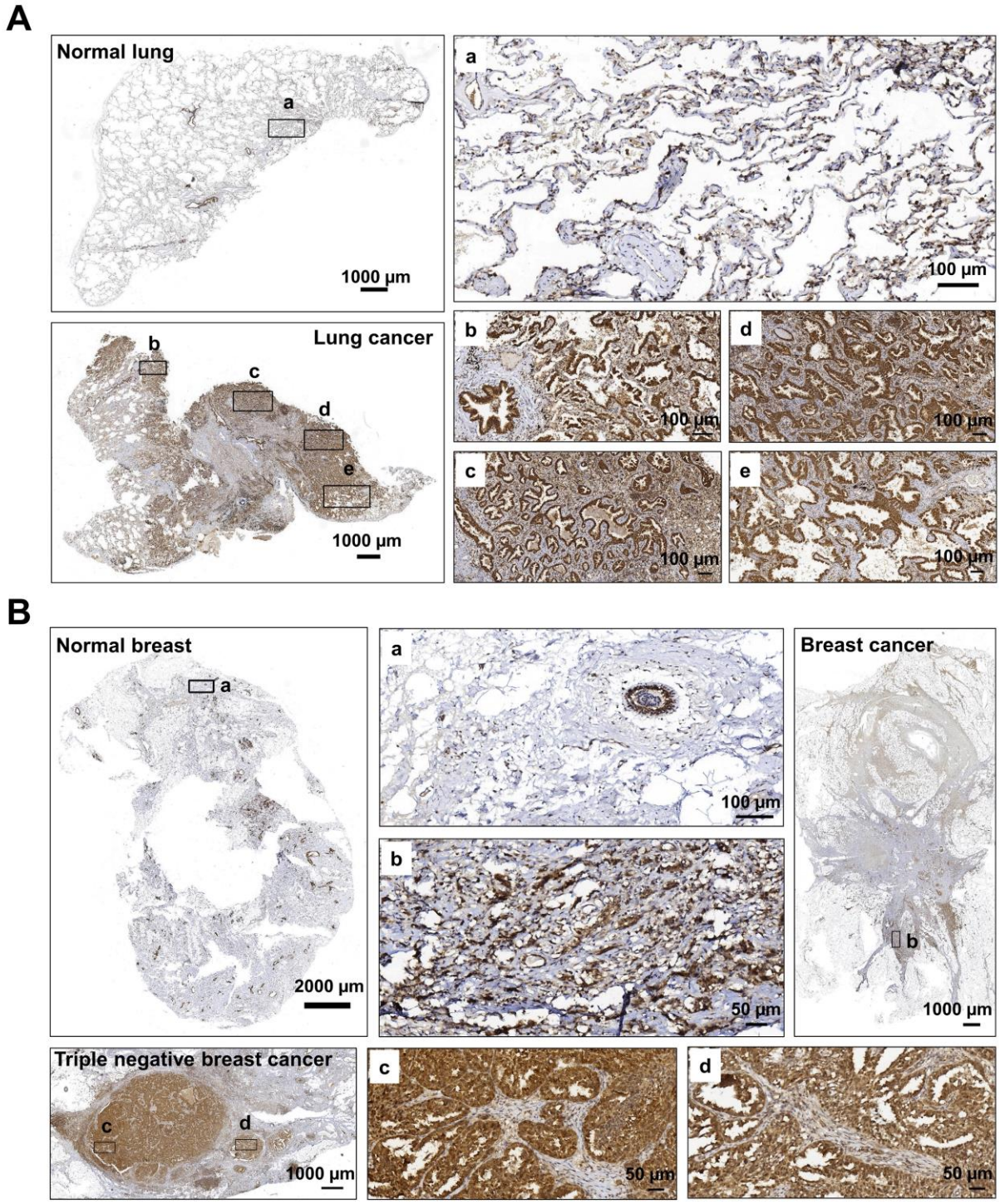

Fig. 2. High expression of msdP0188 in human lung and breast cancers. (**A**) Representative immunohistochemical staining of msdP0188 in normal lung tissues (n = 9) and lung cancer tissues (n = 15) using the anti-msdP0188 polyclonal antibody (1:200). The left side displays the uncropped full immunohistochemical section, with enlarged typical staining areas shown on the right side. (**B**) Representative immunohistochemical staining of msdP0188 in normal breast tissue (n = 8), general breast

cancer tissues (n = 10), and triple-negative breast cancer tissues (n = 12) using the anti-msdP0188 polyclonal antibody (1:200). Zoomed-in images depict typical staining areas from the full sections.

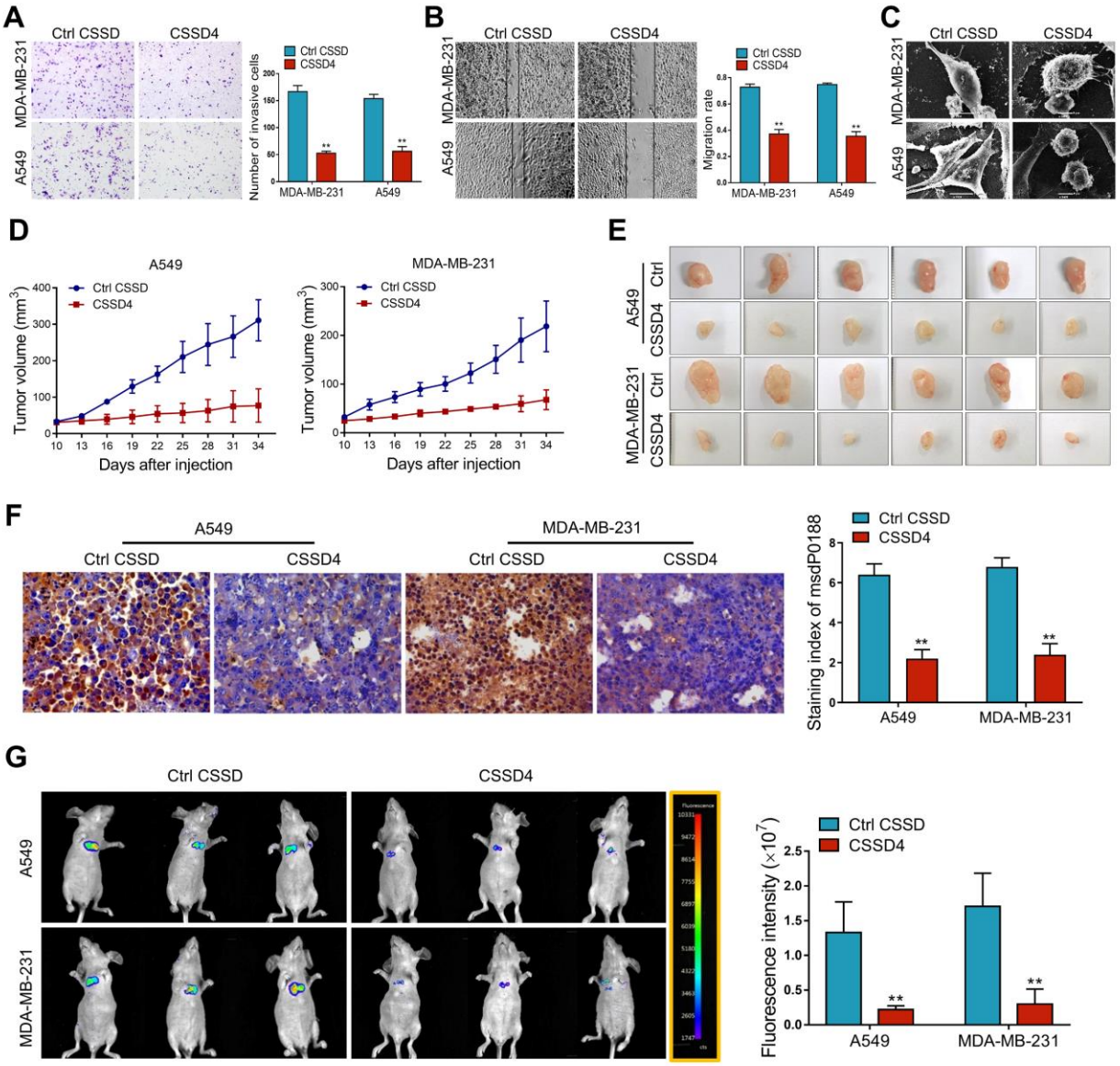

Fig. 3. CSSD4 inhibits lung and breast cancer cell progression both *in vitro* and *in vivo*. (**A**) Effects of CSSD on A549 and MDA-MB-231 cell invasion. (**B**) Effects of CSSD on A549 and MDA-MB-231 cell migration. (**C**) Changes in cell phenotype detected by scanning electron microscopy after CSSD transfection. (**D**) Solid tumors formed by A549 and MDA-MB-231 cells in nude mice following CSSD treatment. (**E**) Tumor volume as depicted in Fig. 3D. (**F**) Intensity of msdP0188 immunohistochemical staining in solid tumors. (**G**) Lung metastasis of A549 and MDA-MB-231 cells after CSSD treatment. In Fig. 3D, error bars represent standard deviations of n = 6 independent biological replicates; n = 3 otherwise. $^{*}P < 0.05$; $^{**}P < 0.01$; Student's *t*-test.

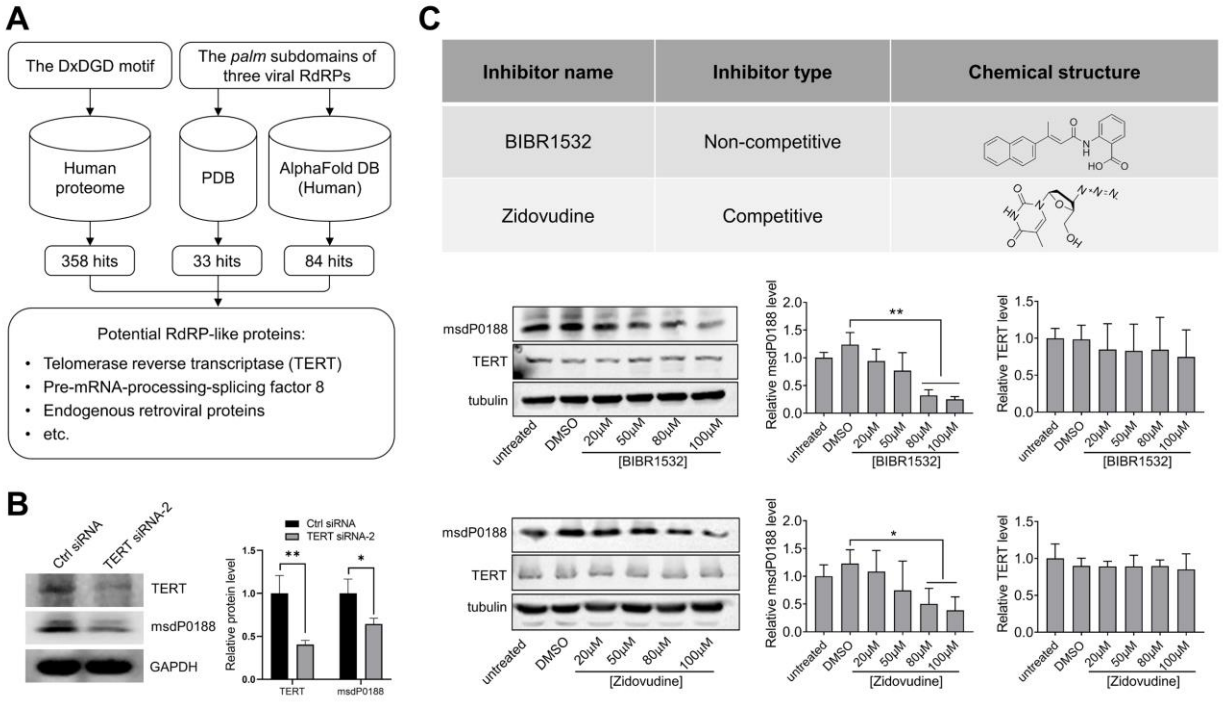

Fig. 4. Human telomerase reverse transcriptase (TERT) is implicated in msdR generation. (**A**) Bioinformatic prediction of RdRP-like proteins deriving msdRs in human cells. (**B**) Relative expression levels of TERT and msdP0188 after TERT knockdown using siRNA-2 and Ctrl siRNA. (**C**) Relative levels of msdP0188 and TERT expression in A549 cell lines following treatment with TERT inhibitors BIBR1532 and zidovudine at various concentrations. Untreated and DMSO-treated A549 cell lines are used as controls. In all figures, error bars represent standard deviations of n = 3 independent biological replicates. *$P < 0.05$; **$P < 0.01$; Student's *t*-test in Fig. 4B; one-way ANOVA and Tukey's posthoc test in Fig. 4C.

Supplementary Materials for

Discovery of microsized RNA-derived RNAs and proteins in human cells

Xiaoqiang Huang[1,#,*], Haidong Wu[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[2], Xinze Lv[2], Yue Lv[2], Jingyu Bao[2], Meimei Wang[5], Yanan Xiong[5], Lijie Dong[4], Jifeng Zhang[1], Guangling Zhang[5,*], Yuanjing Hu[12,*], Jie Xu[1,*], Y. Eugene Chen[1,*], Shuang Chen[1,2,*]

*Correspondence to: shuanche@umich.edu (S.C., lead contact), xiaoqiah@umich.edu (X.H.), haidong123abc@sina.com (H.W.), zhanggl@ncst.edu.cn (G.Z.), julianna_hu@163.com (Y.H.), jiex@umich.edu (J.X.), echenum@umich.edu (Y.E.C.)

The PDF file includes:

Methods
Figs. S1 to S17
Tables S1 to S13
References

Other Supplementary Materials for this manuscript include the following:

Data S1 to S15

Methods

Prediction of msdRs and msdPs

We retrieved a total of 2,656 human miRNAs from miRbase (*https://www.mirbase.org/*; accessed on December 20, 2020), the largest repository for miRNA sequences and annotations. After discarding 24 redundant miRNAs (table S1), the remaining 2,632 unique miRNAs were used as example msRNAs to derive msdRs and msdPs (Fig. 1).

The 3'-end, such as a 15-nt fragment, of each msRNA (U → T) was queried against the NCBI RefSeq RNA database (*17*) using BLASTN (*18*) to identify complementary RNA templates for generating msdRs:

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

Given that the queries are short, the "*-task blastn-short*" option was utilized, employing default parameters "*-gapopen 5 -gapextend 2 -reward 1 -penalty -3*" for "*blastn-short*" tasks. To ensure alignment without gaps, the "*-ungapped*" option was applied, while restricting the search to reverse matches was achieved with "*-strand minus*". Further restriction to human RNAs was enacted via the "*-taxids 9606*" parameter.

We enforced strict reverse base pairing at the last bases of the 3'-end of msRNA with the RNA target through the following conditions:

1) The "-ungapped" option ensures no gaps in the pairing between msRNA and RNA.
2) We elaborately adjusted e-value thresholds, setting them at specific values such as 1, 5, 20, 100, 500, and 1,000 for queries of 15, 14, 13, 12, 11, and 10 nt, respectively. For instance, when utilizing the parameters "*-gapopen 5 -gapextend 2 -reward 1 -penalty -3 -ungapped*," an exact, ungapped 15-nt match yields a calculated e-value of 0.941241. Therefore, we employed a slightly higher e-value cutoff of "*-evalue 1*" for 15-nt queries. Hits with mismatches for the 15-nt query result in e-values larger than 1, hence such hits are excluded.
3) We mandated that the *Hsp_identity* (found in the *out_15nt.xml* file) be exactly equal to the query length (e.g., 15-nt) to ensure no mispairing.
4) We stipulated that "*-strand minus*" and "*Hsp_hit-frame = −1*" and "*Hsp_query-frame = 1*" (found in the *out_15nt.xml* file) to ensure that the msRNA is in the 5'→3' direction while the paired fragment of RNA hit in the 3'→5' direction.

With these BLASTN parameter settings, a sufficiently high cutoff of "*-max_target_seqs 10000*" was established to generate RNA hits; however, this threshold was not surpassed for queries ranging from 10 to 15 nt.

For clarity, the following discussion employs the 15-nt match as an example. Relevant information for a hit, including the GenInfo Identifier (GI) number, NCBI Reference Sequence number, start and end

positions of the match, and gene symbol, was documented.

The sequence of a hit RNA was retrieved by utilizing its GI number:

blastdbcmd -entry gi -db db_refseq_rna

The msdR sequence was produced through reverse transcription of the RNA hit from the end-matching position to its 5'-end. The number of msdRs derived from each msRNA varies substantially. For instance, certain msRNAs (e.g., *hsa-miR-99b-5p*, *hsa-miR-9985*, and *hsa-miR-9903*, to name a few) do not produce any msdR, while *hsa-miR-619-5p* alone generates 3,421 msdRs. In total, 12,953 msdRs were obtained from all 2,632 human msRNAs (table S1). Of these, 11,121 unique msdRs were utilized to make a database:

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

The unique msdRs were translated into proteins if the following criteria were met: the msdR possesses a predicted IRES (*19*), and a start codon is present within 100 nt downstream of the IRES. The first criterion was verified by comparing msdRs with the experimental IRESs sourced from IRESbase (*20*), which encompasses 554 viral IRESs, 691 human IRESs, and 83 IRESs from other eukaryotes. Due to the considerable variation in sequence length, the experimental IRESs were searched against the msdR database using distinct BLASTN parameters:

If an IRES has ≤15 nt:

blastn -db db_uniq_msdR -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_msdR -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_msdR -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 "*-strand plus*" option confines the search to forward matches. For short IRESs (≤15 nt), a high gap opening penalty "*-gapopen 5*" and the parameter *"-ungapped*" were applied to prohibit gaps; "*-penalty -3*" was employed to discourage mismatches. For IRES spanning (15, 50) nt, penalty scores were adjusted to permit more mismatches and gap openings. For IRES >50 nt, default parameters for a standard "*blastn*" task were utilized. In all scenarios, large cutoffs for e-value ("*-evalue 1e5*") and the number of targets ("*-max_target_seqs 2500000*", exceeding the number of msdRs) were utilized.

An IRES site was considered to exist if *coverage* ≥ 1 and *aligned identity* ≥ 1 for IRES queries of ≤15 nt,

or if *coverage* ≥ 0.7 and *aligned identity* ≥ 0.75 for queries >15 nt. Here, *coverage* and *aligned identity* were defined as the ratio of *aligned length* to *query length* and *identity* to *aligned length*, respectively. These metrics (*aligned length* and *identity*) were extracted from the BLASTN output (in the *.xml* file).

The msdRs were subsequently translated according to the second criterion. One msdR may encompass multiple IRESs, and there may be multiple start codons following an IRES. All potential scenarios were taken into account during the translation process from msdR to msdP(s). In total, 5,369 msdPs with more than 10 AAs were generated, out of which 1,239 are unique. We conducted a BLASTP (*18*) search against the NCBI nr protein database to ascertain whether these msdPs have counterparts in the human protein database with significant similarity using distinct parameters:

If a msdP has ≤15 AAs:

blastp -db nr -query msdP_le15.fa -out msdP_le15.xml -outfmt 5 -evalue 1 -task blastp-short -taxids 9606

Else if a msdP has <30 AAs:

blastp -db nr -query msdP_lt30.fa -out msdP_lt30.xml -outfmt 5 -evalue 1e-4 -task blastp-short -taxids 9606

Otherwise:

blastp -db nr -query msdP_ge30.fa -out msdP_ge30.xml -outfmt 5 -evalue 1e-4 -task blastp -taxids 9606

In this context, "*significant similarity*" is characterized by *coverage* ≥ 1 and *aligned identity* ≥ 0.9 for msdPs with ≤15 AAs, or *coverage* ≥ 0.9 and *aligned identity* ≥ 0.9 otherwise (*coverage* and *aligned identity* are parameters defined as previously described).

The same procedures were carried out for matching thresholds of 10–14 nt.

Estimation of the potential false discovery rate (FDR) for msdR prediction

To evaluate the reliability of our bioinformatic pipeline, we estimated the potential false discovery rates (FDRs) for msdR prediction for different matching lengths. FDR was calculated as the ratio of false positive (FP) to predicted positive (PP), in which the number of msdRs predicted using database-documented msRNAs was regarded as PP whereas the number of msdRs predicted using computer-simulated msRNAs as FP. We carried out a series of computational simulations to estimate FDRs and their bounds.

As we predicted msdRs and msdPs using miRBase miRNAs, it was reasonable to take the number of msdRs predicted with unique human miRNAs (n = 2,632) in miRBase as PP. Inspired by previous studies utilizing shuffled miRNAs to estimate the FDRs for miRNA target prediction (*22, 23*), we calculated FP as the mean number of predicted msdRs using n = 9 sets of msRNA artifacts (n = 2,632 of each set) by shuffling the 3'-end *l* (*l* = 10–15) nt of each parental miRBase miRNA (fig. S1A). The 3'-end nucleotide composition was held constant during shuffling, and therefore the msRNA artifacts generated in this way would exhibit

identical binding affinities compared to the parental miRNAs when bound to RNAs with their 3'-end *l* nt.

We further explored different ways to generate msRNA artifacts to estimate FDR ranges. As expected, using completely randomly generated msRNA artifacts (fig. S1A) yielded lower FDR for each matching length (table S4). Since the 3'-end shuffling and random generation methods do not control for di- or tri-nucleotide composition, we generated another two sets of msRNA artifacts by shifting some nucleotides (s = 1, 2, …, 9) of the 3'-ends of miRBase miRNAs from either the 3'→5' direction (Rshift) or the 5'→3' direction (Lshift) (fig. S1A) and as expected, both cases, which were less-random compared to shuffling and random-generating, dramatically increased the mean numbers of predicted msdRs, thereby resulting in higher FDRs for each matching length (table S5, table S6).

In addition, we examined if our bioinformatic pipeline was able to tell the database-documented miRNAs apart from their reverse complementary strands or passenger miRNAs. In the former simulation, we utilized the miRBase human miRNAs to generate reverse complements (fig. S1A), while for the latter ex, we took human mature and passenger miRNAs from mirGeneDB (*24-26*). Interestingly, both the reverse complements and passenger miRNAs produced more predicted msdRs than the parental miRNAs (table S7, table S8).

Nomenclature of msdR and msdP identifiers and names

We have devised a nomenclature for assigning identifiers to msdRs and msdPs, reflecting the nature of their derivation from the elongation of msRNAs along template RNAs. The msdR identifier is structured as "msdR.{sRNA name}.{GenInfo Identifier number of the RNA (gene symbol of the RNA)}.{the end matching position on RNA}", where "msdR" is a prefix. The msdP identifiers can be similarly established by extending the msdR identifier with {start codon position on msdR}. However, we observed that a msdP can be fully delineated by the template RNA alongside the position on RNA corresponding to the start-codon position on msdR. Therefore, a msdP identifier is formulated as "msdP.{GenInfo Identifier number of the RNA (gene symbol of the RNA)}.{the msdR start codon-paired position on RNA}", where "msdP" is a prefix. This method largely avoids using multiple msdR-based msdP identifiers for the same msdP encoded by msdRs that originated from different msRNAs bound to the same template RNA.

Despite the meaningfulness, the above-defined identifiers may still not be unique for some msdRs/msdP due to RNA variants from splicing. To facilitate data presentation, we also used sequential names to describe certain msdRs and msdPs (table S9, table S10, data S4).

Estimation of the presence of msdPs via proteomics database search

The sequences of 1,193 new msdPs were used to create a FASTA database, which was then compared to mass spectrometry (MS) proteomics data obtained from four human tumor cell lines (refer to the "*Mass Spectrometry*" section) using Proteome Discoverer 2.1 (Thermo Scientific) with the Sequest-HT search engine. The S/N (signal-to-noise ratio) threshold was set to 1.5, the maximum missed cleavage sites to 2,

the minimum peptide length to 6, the maximum peptide length to 144, the precursor mass tolerance to 10 ppm, and the fragment mass tolerance to 0.6 Da, respectively. Default parameters were used for other settings. The comparison suggested the potential presence of 941 msdPs, identified through a scoring metric that considers protein sequence coverage, the number of peptide-spectrum matches (PSMs), and the Sequest-HT score (data S8).

We also used the standalone PepQuery2 (*30*) to search for peptides within the 1,193 msdPs against the PepQueryDB (*30*) database. PepQuery serves as a targeted peptide search engine for the identification or validation of both known and novel peptides of interest in available MS-based proteomics datasets. PepQueryDB constitutes a comprehensive database of indexed MS/MS spectra, encompassing 48 datasets, 25,405 raw MS files, and a total of 1,012,696,284 MS/MS spectra. Following the PepQuery2 instructions, we executed the command "*java -jar pepquery-2.0.2.jar -b all -db gencode:human -hc -o $outdir/ -i $prot -fast -cpu $ncpu -n 1000 -t protein*", where *$prot* denotes the query msdP protein sequence, *$outdir* specifies the output path, and *$ncpu* represents the number of CPUs employed for computation, respectively. The *"-t protein*" option specifies the input sequence as a digestible protein. The "*-n 1000*" parameter generates 1,000 random peptides to estimate the statistical significance of a PSM hit. The "*-fast*" option specifies the fast mode for searching. Confident PSMs were identified for 1,095 peptides originating from 555 msdPs (data S9). Among them, peptides from 459 msdPs were also detected by the Sequest-HT search. Two peptide fragments that exhibit confident PSMs were discovered for msdP0188: LLETEVSLLSCGPALSCPCGVPAWGLHR and CLPRLLPSSWAPVGLLAFLWAR; notably, the latter encompasses the peptide LLPSSWAPVGLLAFLW determined by our MS analysis for verifying msdP0188. Similarly, two peptides featuring confident PSMs were identified for msdP0112: HPVFSLK and MDFSTKHPVFSLK.

Prediction of msdR-deriving RdRP in human cells

We conducted a search of the human proteome (https://www.uniprot.org/proteomes/UP000005640, accessed on May 19, 2023), which encompasses 82,492 proteins. Within this dataset, we considered proteins containing a linear D*x*DGD motif (*x* can be any AA) and consisting of >500 AAs, as we reason that human RdRP-like proteins tend to have a relatively larger size. Subsequently, we manually analyzed these proteins, considering factors such as their function, subcellular location, structural information derived from experimental studies or predictions by AlphaFold2, as well as other features annotated in UniProt (data S12).

On the other hand, we posited that it supplements the D*x*DGD motif-searching method by leveraging the *palm* subdomains of three viruses (SARS-CoV-2, human picobirnavirus strain Hy005102, and *Thosea asigna* virus) to potentially identify human RdRP-like proteins that lack the linear, signature D*x*DGD motif. We employed TM-align to explore structural analogs of the three viral *palm* subdomains across the entire Protein Data Bank (PDB) (*36*) database (204,826 structures; accessed on May 19, 2023) and the AlphaFold Protein Structure Database (*37, 38*) for *Homo sapiens* (20,504 nonredundant models; accessed on May 19, 2023), utilizing a cutoff of TM-score > 0.5 (data S13, data S14).

Cell line and transfection

The 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 Sciences and authenticated through short tandem repeat analysis. All cell lines tested negative for mycoplasma contamination. A549 cells were cultured in minimum Eagle's medium with 10% heat-inactivated fetal calf serum (A5669801, Gibco, USA) and supplemented with penicillin-streptomycin (C0222, Beyotime, China). MDA-MB-231 cells were cultured in Dulbecco's modified Eagle's medium with 10% heat-inactivated fetal calf serum (A5669801, Gibco, USA) and supplemented with penicillin-streptomycin (C0222, Beyotime, China). Cells were maintained at 37 °C with 5% CO_2.

For the knockdown of msdP0188, 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 interfering sequences are as follows: 5'-GGAGTCATCTCCACTGCATTT-3' for shRNA-1, 5'-GCGTCTACTTCCTTCGTCTTG-3' for shRNA-2, 5'-GCTAGGTGCGTCTGCTTCATT-3' for shRNA-3, and 5'-CCTAAGGTTAAGTCGCCCTCG-3' for control shRNA.

For the overexpression of msdP0188, the msdP0188-encoding sequence was cloned into PiggyBac Dual promoter to generate msdP0188-PiggyBac Dual promoter (GFP-Puro) plasmid by Genewiz Biotechnology (Beijing, China). Following the instructions provided by Neofect (Neofect Biotech Co., Ltd), the msdP0188-PiggyBac Dual promoter (GFP-Puro) plasmid and PiggyBac transposase (GFP-Puro) were transfected into A549 cells. After 48 hours, puromycin (1 μg/ml) was added to the culture for positive clone selection over the next 7 days. Confirmation of msdP0188-overexpressed A549 cell lines was conducted through Western blot analysis.

Transwell assay

Transwell chambers (Corning, USA) were placed in a 24-well plate and 300 μl pre-warmed serum-free medium was added to the upper chamber. After incubating for 15 minutes at room temperature, the medium was discarded. The treated cells were then trypsinized into single cells and resuspended in a serum-free medium. After counting, the cells were seeded into the upper chamber at a density of 10^5 cells/well. In the lower chamber, 500 μl of medium containing 10% fetal bovine serum medium was added. All cells were cultured in an incubator with CO_2 at 37 °C for 24 hours. After the removal of cells from the upper chamber, the remaining cells were fixed with 4% formaldehyde for 30 minutes. Subsequently, the cells were washed with phosphate-buffered saline (PBS) and stained with 0.1% crystal violet. The invaded cells were then observed under a microscope at 200× magnification. All tests were done in triplicate.

Wound healing assay

Cells were digested and seeded in 24-well plates at a density of 2×10^5 cells/well. Upon cell density reaching

90%, straight scratches were made using a 100 μl pipette tip. The culture medium was then discarded, and cells were washed three times with PBS to eliminate debris. Subsequently, cells were cultured in serum-free medium in an incubator with 5% CO_2 at 37 °C. After 48 hours, photographs were captured, and cell migration rates were calculated using the formula: (*initial wound distance - wound distance at 48 hours*) / *initial wound distance*. All tests were done in triplicate.

Scanning electron microscope imaging

The cells were first fixed in pre-chilled 2.5% glutaraldehyde at 4 °C for 2 hours, followed by dehydration with a gradually increasing concentration of ethanol (30%, 50%, 70%, 80%, 90%, 95%, and 100% v/v) for 10 minutes at each concentration. Following this, cells underwent treatment with a gradually increasing concentration of tert-butanol (30%, 50%, 70%, 80%, 90%, 95%, and 100% v/v) for 15 minutes at each concentration. After removing the tert-butanol under a low-temperature vacuum, the cell surface was coated with gold. Finally, cell phenotypes were observed under a scanning electron microscope (JEOL, Japan).

msdR RT-PCR

Total cellular RNAs were extracted using trizol (R0016, Beyotime, China) following the manufacturer's instructions. The extracted total RNAs underwent treatment with DNase I (2270A, Takara, Japan) at 37 °C for 30 minutes to eliminate genomic DNA. Subsequently, 3 μg of total RNA was utilized for reverse transcription (RT) to obtain cDNA using the PrimeScript RT-PCR Kit (RR014A, Takara, Japan), employing msdR-specific RT primers rather than oligo dT or random primers. These RT primers, tailored to msdRs, were employed to minimize non-specific binding. Next, primers spanning at least two exons were designed for PCR amplification with the cDNA as a template. Amplification was carried out using the PrimeScript RT-PCR Reagent Kit (RR014A6110A, Takara, Japan) through 30 cycles at 94 °C for 30 seconds, 53 °C for 30 seconds, and 72 °C for 60 seconds. The PCR products were subsequently analyzed via 1% agarose gel electrophoresis and confirmed by DNA sequencing. All primer sequences are provided in table S11.

msRNA qRT-PCR

For msRNA qRT-PCR, total cellular RNAs were extracted using trizol (R0016, Beyotime, China) following the manufacturer's instructions. 3 μg of total RNA was utilized for reverse transcription (RT) to obtain cDNA using the PrimeScript RT-PCR Kit (RR014A, Takara, Japan), employing msRNA-specific RT primers. U6 snRNA was used as an internal control. The resulting RT product was subsequently used as a template for quantitative real-time PCR using SYBR Green qPCR Mix (D7260, Beyotime, China) according to the manufacture's instructions. The ΔΔCt method was used to analyze the expression of msRNAs. All primers used in msRNA qRT-PCR are provided in table S13.

Western blot

Cellular proteins were extracted using RIPA lysis buffer (composed of 50 mM Tris, 150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, and 0.1% SDS) containing protease inhibitors (P0013C, Beyotime, China). Following quantification with the BCA Protein Assay Kit (P0009, Beyotime, China), the proteins were separated via 10% SDS-PAGE and transferred onto a PVDF membrane (IPVH00010, Millipore, USA). The membrane was then immersed in QuickBlock blocking buffer (P0220, Beyotime, China) and subsequently incubated with the following antibodies for 2 hours at room temperature: anti-msdP0188 pAb (1:500), anti-TERT antibody (1:500, ABclonal, China), anti-tubulin antibody (1:1000, Affinity, USA), or beta-actin antibody (1:2000, Affinity, USA). Following this, the membranes were incubated with horseradish peroxidase-labeled secondary antibody (1:3000, Affinity, USA) at room temperature for 1 hour. After washing the membrane three times with PBS containing Tween20, protein bands were detected using the ECL Chemiluminescence Kit (P0018S, Beyotime, China). All tests were done in triplicate.

Human telomerase reverse transcriptase (TERT) inhibition and knockdown

Two TERT inhibitors, BIBR-1532 (S1186, Sellect, USA) and zidovudine (HY-17413, MedChemExpress, USA), were used in the TERT inhibition assay. A549 cells were seeded in 24-well plates at a density of 2×10^5 cells/well. After cell density reached 70%, they were treated individually with each inhibitor for specific durations and concentrations: BIBR1532 for 24 hours at 20, 50, 80, and 100 μM, respectively; and zidovudine for 60 hours at 20, 50, 80, and 100 μM, respectively. Untreated cells and dimethyl sulfoxide (DMSO) treated cells were used as controls.

Alternatively, TERT was knocked down using siRNAs, including two reported and three self-designed siRNAs. The sequences for indicated siRNAs were as follows: 5'-GAACUUCCCUGUAGAAGACGATT-3' for siRNA-1 (*69*), 5'-GCAUUGGAAUCAGACAGCATT-3' for siRNA-2 (*70*), 5'-GCGACGACGUGCUGGUUCACCTT-3' for siRNA-432, 5'-CGGUGUACGCCGAGACCAAGCTT-3' for siRNA-966, 5'-GGAAGAGUGUCUGGAGCAAGUTT-3' for siRNA-1728, and 5'-UUCUCCGAACGUGUCACGUTT-3' for control siRNA. A549 cells seeded in 24-well plates were transfected with siRNAs at a final concentration of 50 nM using Xfec RNA Transfection Reagent (631450, Takara, Japan). After 72 hours, cells were collected for protein extraction. Total proteins were extracted using the RIPA lysis buffer containing protease inhibitors (P0013C, Beyotime, China). Following quantification using the BCA Protein Assay Kit (P0009, Beyotime, China), proteins were separated by 10% SDS-PAGE and transferred to a PVDF membrane (IPVH00010, Millipore, USA). The membrane was then placed in 5% bovine serum albumin and incubated with the following antibodies: anti-msdP0188 pAb (1:500), anti-TERT antibody (1:500, ABclonal, China), or anti-tubulin antibody (1:1000, Affinity, USA), for 2 hours at room temperature, followed by incubation with horseradish peroxidase-labeled secondary antibody (1:3000, Affinity, USA) at room temperature for 1 hour. After washing the membrane three times with PBS-containing Tween20, protein detection was performed using the ECL Chemiluminescence Kit (P0018S, Beyotime, China). All tests were done in triplicate.

Silver staining

After washing the cells twice with 1× PBS, proteins were extracted using 500 µl cell lysate (P0013C, Beyotime, China). A total of 50 µl protein A/G magnetic beads (HY-K0202, MCE, USA) were then incubated with the anti-msdP0188 pAb or rabbit IgG (negative control) at room temperature for 30 minutes. Subsequently, the harvested beads were incubated with the extracted proteins at 4 °C for 2 hours. After elution using 1× SDS-PAGE loading buffer, the protein complex was separated by 12% SDS-PAGE electrophoresis. Silver staining was performed following the provided instructions (P0017S, Beyotime, China). The gel was fixed with 100 ml of fixative solution at room temperature for 20 minutes, followed by washing with 30% ethanol and pure water for 10 minutes each. The silver solution was then applied for 10 minutes at room temperature, followed by washing with pure water. The reaction was terminated by treatment with the color-developing solution for 10 minutes. After photography, the ~18 kDa protein band of interest was excised for subsequent analysis.

Mass spectrometry (MS)

A. Generation of MS spectra data from human tumor cell lines

After harvesting, the four tumor cell lines (HeLa, PANC-1, U118, and HCT116) underwent a triple wash with 1× PBS followed by the addition of RIPA buffer (R0010, Solarbio, China) containing a mixture of protease inhibitors (P6730, Solarbio, China) to induce cell lysis. After centrifugation, the resulting supernatant was subjected to protein concentration measurement using a BCA kit (PC0020, Solarbio, China). Subsequently, 5× loading buffer (P1040, Solarbio, China) was added to the cell lysate, which was then boiled for 10 minutes. Following this, 100 µg of protein was subjected to separation via 10% SDS-PAGE electrophoresis. The gel was then stained with Coomassie brilliant blue R250 solution for 20 minutes, and the color was removed using a decolorization solution comprising 50 mM NH_4HCO_3 and acetonitrile (1:1, v/v) until the background was clear. The gel was cut into approximately 1-mm^3 pieces, and decolorization continued until the blue coloration was completely faded, followed by dehydration with acetonitrile. Reduction alkylation of proteins was achieved by sequential addition of 10 mM DTT (treated at 56 °C for 1 hour) and 55 mM IAA (treated at room temperature in the dark for 45 minutes) to the gel. The gel was then rinsed twice with 25 mM NH_4HCO_3 and a mixture of acetonitrile and 50 mM NH_4HCO_3 (1:1, v/v), respectively, before being dehydrated again with acetonitrile. Trypsin was then added, and the reaction proceeded for 16 hours at 37 °C. The resulting digested peptides were extracted with a solution of acetonitrile, formic acid, and water (7:1:2, v/v/v), and subsequently dried using a vacuum concentrator. The peptides were redissolved in 0.1% formic acid aqueous solution for liquid chromatography with tandem mass spectrometry (LC-MS/MS) detection. The LC-MS/MS system comprises an EASY-nLC system and an Orbitrap Fusion Tribird mass spectrometer. Separation was conducted using an Acclaim PepMap 100 column (75 µm × 15 cm, nanoViper, C18, 3 µm, 100 Å; Thermo Scientific), with a mobile phase consisting of acetonitrile (B) and 0.1% formic acid aqueous solution (A) at a flow rate of 300 nl/min. The composition of the mobile phase starts at 97% A, linearly decreasing with the following schedule: 97%–92% A (0–5 minutes), 92%–82% A (5–75 minutes), 82%–72% A (75–103 minutes), 72%–10% A (103–115 minutes), and finally keeping at 10% A for 5 minutes. MS data acquisition utilized a data-dependent acquisition (DDA) mode, with scanning and fragmentation modes of MS being orbitrap (OT)–higher energy collisional

dissociation (HCD)–iontrap (IT).

B. Identification of msdP0188 using MS

To identify msdP0188, we utilized an Acuity UPLC system (Thermo Scientific) coupled with a Q Exactive orbitrap mass spectrometer (Thermo Scientific) for protein identification through peptide mapping. Following silver staining of the SDS-PAGE gel, the ~18 kDa band of interest was excised, de-stained, reduced, and alkylated. Subsequently, in-situ trypsin digestion was performed for 16 hours at 37 °C. The resulting digested peptides were then extracted using a solution of acetonitrile, formic acid, and water (7:1:2, v/v/v), and dried directly via vacuum concentration. Peptide separation was achieved using a UPLC system (Thermo Scientific) equipped with a reversed-phase CSH C18 column (150 mm, 1.7 µm, Waters), with detection performed by a Q Exactive orbitrap mass spectrometer (Thermo Scientific) operating in high-sensitivity mode. The flow rate was set at 200 µl/min, with a column temperature maintained at 40 °C. Solvent A consisted of 0.1% formic acid (FA) in water, while solvent B consisted of 0.1% FA in acetonitrile. During the 120-min gradient, the percentage of solvent B level was increased progressively: from 2% to 29% over 60 minutes, from 29% to 55 % in 5 minutes, from 55% to 75% in 2 minutes, and held at 75% for 5 minutes. Subsequently, it was returned to 2% over 2 minutes and maintained at 2% for the remaining 6 minutes. The digestion products by Glu-C (staphylococcal V8 protease) and chymotrypsin were similarly analyzed.

Synthesis of Circular single-stranded DNA (CSSD)

CSSD was designed to contain multiple complementary antisense sequences for msRNA absorption. Here CSSD4 contains reverse complementary sequences of four msRNAs, while the control CSSD (Ctrl CSSD) comprises four repetitive scrambled sequences (see below, underlined). All nucleic acids have phosphorylated 5'-ends.

For Ctrl CSSD, the single-stranded DNA 5'-AATTCAAAGAATTAACCTTAATTGAAGGGGAGGGTTCAGTACTTTTGTGTAGTACAAATATCAGTACTTTTGTGTAGTACAAAAGGGAGGGCTTCAATTAAGGTTAATTCTTTG-3' was diluted in annealing buffer (Beyotime, China), heated at 95 °C for 2 minutes, and then cooled to 4 °C at a rate of 0.1 °C per 8 seconds. Subsequently, the annealed DNAs were treated with T4 DNA ligase (Takara, Beijing, China) for 2 hours at room temperature to form a complete circular DNA.

For CSSD4, two single-stranded DNAs were synthesized: 5'-AATTCAAAGAATTAACCTTAATTGAAGGGGAGGGTTCACCACTGCACTCGCTCCAGCCATATGGCTCATGCCTGTAATCCCAGCAAGGGAGGGCTTCAATTAAGGTTAATTCTTTG -3' containing binding sequences that are reverse complementary to *miR-619-5p* and *miR-3135b* (underlined) and 5'-AATTCAAAGAATTAACCTTAATTGAAGGGGAGGGTTCTGGGGCAGGAGGGAGGTCAGCATGATATACCTGTAGTCCCAGCTATTCAGAAGGGAGGGCTTCAATTAAGGTTAATTCTTT-3' containing binding sequences that are reverse complementary to *miR-4728-3p* and *miR-5585-3p* (underlined). After separate annealing with annealing buffer (Takara, Beijing, China), the two DNA strands were treated with

T4 DNA ligase (Takara, Beijing, China) to form a completely closed circular DNA (fig. S12). All single-stranded DNAs were synthesized by Genewiz Biotechnology (Beijing, China).

Immunohistochemistry (IHC) assay

Various tissue samples including 15 lung cancer tissues, 9 normal lung tissues, 10 general breast cancer tissues, 12 triple-negative breast cancer (TNBC) tissues, and 8 normal breast tissues were analyzed using IHC assays. Tumor tissues were fixed with 10% formalin and embedded in paraffin. For the IHC analysis, the embedded tissues were sectioned into 4 μm-thick slices. Antigen retrieval was performed using sodium citrate solution (Beyotime, China), followed by blocking with 3% H_2O_2. Subsequently, the sections were incubated overnight at 4 °C in the dark with anti-msdP0188 pAb (1:200). Negative controls were prepared by omitting the anti-msdP0188 pAb during the incubation step. After washing with PBS, the samples were incubated with a horseradish peroxidase-conjugated secondary antibody (1:2000, Abcam, UK). Subsequently, staining development was carried out using the DAB kit (Beyotime, China), followed by counterstaining with hematoxylin. As msdP0188 is a novel protein and its tissue-specific expression is not yet fully understood, establishing a positive control for IHC analysis is currently not feasible.

Polyclonal Antibody (pAb)

For selected msdPs, specific immunogenic peptides were designed and synthesized (Apeptide, Shanghai, China) and confirmed through mass spectrometric analysis. These immunogenic peptides have no homologous sequences with the annotated proteins after blast comparison. An additional Cys residue was appended at the C-terminal for conjugation with the keyhole limpet hemocyanin protein. The immunogenic peptide sequences are provided in table S9.

The immunogens were then used to immunize rabbits. In the initial immunization, 350 μg of antigen was mixed with 1,000 μl of Freund's complete adjuvant (Sigma, F5881) and subcutaneously injected at multiple sites. Subsequently, for three follow-up immunizations, 200 μg of antigen mixed with 1,000 μl of Freund's incomplete adjuvant (Sigma, F5506) was administered subcutaneously using the same procedure. The interval between immunizations was 14 days. On day 7 following the final immunization, the rabbits were euthanized, and blood was collected and centrifuged to obtain antiserum for subsequent antigen affinity purification, a method that is widely used to obtain the purest pAbs with the least amount of cross-reactivity. The antiserum from the immunized rabbits was incubated in the column filled with epoxy-activated agarose (Wechsler Chromatography, Beijing, China) conjugated with peptide immunogens and then eluted to obtain relatively purified antigen-specific pAbs.

We demonstrated the pAbs' sensitivity and specificity through a series of experiments, using anti-msdP0118 pAb as an example. We constructed a vector, pET28a-msdP0188, to overexpress msdP0188 in *E. coli* and obtain recombinant msdP0188. The purified msdP0188 protein was used as a positive control in Western blot analysis to show the sensitivity of the anti-msdP0188 pAb, as we expected that the anti-msdP0188 pAb must be able to detect msdP0188 effectively. SDS-PAGE was run to separate recombinant msdP0188 and

A549 cell lysate simultaneously, and in all the groups, ~18 kDa bands that match the calculated molecular weight were detected (fig. S5). Following this, we conducted IHC staining by treating human lung cancer and triple-negative breast cancer tissues with anti-msdP0188 pAb, using horseradish peroxidase-labeled goat anti-rabbit secondary IgG (isotype IgG)-treated or untreated (PBS) human cancer tissues as negative controls (fig. S6). Robust staining signals were observed in the anti-msdP0188 pAb group, whereas the two control groups showed no or much weaker discernible signals (fig. S6). Besides, in the silver staining experiment we also detected a ~18 kDa protein band in the A549 cell lysate using protein A/G magnetic beads that are capable of binding anti-msdP0188 pAb, and confirmed the identity of proteins within this band as msdP0188 through MS analysis (Fig. 1). Following the same bioinformatic pipeline, we did not anticipate the presence of msdP0188 in mice. As expected, in contrast to the robust signals observed for the anti-msdP0188 pAb in human cancer tissues (fig. S6), IHC assays in mouse liver and spleen with anti-msdP0188 pAb or PBS showed no detectable staining signals (fig. S7). Together, these data demonstrate the high sensitivity and specificity of the anti-msdP0188 pAb to target msdP0188.

Clinical specimens

Tumor samples embedded in paraffin from Qinhuangdao Municipal No. 1 Hospital were reviewed by a pathologist. The study was conducted under the guidelines outlined in the Declaration of Helsinki and its subsequent amendments, or comparable ethical standards.

Xenograft tumor in nude mice

BALB/C nude mice were utilized to assess tumor growth and metastasis. For the tumor growth experiments, 24 BALB/C nude 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 subsequently treated with CSSD4 or Ctrl CSSD via intratumoral injection at a dosage of 20 μg/mouse, administered once every seven days for a total of three treatments. The same procedures were repeated with the other two groups using MDA-MB-231 cells. Tumor sizes were measured every three days, and tumor volume was calculated using the formula: $tumor\ volume = (major\ axis\ of\ tumor)\times(minor\ axis\ of\ tumor)^2/2$. At the end of the experiment, all mice were euthanized on day 34 via CO_2 asphyxiation. Tumor tissues were fixed and embedded in paraffin for subsequent analysis. The expression of msdP0188 in solid tumors was detected using immunohistochemical staining.

For the tumor metastasis experiments, 12 BALB/C nude 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 a dosage of 30 μg/mouse, administered once every seven days for a total of three treatments.

All animal experiments were conducted in compliance with the regulations outlined by the Institutional Animal Care and Use Committee (IACUC) of the Tianjin International Joint Academy of Biomedicine.

Statistical analysis

Statistical analyses were performed using the GraphPad Prism software 9.0 (La Jolla, CA, USA), and results were presented as the mean ± standard deviation. Comparison between two groups was subjected to unpaired, two-tailed Student's *t*-test, while comparisons among multiple groups were conducted via one-way ANOVA followed by Tukey's posthoc test. $^{*}P < 0.05$; $^{**}P < 0.01$; ns, not significant.

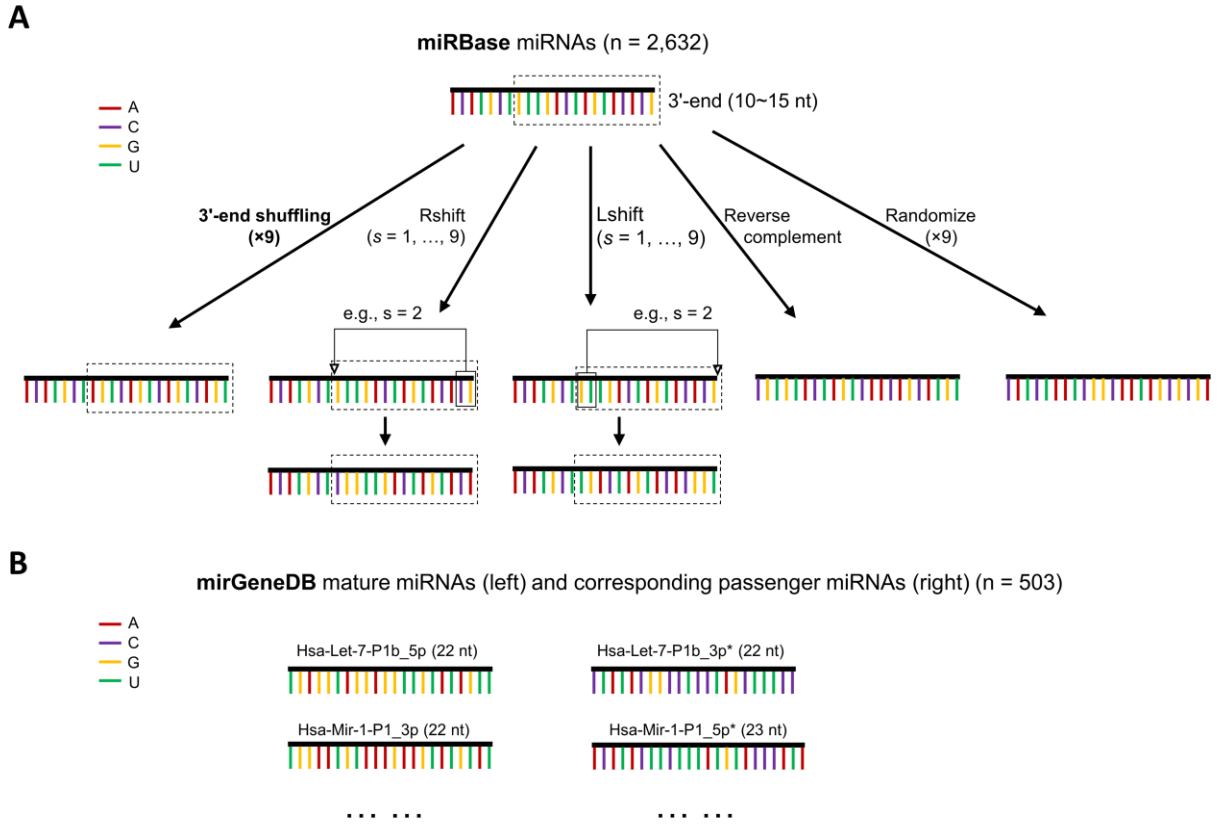

Fig. S1. Diagrams of methods to generate control msRNAs for estimating the false discovery rate of msdR prediction. (**A**) 3'-end shuffling, shifting (Rshift and Lshift), reverse complement, and random generation based on the unique human miRNAs in the miRBase database (n = 2,632). (**B**) Human mature miRNAs and passenger miRNAs (or miRNA* strands) in the mirGeneDB database (n=503).

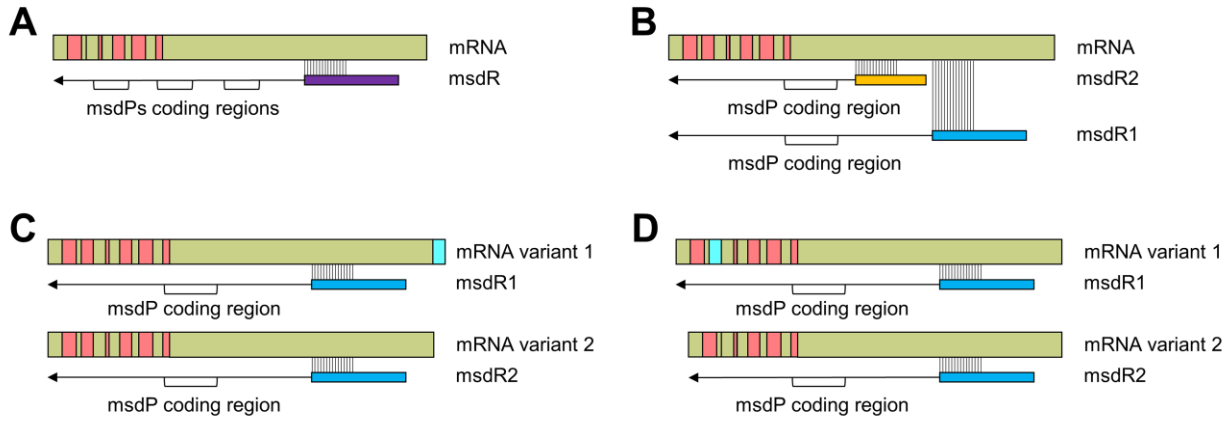

Fig. S2. Scenarios for msdR and msdP derivation. (**A**) A single msdR may encode multiple msdPs. (**B**) Different msRNAs bound to the same RNA yield different msdRs but may encode identical msdPs. (**C**) RNA splicing outside of the msRNA elongation region yields identical msdRs. (**D**) RNA splicing within the msRNA elongation region leads to different msdRs, yet they may encode identical msdPs. In (C) and (D), the excised exon in mRNA variant 1 is depicted in cyan.

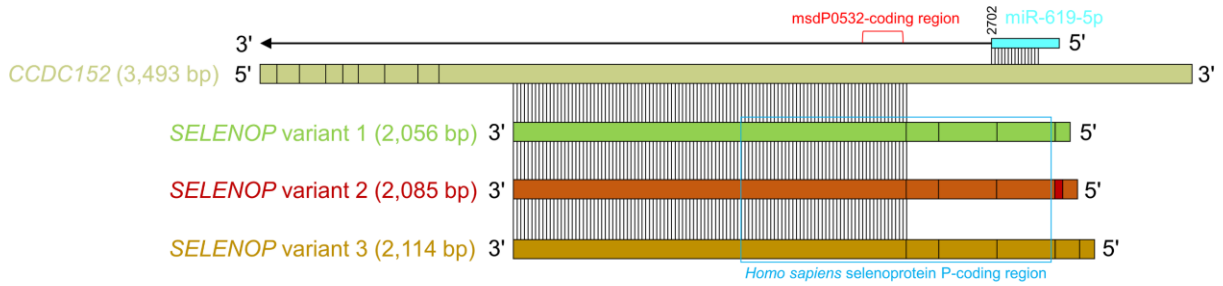

Fig. S3. Derivation of msdP0532 and its associated msdR is grounded in the interaction between *miR-619-5p* and mRNA *CCDC152*. Notably, *CCDC152* exhibits a reverse complementarity spanning 1,453 base pairs with all three SELENOP variants capable of encoding the protein, selenoprotein P. msdP0532 corresponds to selenoprotein P, plasma, 1 as encoded by *SEPP1*, wherein *SEPP1* comprises a truncated coding region of *SELENOP*, elucidated through conceptual translation. The exons in each mRNA are depicted as rectangle blocks.

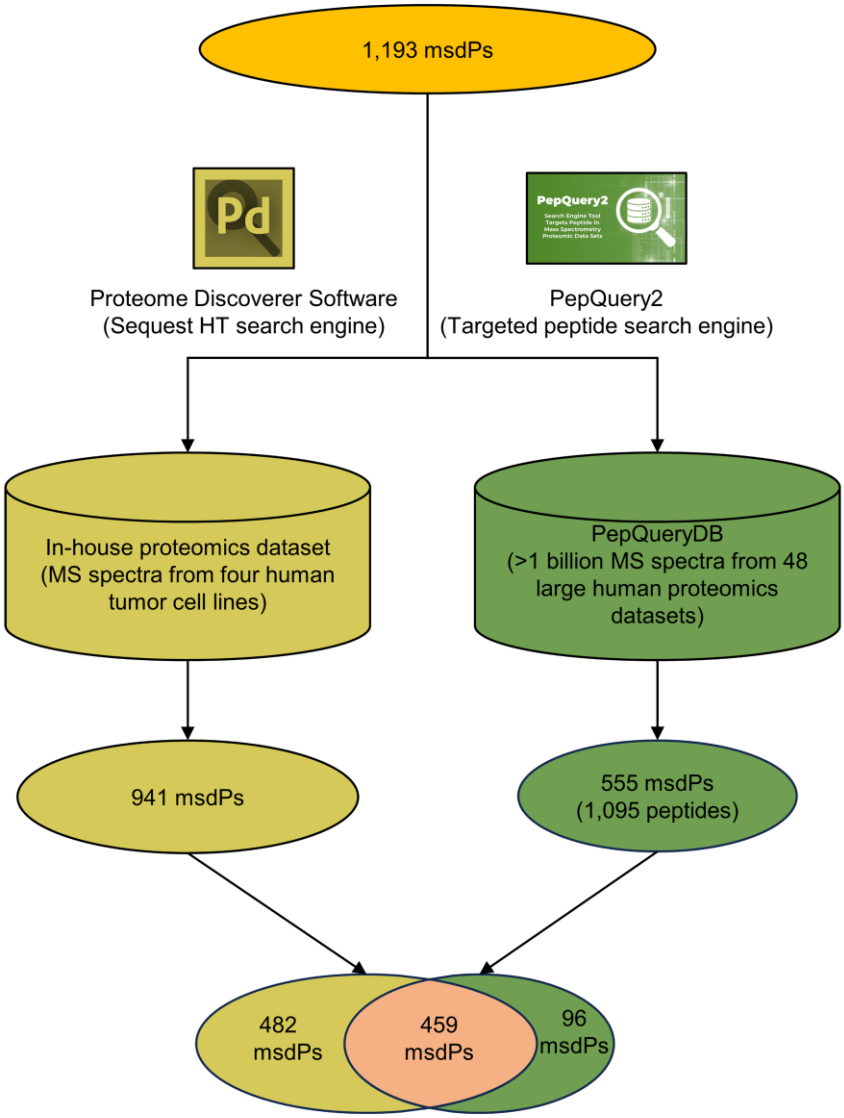

Fig. S4. Workflow for searching msdP peptide fragments in in-house and public proteomics databases. MS, mass spectrometry.

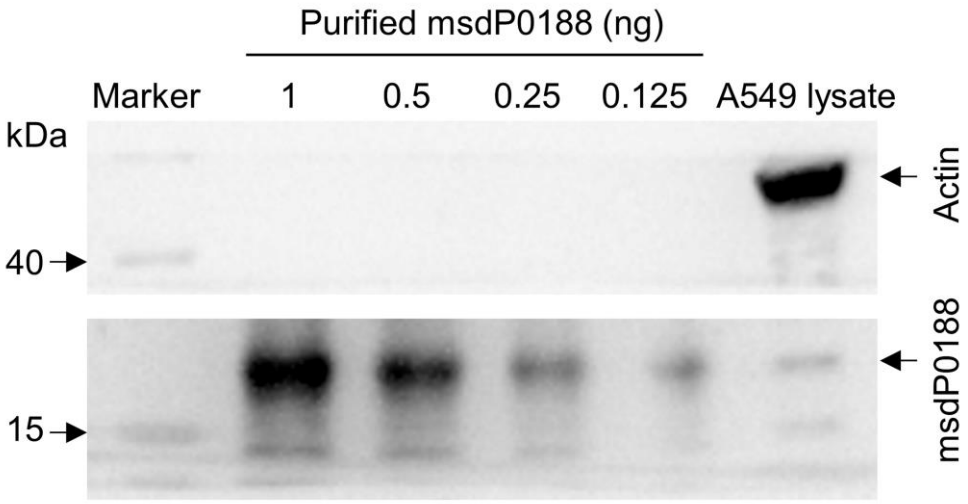

Fig. S5. Western blot analysis of purified msdP0188 protein and A549 cell lysate using anti-msdP0188 polyclonal antibody. Serial dilutions of purified msdP0188 were loaded for analysis.

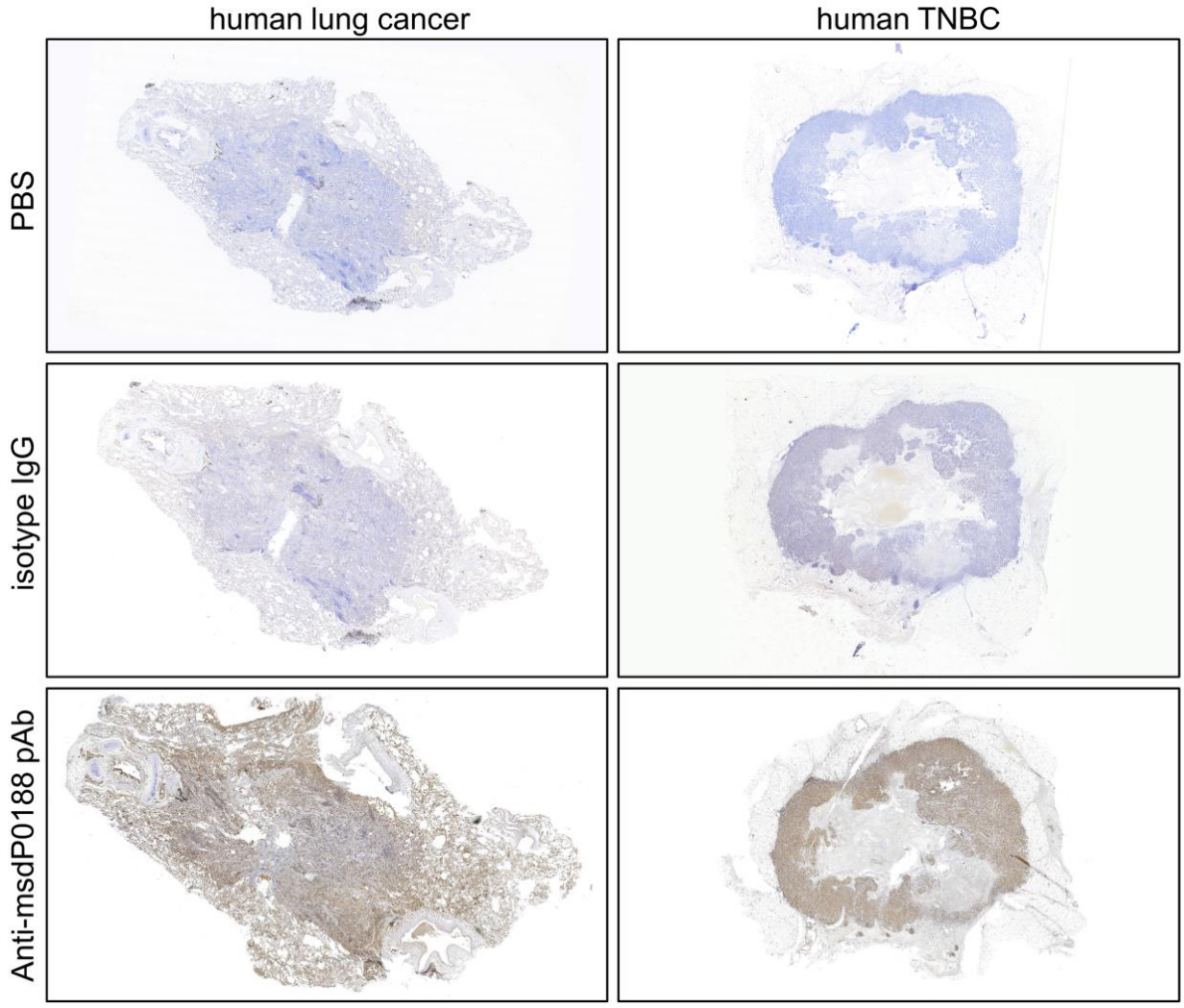

Fig. S6. Immunohistochemistry staining of human cancer tissues. pAb, polyclonal antibody. PBS, phosphate-buffered saline. TNBC, triple-negative breast cancer.

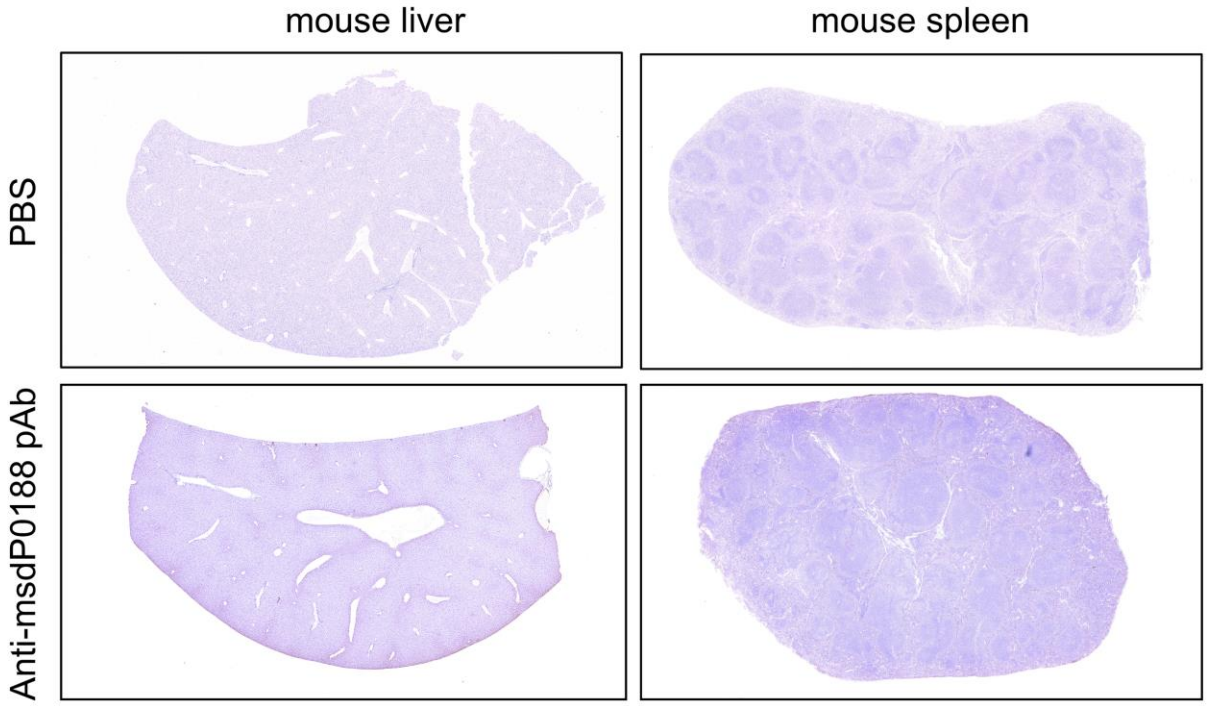

Fig. S7. Immunohistochemistry staining of mouse tissues. pAb, polyclonal antibody. PBS, phosphate-buffered saline.

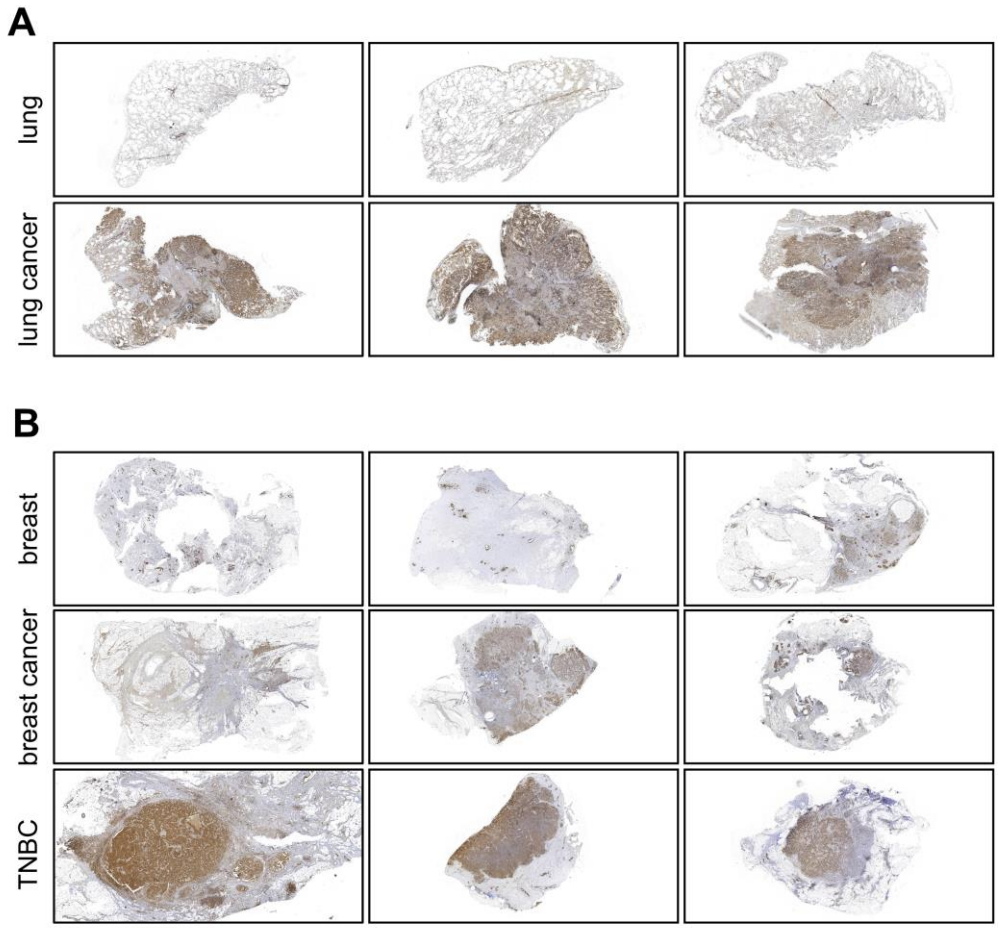

Fig. S8. Immunohistochemistry staining of tissues from various patients. (**A**) anti-msdP0188 polyclonal antibody (pAb) stained full-section images depicting normal lungs (n = 3) and lung cancer tissues (n = 3). (**B**) anti-msdP0188 pAb stained full-section images depicting normal breast tissues (n = 3), breast cancer tissues (n = 3), and triple-negative breast cancer (TNBC) tissues (n = 3). Each image corresponds to a different patient.

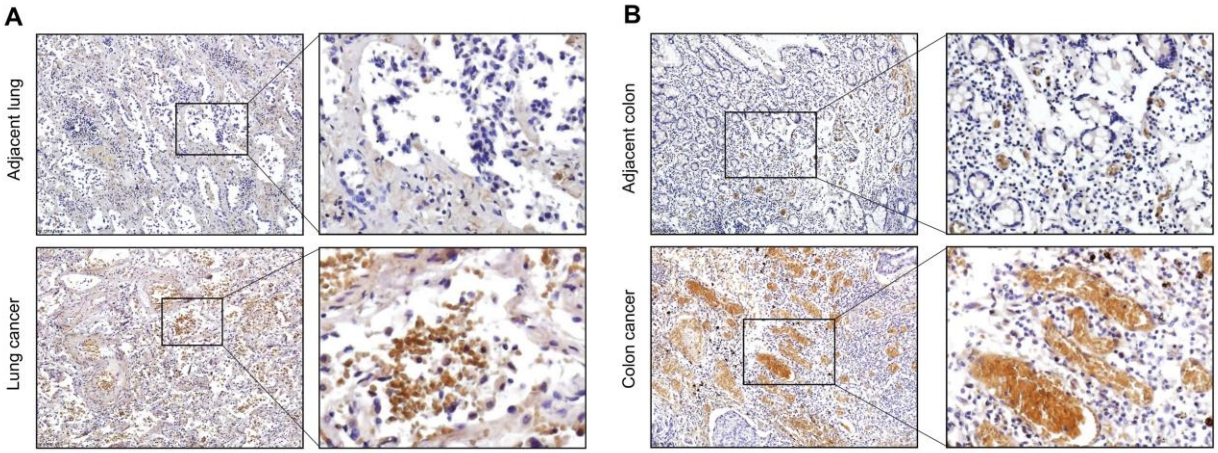

Fig. S9. Expression of msdP0112 in lung and colon cancer tissues. (**A**) Immunohistochemistry staining of lung cancer and adjacent lung tissues using the anti-msdP0112 polyclonal antibody (left). Sections enclosed by black rectangles are magnified to 100× (right) for each core section. (**B**) Immunohistochemistry staining of colon cancer and adjacent colon specimens using the anti-msdP0112 polyclonal antibody (left). Sections enclosed by black rectangles are zoomed in to 100× (right) for each core section.

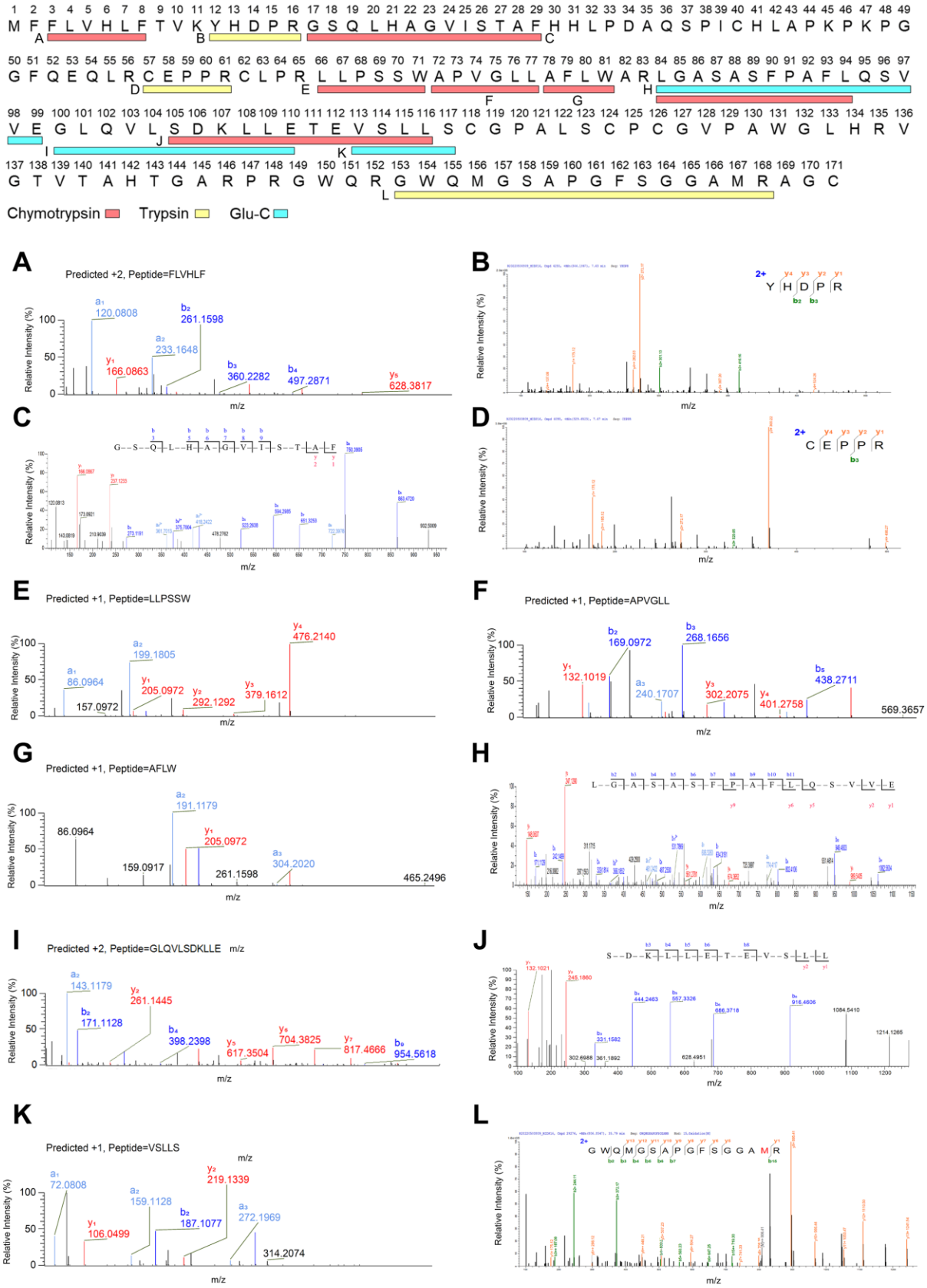
M F F L V H L F T V K Y H D P R G S Q L H A G V I S T A F H H L P D A Q S P I C H L A P K P K P G
G F Q E Q L R C E P P R C L P R L L P S S W A P V G L L A F L W A R L G A S A S F P A F L Q S V
V E G L Q V L S D K L L E T E V S L L S C G P A L S C P C G V P A W G L H R V
G T V T A H T G A R P R G W Q R G W Q M G S A P G F S G G A M R A G C
Chymotrypsin
Trypsin
Glu-C
A
Predicted +2, Peptide=FLVHLF
B
C
D
E
Predicted +1, Peptide=LLPSSW
F
Predicted +1, Peptide=APVGLL
G
Predicted +1, Peptide=AFLW
H
I
Predicted +2, Peptide=GLQVLSDKLLE
J
K
Predicted +1, Peptide=VSLLS
L
Relative Intensity (%)
m/z

Fig. S10. Mass spectrometry assay authenticates msdP0188 peptide fragments after protease digestion. The protein bound to the anti-msdP0188 polyclonal antibody during immunoprecipitation was separated via SDS-PAGE and silver staining to identify a potential target band. This band was isolated and subjected to digestion individually using various proteases. The digested samples were then analyzed using mass spectrometry. Thirteen peptide fragments within msdP0188 were validated, of which three were from trypsin digestion (highlighted in yellow), three from Glu-C digestion (cyan), and seven from chymotrypsin digestion (red).

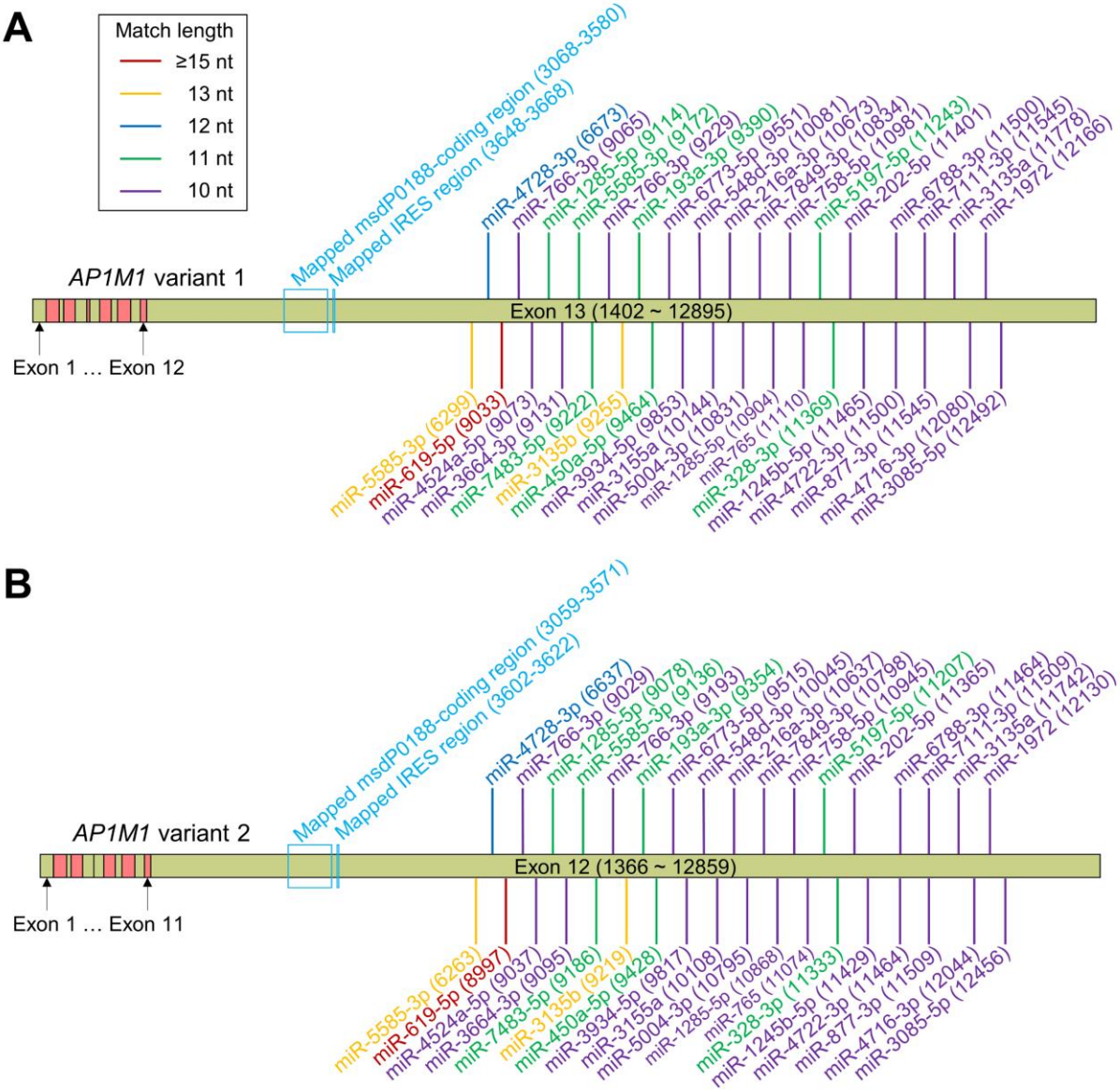

Fig. S11. Example msRNAs bound to the *AP1M1* mRNA with 10–15 nt matches. (**A**) In variant 1, all the ≥10-nt matches are displayed after position 9033, with only the ≥12-nt matches shown before this position. (**B**) In variant 2, all the ≥10-nt matches are displayed after position 8997, with only the ≥12-nt matches shown before this position.

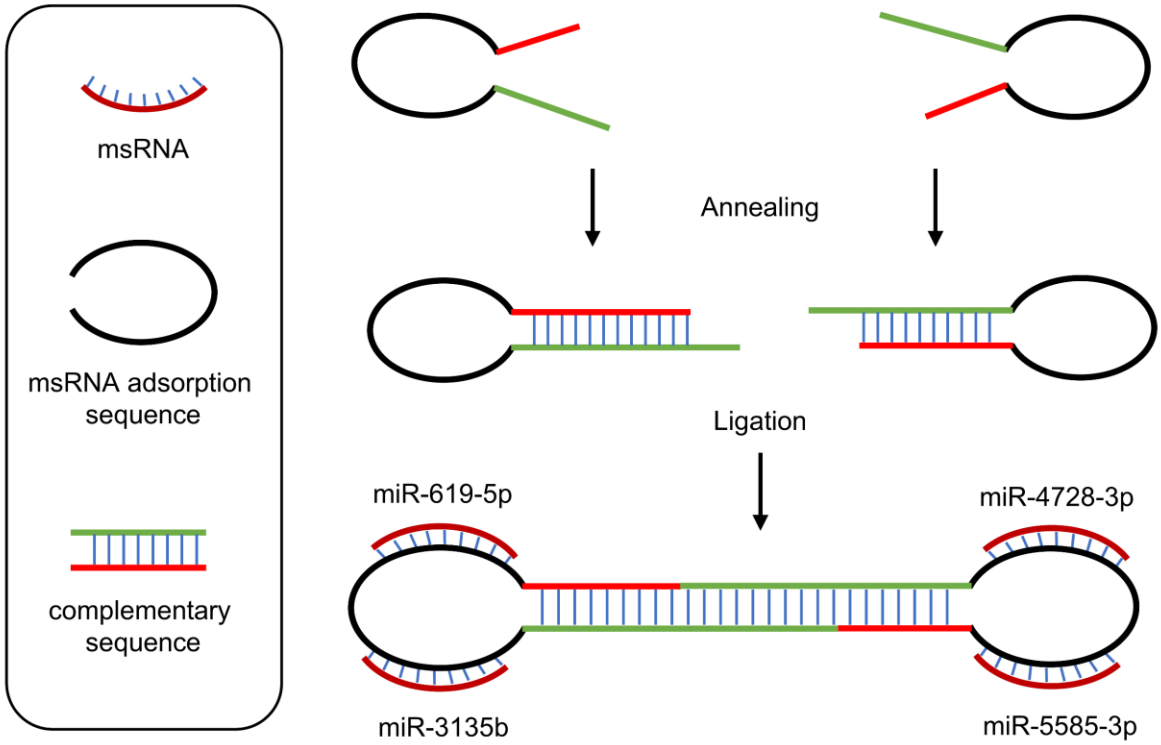

Fig. S12. Diagram of the route for CSSD4 design. CSSD4 is designed to simultaneously bind *miR-619-5p*, *miR-3135b*, *miR-4728-3p*, and *miR-5585-3p*. Both CSSD4 (comprising multiple complementary binding sites for msRNAs) and Ctrl CSSD (containing a scrambled DNA sequence) consist of two DNA half-rings. Single-stranded DNAs were initially synthesized and annealed to form half-rings. Subsequently, these half-rings were treated with T4 DNA ligase to create fully closed circular DNA structures.

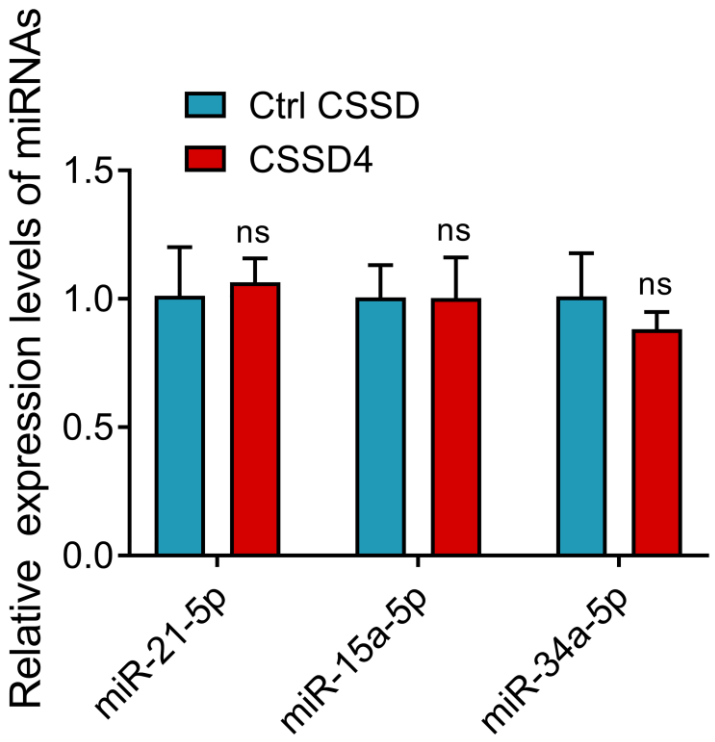

Fig. S13. Expression levels of *miR-21-5p*, *miR-15a-3p*, and *miR-34a-5p* in A549 cells as determined by qRT-PCR following CSSD transfection. Error bars represent standard deviations of n = 3 independent biological replicates. ns, not significant; Student's *t*-test.

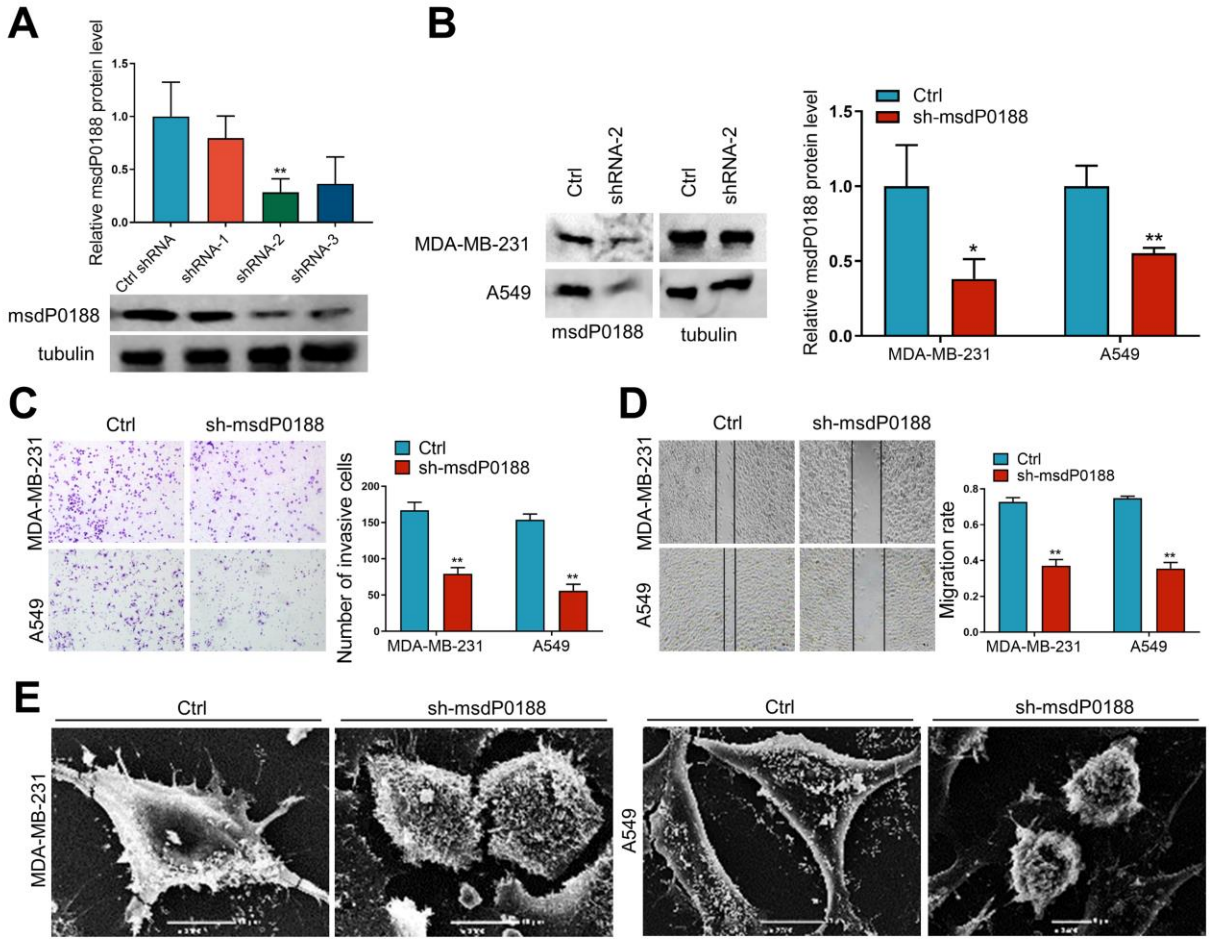

Fig. S14. Downregulation of msdP0188 via shRNA transfection inhibits the progression of lung and breast cancer cells *in vitro*. (**A**) A549 cells were transfected with three interfering lentiviruses targeting msdR0188, along with a control lentivirus carrying a scrambled sequence, and the msdP0188 protein level was assessed by Western blot to compare the efficacy of the three shRNAs. (**B**) Downregulation of msdP0188 levels in A549 and MDA-MB-231 cells. (**C**) Effects of sh-msdP0188 on the invasion of A549 and MDA-MB-231 cells. (**D**) Effects of sh-msdP0188 on the migration of A549 and MDA-MB-231 cells. (**E**) Visualization of cell phenotype using scanning electron microscopy after sh-msdP0188 transfection. In all figures, error bars represent standard deviations of n = 3 independent biological replicates. *$P < 0.05$; **$P < 0.01$; Student's *t*-test.

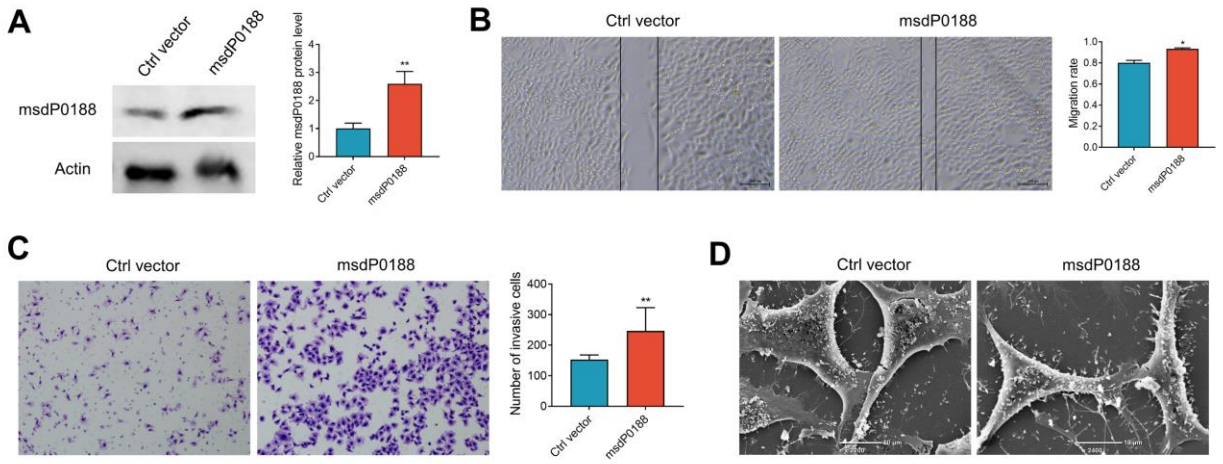

Fig. S15. Overexpression of msdP0188 promotes migration and invasion of A549 cells. A549 cells were transfected with either an empty vector (Ctrl) or the PiggyBac Dual promoter plasmid containing the msdP0188 coding sequence. (**A**) Western blot analysis to detect msdP0188 protein levels. (**B**) Effects of msdP0188 overexpression on A549 cell migration. (**C**) Effects of msdP0188 overexpression on A549 cell invasion. (**D**) Phenotypic changes in A549 cells due to msdP0188 overexpression visualized by scanning electron microscopy. In all figures, error bars represent standard deviations of n = 3 independent biological replicates. *$P < 0.05$; **$P < 0.01$; Student's *t*-test.

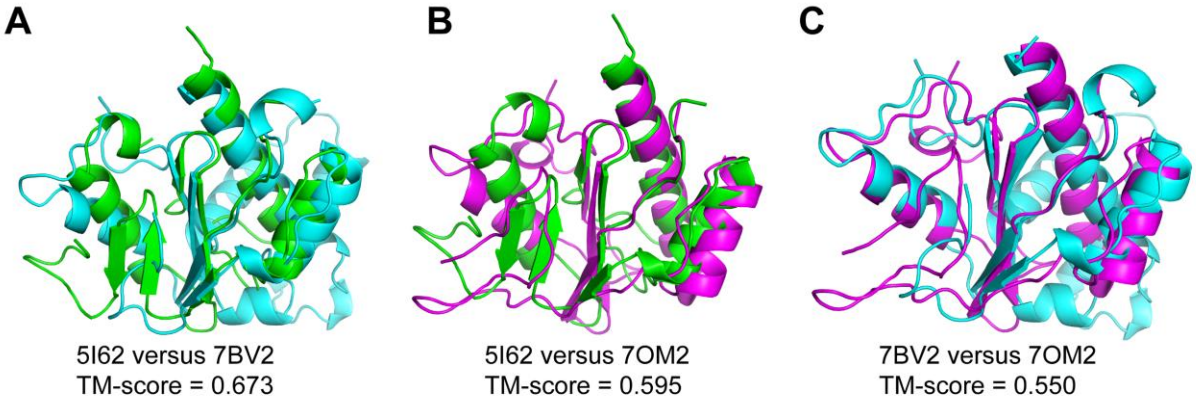

Fig. S16. Comparison of the *palm* subdomains of three viral RdRPs. (**A**) Human picobirnavirus strain Hy005102 (green, PDB ID: 5I62) versus SARS-CoV-2 (cyan, PDB ID: 7BV2). (**B**) Human picobirnavirus strain Hy005102 (green) versus *Thosea asigna* virus (magenta, PDB ID: 7OM2). (**C**) SARS-CoV-2 (cyan) versus *Thosea asigna* virus (magenta). The *palm* subdomain of human picobirnavirus strain Hy005102 spans amino acids 231–267 and 325–414. The *palm* subdomain of SARS-CoV-2 spans amino acids 581–627 and 687–812. The *palm* subdomain of *Thosea asigna* virus spans amino acids 304–374 and 444–519.

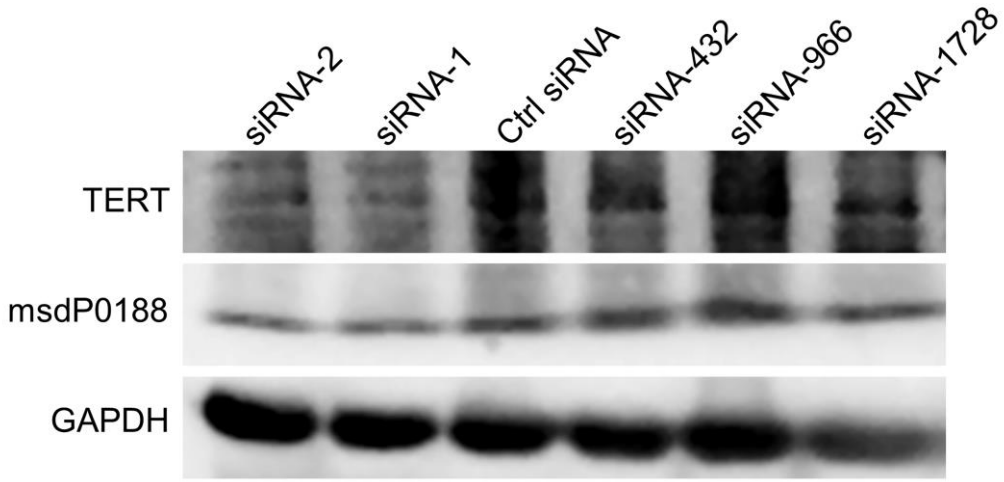

Fig. S17. Knockdown of human telomerase reverse transcriptase (TERT) using different siRNAs.

Table S1. Redundant miRNA sequences in the miRBase database.

Names of redundant miRNAs[a]	Sequence (5'→3')
hsa-miR-199a-3p, hsa-miR-199b-3p	ACAGUAGUCUGCACAUUGGUUA
hsa-miR-365a-3p, hsa-miR-365b-3p	UAAUGCCCCUAAAAAUCCUUAU
hsa-miR-518e-5p, hsa-miR-519a-5p, hsa-miR-519b-5p, hsa-miR-519c-5p, hsa-miR-522-5p, hsa-miR-523-5p	CUCUAGAGGGAAGCGCUUUCUG
hsa-miR-519a-2-5p, hsa-miR-520b-5p	CCUCUACAGGGAAGCGCUUUC
hsa-miR-518d-5p, hsa-miR-520c-5p, hsa-miR-526a-5p	CUCUAGAGGGAAGCACUUUCUG
hsa-miR-517a-3p, hsa-miR-517b-3p	AUCGUGCAUCCCUUUAGAGUGU
hsa-miR-516a-3p, hsa-miR-516b-3p	UGCUUCCUUUCAGAGGGU
hsa-miR-518a-5p, hsa-miR-527	CUGCAAAGGGAAGCCCUUUC
hsa-miR-548ai, hsa-miR-570-5p	AAAGGUAAUUGCAGUUUUUCCC
hsa-miR-548am-5p, hsa-miR-548c-5p, hsa-miR-548o-5p	AAAAGUAAUUGCGGUUUUUGCC
hsa-miR-548aj-5p, hsa-miR-548g-5p, hsa-miR-548x-5p	UGCAAAAGUAAUUGCAGUUUUUG
hsa-miR-548h-3p, hsa-miR-548z	CAAAAACCGCAAUUACUUUUGCA
hsa-miR-548aa, hsa-miR-548t-3p	AAAAACCACAAUUACUUUUGCACCA
hsa-miR-3689a-5p, hsa-miR-3689b-5p, hsa-miR-3689e	UGUGAUAUCAUGGUUCCUGGGA
hsa-miR-3689b-3p, hsa-miR-3689c	CUGGGAGGUGUGAUAUUGUGGU
hsa-miR-548ad-5p, hsa-miR-548ae-5p	AAAAGUAAUUGUGGUUUUUG

[a]For each set of redundant miRNAs, the one in italics was used for matching.

Table S2. Summary of number of matched RNAs and predicted msdRs and msdPs.

Match length (nt)	Number of matched RNAs[a]	Number of unique predicted msdRs[b]	Number of unique (redundant) predicted msdPs[c]
15	12,953	11,121	1,239 (5,369)
14	28,531	24,048	2,135 (9,629)
13	80,754	66,794	4,014 (22,829)
12	243,976	200,386	4,252 (34,225)
11	828,305	679,585	4,397 (71,208)
10	2,932,303	2,399,265	3,784 (162,694)

[a]miRNAs within the NCBI RefSeq RNA database were excluded from the matched RNAs.

[b]Number of redundant msdRs is identical to that of matched RNAs (the 2nd column). Unique predicted msdRs stand for nonredundant msdR sequences.

[c]Theoretically, msdPs derived from shorter matches will always include those 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 possible alignments, and therefore some hits from longer matches may be missing as the msdR database's size increases.

Table S3. Estimation of potential false discovery rates (FDRs) for msdR prediction using 3'-end shuffled msRNA artifacts based on miRBase miRNAs (n = 2,632).

Match length (nt)	Number of predicted msdRs with miRBase miRNAs[a]	Mean and standard deviation of number of predicted msdRs with msRNA artifacts[b]	FDR (%)[c]
15	11,121	1,695 ± 245	15.2 ± 2.2
14	24,048	6,206 ± 396	25.8 ± 1.6
13	66,794	23,260 ± 1,033	34.8 ± 1.5
12	200,386	91,330 ± 2,328	45.6 ± 1.2
11	679,585	357,048 ± 7,124	52.5 ± 1.0
10	2,399,265	1,393,073 ± 20,414	58.1 ± 0.9

[a]miRNAs within the NCBI RefSeq RNA database were excluded from being employed as templates for msdR prediction.

[b]The 3'-end (10–15 nt) of each miRBase miRNA was independently shuffled n = 9 times. Data are presented as mean ± standard deviation. The original numbers of msdRs predicted using each set of msRNA artifacts are presented in data S5.

[c]The deviations were calculated through error propagation.

Table S4. Estimation of potential false discovery rates (FDRs) for msdR prediction using completely randomly generated msRNA artifacts (n = 2,632).

Match length (nt)	Number of predicted msdRs with miRBase miRNAs[a]	Mean and standard deviation of number of predicted msdRs with msRNA artifacts[b]	FDR (%)[c]
15	11,121	1,214 ± 102	10.9 ± 0.9
14	24,048	4,724 ± 167	19.6 ± 0.7
13	66,794	19,638 ± 1,442	29.4 ± 2.2
12	200,386	77,375 ± 3,399	38.6 ± 1.7
11	679,585	302,461 ± 9,210	44.5 ± 1.4
10	2,399,265	1,191,447 ± 72,579	49.7 ± 3.0

[a]miRNAs within the NCBI RefSeq RNA database were excluded from being employed as templates for msdR prediction.

[b]A set of n = 9 msRNA artifacts of the same length was randomly generated for each miRBase miRNA. Data are presented as mean ± standard deviation. The original numbers of msdRs predicted using each set of msRNA artifacts are presented in data S5.

[c]The deviations were calculated through error propagation.

Table S5. Estimation of potential false discovery rates (FDRs) for msdR prediction using 3'-end right-shifted (Rshift) msRNA artifacts based on miRBase miRNAs (n = 2,632).

Match length (nt)	Number of predicted msdRs with miRBase miRNAs[a]	Mean and standard deviation of number of predicted msdRs with msRNA artifacts[b]	FDR (%)[c]
15	11,121	3,918 ± 302	35.2 ± 2.7
14	24,048	13,496 ± 595	56.1 ± 2.5
13	66,794	47,442 ± 2,093	71.0 ± 3.1
12	200,386	168,421 ± 4,588	84.0 ± 2.3
11	679,585	581,686 ± 10,654	85.6 ± 1.6
10	2,399,265	2,177,103 ± 40,912	90.7 ± 1.7

[a]miRNAs within the NCBI RefSeq RNA database were excluded from being employed as templates for msdR prediction.

[b]For each matching length, nine sets of msRNA artifacts were generated by shifting *s* nucleotides (*s* = 1, 2, …, 9) of the 3'-end (10–15 nt) of each miRBase miRNA from the 3'→5' direction (see fig. S1). Data are presented as mean ± standard deviation. The original numbers of msdRs predicted using each set of msRNA artifacts are presented in data S5.

[c]The deviations were calculated through error propagation.

Table S6. Estimation of potential false discovery rates (FDRs) for msdR prediction using 3'-end left-shifted (Lshift) msRNA artifacts based on miRBase miRNAs (n = 2,632).

Match length (nt)	Number of predicted msdRs with miRBase miRNAs[a]	Mean and standard deviation of number of predicted msdRs with msRNA artifacts[b]	FDR (%)[c]
15	11,121	4,383 ± 845	39.4 ± 7.6
14	24,048	14,314 ± 1976	59.5 ± 8.2
13	66,794	47,718 ± 2,850	71.4 ± 4.3
12	200,386	169,293 ± 4,708	84.5 ± 2.3
11	679,585	583,745 ± 13,722	85.9 ± 2.0
10	2,399,265	2,177,103 ± 40,912	90.7 ± 1.7

[a]miRNAs within the NCBI RefSeq RNA database were excluded from being employed as templates for msdR prediction.

[b]For each matching length, nine sets of msRNA artifacts were generated by shifting *s* nucleotides (*s* = 1, 2, …, 9) of the 3'-end (10–15 nt) of each miRBase miRNA from the 5'→3' direction (see fig. S1). Data are presented as mean ± standard deviation. The original numbers of msdRs predicted using each set of msRNA artifacts are presented in data S5.

[c]The deviations were calculated through error propagation.

Table S7. Estimation of potential false discovery rates (FDRs) for msdR prediction using reverse complementary strands of miRBase miRNAs (n = 2,632).

Match length (nt)	Number of predicted msdRs with miRBase miRNAs[a]	Mean and standard deviation of number of predicted msdRs with control msRNAs[b]	"FDR" (%)[c]
15	11,121	16,872	151.7
14	24,048	30,945	128.7
13	66,794	73,711	110.4
12	200,386	213,224	106.4
11	679,585	727,842	107.1
10	2,399,265	2,452,747	102.2

[a]miRNAs within the NCBI RefSeq RNA database were excluded from being employed as templates for msdR prediction.

[b]Control msRNAs are the reverse complements of miRBase miRNAs.

[c]Note that we provided the "FDR" values for comparison with those in tables S3–6, but they should not be interpreted as meaningful FDRs (see main text).

Table S8. Estimation of potential false discovery rates (FDRs) for msdR prediction using mirGeneDB's human mature miRNAs and corresponding passenger miRNAs (n = 503).

Match length (nt)	Number of predicted msdRs with miRBase miRNAs[a]	Mean and standard deviation of number of predicted msdRs with control msRNAs[b]	"FDR" (%)[c]
15	616	858	139.3
14	1,661	2,407	144.9
13	5,702	7,799	136.8
12	20,890	25,662	122.8
11	75,594	94,758	125.4
10	277,495	345,705	124.6

[a]miRNAs within the NCBI RefSeq RNA database were excluded from being employed as templates for msdR prediction.

[b]Control msRNAs are the passenger strands (also known as miRNA* strands) of the same set of human mature miRNAs taken from the mirGeneDB database.

[c]Note that we provided the "FDR" values for comparison with those in tables S3–6, but they should not be interpreted as meaningful FDRs (see main text).

Table S9. Summary of selected msdPs for immunohistochemistry assay.

msdP name	Length (AA)	Sequence	Immunogen sequence
msdP0045	15	MANIFPVFAPVTFRT	CANIFPVFAPVTFRT
msdP0112	25	MDFSTKHPVFSLKHKCHHVTGSCLT	PVFSLKHKCHHVTGSC
msdP0188	171	MFFLVHLFTVKYHDPRGSQLHAGVIST AFHHLPDAQSPICHLAPKPKPGGFQEQ LRCEPPRCLPRLLPSSWAPVGLLAFLW ARLGASASFPAFLQSVVEGLQVLSDKL LETEVSLLSCGPALSCPCGVPAWGLHR VGTVTAHTGARPRGWQRGWQMGSAP GFSGGAMRAGC	LAPKPKPGGFQEQLRC
msdP0323	68	MGFYHVGQAGLELLNSSDPPATASQSA GITGMGHRAQPLRSDFKQLYICELIYSG THSVFIIIIVCIF	CGMGHRAQPLRSDFKQ

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

Table S10. Summary of identifiers and names for selected msdPs and msdRs.

msdP name	msdP and msdR identifier(s)[a]	msdR name[b]
msdP0045	msdP.1370471854(RPS6KB1).414, *msdR.hsa-miR-1972.1370471854(RPS6KB1).2511*	msdRs0045
msdP0112	msdP.1370451977(NAXE).1372, *msdR.hsa-miR-5585-3p.1370451977(NAXE).3203*	msdRs0112
msdP0188	msdP.1677479768(AP1M1).3571, *msdR.hsa-miR-619-5p.1677479768(AP1M1).8997*; msdP.1677502062(AP1M1).3607, *msdR.hsa-miR-619-5p.1677502062(AP1M1).9033*	msdRs0188
msdP0323	msdP.1819229406(C9orf129).1643, *msdR.hsa-miR-3686.1819229406(C9orf129).4888*	msdRs0323

[a]The nomenclature of msdP and msdR identifiers can be referred to in Methods. Different msdR sequences may encode the same msdP sequence; such msdRs 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, the *AP1M1* variants 1 and 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. S9). 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 msdRs which may encode the same protein with different identifiers by our nomenclature (e.g., both msdP.1677479768(AP1M1).3571 and msdP.1677502062(AP1M1).3607 are identifiers for msdP0188).

[b]The number in msdRs name was arbitrarily set identical to that in msdP for convenience of data presentation.

Table S11. Summary of RT-PCR amplified regions of selected msdRs.

msdR name[a]	Amplified region[b]	RT primer[c]	Forward primer[c]	Reverse primer[c]
msdRs0045	442–831 (exons 4–8)	ATGGCGGCAGCGGCTGTGGT	AAACTCCACCAATCCACAGC	AGTAAAGCATCCCTTCATCG
msdRs0112	186–675 (exons 1–5)	ATGGCGAGCACGGTGGTGAAG	TGGCAATGGGCACAGTGAGT	GCGAGCACGGTGGTGAAGTA
msdRs0188.1	602–1318 (exons 5–11)	GGTTGCCACATCCAAGAAGA	CCGGAGGTAGTGAAGTAAGG	AGGCATCAAGTATCGGAAGA
msdRs0188.2	594–969 (exons 5–9)	AGACCACCGACAGCAAGAT	CGTGGGTGTTGAGACGGTAGG	CGGTCCGAAGGCATCAAGTA
msdRs0323	156–577 (exons 1–3)	GCATCTTCCAGATCCCAAT	CGATGTAGAAGAGCCAGGGAAGTA	TGGGATGTAACAGAGCAAGCG

[a]Two fragments (msdRs0188.1 and msdRs0188.2) were amplified for msdRs0188 for dual verification.

[b]The amplified regions are indexed on template RNAs.

[c]All primers are presented in the direction of 5'→3'.

Table S12. BLASTP search results for four peptides within msdP0188.

Peptide name	Sequence	Result 1	Result 2
msdP0188_pep1	YHDPRGSQLHAGVISTAF	nssf	nhf
msdP0188_pep2	LLPSSWAPVGLLAFLW	nssf	nhf
msdP0188_pep3	LGASASFPAFLQSVVEGLQVLSDKLLET EVSLL	nssf	nhf
msdP0188_pep4	GWQMGSAPGFSGGAMR	nssf	nhf

Result 1 was obtained using the BLASTP server with default parameters for short queries, whereas Result 2 was obtained using the standalone BLASTP program with the local NCBI nr protein database. nssf, no significant similarity found; nhf, no hit found.

Table S13. Summary of qRT-PCR primers for msRNA qualification.

Primer name	Sequence (5' → 3')
hsa-miR-619-5p-RT	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACT GGATACGACGGCTCA
hsa-miR-619-5p-forward	AACAGTGGCTGGGATTACAGG
hsa-miR-4728-3p-RT	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACT GGATACGACCTGGGG
hsa-miR-4728-forward	AACAGACATGCTGACCTCCC
hsa-miR-3135b-RT	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACT GGATACGACCACCAC
hsa-miR-3135b-forward	AATAGTGGCTGGAGCGAGTG
hsa-miR-5585-3p-RT	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACT GGATACGACACCTGT
hsa-miR-5585-3p-forward	AACCGGCTGAATAGCTGGG
U6-RT	GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACT GGATACGACAAAATA
U6- forward	AACAATGTGCTCGCTTCGG
Universal reverse primer	GTCGTATCCAGTGCAGGGT

Other Supplementary Data

Data S1. Summary of unique miRBase human miRNAs (n=2,632) and unique mirGeneDB human mature miRNAs and corresponding miRNA passenger strands (n = 503) used in this study.
Data S2. Unique msdRs predicted using 10–15 nt matches between msRNA and template RNAs.
Data S3. Unique msdPs translated from msdRs.
Data S4. Simulated msRNA artifacts for estimating the false discovery rates for msdR prediction.
Data S5. Summary of the number of msdRs predicted using either database-documented miRNAs or control msRNA artifacts.
Data S6. Summary of 46 msdPs with homologs in the NCBI nr protein database.
Data S7. Summary of six msdPs with identical hit sequences in the NCBI nr protein database.
Data S8. Summary of 1,193 new msdPs.
Data S9. Summary of 941 msdPs matching with mass spectra from cancer cell lines.
Data S10. Summary of 555 msdPs with confident peptide-spectrumes via PepQuery2 search.
Data S11. DNA sequencing source data for msdRs0122 and msdRs0188.
Data S12. Complete mass spectrometry report for msdP0188.
Data S13. Summary of 358 D*x*DGD motif-containing human proteins with >500 amino acids.
Data S14. Summary of 33 human proteins in the PDB with TM-score > 0.5 to the *palm* subdomain of three viral RdRPs.
Data S15. Summary of 84 human proteins in the AlphaFold Protein Structure Database with TM-score > 0.5 to the *palm* subdomain of three viral RdRPs.
Data S1 to S15 are publicly available at https://doi.org/10.5281/zenodo.7826531.