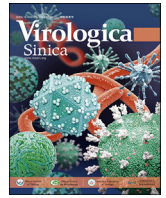




Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.



Research Article

Translation landscape of SARS-CoV-2 noncanonical subgenomic RNAs

Kai Wu^{a,1}, Dehe Wang^{a,1}, Junhao Wang^a, Yu Zhou^{a,b,c,d,*}^a State Key Laboratory of Virology, College of Life Sciences, Wuhan University, Wuhan, 430072, China^b TaiKang Center for Life and Medical Sciences, RNA Institute, Wuhan University, Wuhan, 430072, China^c Institute for Advanced Studies, Wuhan University, Wuhan, 430072, China^d Frontier Science Center for Immunology and Metabolism, Wuhan University, Wuhan, 430072, China

ARTICLE INFO

Keywords:

SARS-CoV-2

Subgenomic RNA (sgRNA)

Mass spectrometry

Translation

RNA binding

ABSTRACT

The ongoing COVID-19 pandemic is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) with a positive-stranded RNA genome. Current proteomic studies of SARS-CoV-2 mainly focus on the proteins encoded by its genomic RNA (gRNA) or canonical subgenomic RNAs (sgRNAs). Here, we systematically investigated the translation landscape of SARS-CoV-2, especially its noncanonical sgRNAs. We first constructed a strict pipeline, named *vi pep*, for identifying reliable peptides derived from RNA viruses using RNA-seq and mass spectrometry data. We applied *vi pep* to analyze 24 sets of mass spectrometry data related to SARS-CoV-2 infection. In addition to known canonical proteins, we identified many noncanonical sgRNA-derived peptides, which stably increase after viral infection. Furthermore, we explored the potential functions of those proteins encoded by noncanonical sgRNAs and found that they can bind to viral RNAs and may have immunogenic activity. The generalized *vi pep* pipeline is applicable to any RNA viruses and these results have expanded the SARS-CoV-2 translation map, providing new insights for understanding the functions of SARS-CoV-2 sgRNAs.

1. Introduction

Coronavirus disease 2019 (COVID-19) is a worldwide pandemic respiratory disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (Graham, 2020; Liu et al., 2020; Zhu et al., 2020). Up to June 2022, SARS-CoV-2 has infected more than 530 million people and caused more than 6 million deaths worldwide (Dong et al., 2020). SARS-CoV-2 is a positive-stranded enveloped RNA virus in the *Betacoronavirus* genus with 29,000 nucleotide (nt) RNA genome which contains a 5'-cap structure and a 3' poly(A) tail (Wu et al., 2020; Zhou et al., 2020).

The genomic RNA (gRNA) of SARS-CoV-2 contains two large open reading frames (ORF1a/1b) encoding 16 viral nonstructural proteins (nsps) through -1 ribosomal frameshifting, and 13 other open reading frames (ORFs) encoding four structural proteins and 9 accessory factors from homologous gene prediction (Chan et al., 2020; Perlman and Netland, 2009; Zhou et al., 2020). Those non-structure proteins play important roles, such as replicating and transcribing the genome (nsp7, nsp8, and nsp12), capping virus RNA (nsp13 and nsp14), cleaving viral polyprotein (3C-like), etc. (Gordon et al., 2020b). Interestingly, proteins other than those in ORF1a/1b are believed to be translated from subgenomic

RNAs (sgRNAs) generated through discontinuous transcription (Kim et al., 2020; Parker et al., 2021; Stern and Kennedy, 1980). This discontinuous transcription step called "template switch", is mainly mediated by the RNA-RNA interactions between the RNA template and the synthesizing RNA sequence flanking the junction sites during sgRNA biogenesis (Wang et al., 2021). Previously, we have defined three types of sgRNAs including leader-type sgRNAs (canonical sgRNAs), ORF1ab-type, and S/N-type sgRNAs (noncanonical sgRNAs) (Wang et al., 2021). Nomburg et al. also analyzed SARS-CoV-2 noncanonical sgRNAs and found that they accounted for 33% of the total sgRNAs in the transcription level, and the proportions increased along the time after infection in cell culture (Nomburg et al., 2020).

With the widespread application of the bioinformatics method in homology comparison and codon analysis, some new ORFs encoded by canonical sgRNAs are continuously discovered, such as ORF3c and ORF3d (Firth, 2020; Jungreis et al., 2021; Nelson et al., 2020). Furthermore, many potential novel proteins encoded by noncanonical sgRNAs have also been found. Finkel et al. observed that some noncanonical sgRNAs may have translational potential based on ribosome footprints (Finkel et al., 2021). Andrew et al. found a novel S glycoprotein with 8 amino acid (aa) deletion and an N protein with 17 aa deletion by mass spectrometry

* Corresponding author.

E-mail address: yu.zhou@whu.edu.cn (Y. Zhou).¹ Kai Wu and Dehe Wang contributed equally to this work.<https://doi.org/10.1016/j.virs.2022.09.003>

Received 5 July 2022; Accepted 1 September 2022

Available online xxxx

1995-820X/© 2022 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

(MS) (Davidson et al., 2020). We also found that two ORF1ab-type sgRNAs have translation potential using a limited number of MS data (Wang et al., 2021). Those sporadic evidence indicated that noncanonical sgRNAs have coding capacity, but a systematic study on noncanonical sgRNAs using large sets of MS data is still missing for SARS-CoV-2. Considering that the sgRNA levels increase along with the viral infection time and are much higher than gRNA, the protein products of sgRNAs and their potential functions are important questions requiring more comprehensive studies.

Here, we developed a general and strict pipeline called *vipep* to systematically identify novel viral RNA-encoded peptides using RNA-seq and mass spectrometry data. We collected 18 sets of RNA-seq data and 24 sets of mass spectrometry data related to SARS-CoV-2 that are currently available. We used *vipep* pipeline to identify reliable peptides from all ORFs in both strands of gRNAs and sgRNAs. We obtained 473 and 53 peptides derived from canonical and noncanonical sgRNAs, respectively. We found that the translation level of noncanonical sgRNAs stably increases at different time points along with viral infection, consistent with the increase at the RNA level. Furthermore, we explored the biological functions of noncanonical sgRNAs-encoded proteins and found that they have the capability of binding with viral RNAs and being immunogenic.

2. Materials and methods

2.1. Data collection and curation

Public mass spectrometry data were downloaded from iProX (Ma et al., 2019) and PRIDE (Perez-Riverol et al., 2019) databases. We manually screened 24 sets of data with SARS-CoV-2 infection (5 in iProX and 19 in PRIDE). The sgRNA junction sites of SARS-CoV-2 were extracted from 18 RNA-seq [9 next-generation sequencing (NGS) and 9 nanopore] datasets using our published method (Wang et al., 2021). Public Ribo-seq data were downloaded from the NCBI SRA database with accession number: SRP260279. The annotated SARS-CoV-2 protein sequences were collected from the NCBI gene reference (NC_045512.2) and UniProt database (version 20210416, 16 entries) (UniProt Consortium, 2021). The domain annotations of those proteins were downloaded from UCSC SARS-CoV-2 Genome Browser (Fernandes et al., 2020).

2.2. Mass spectrometry data analysis

We constructed a complete and non-redundant MS searching database composed of three parts: 1) 6-frame translation for SARS-CoV-2 gRNA which contains annotated ORFs; 2) host proteome from human UniProt (version 2020/09/28 with 192,656 entries) or monkey UniProt (version 2020/10/22 with 19,525 entries) databases, and 3) sgRNA junction-spanning ORFs from the 3-frame translation of SARS-CoV-2 sgRNAs in both forward and reverse strands.

In our *vipep* workflow, the original raw data were divided into tandem mass tag (TMT) and label-free groups according to their experimental design. Then, each group was searched by MaxQuant (1.6.12.0) with corresponding default parameters. A fixed modification of carbamidomethyl (C) and variable modifications of oxidation (M) and acetyl (protein N-term) were included in the search, and the maximum number of missed cleavages was set to 2. For novel peptide identification, the peptide-spectrum match false discovery rate (FDR) threshold was set as 0.01. In order to obtain more reliable results, potential pollutants or reversed peptides were discarded first. The peptides with leucine and isoleucine content higher than 25% were also discarded due to that the two amino acids could not be distinguished by the mass, which may lead to false positives. The remaining peptides detected in single- or multi-datasets were strictly filtered requiring the search score to be larger than 120 or 90, respectively. Only the junction-spanning peptides with overhang equal to or longer than 6 nt at both ends were kept for further analysis.

2.3. SNPs and RNA editing analysis

The human single nucleotide polymorphisms (SNPs) were downloaded from UCSC common dbSNP (version V153, hg38) (Sherry et al., 2001). For all coding isoforms in GENCODE basic annotation (version 41, hg38), the reference base was replaced with the corresponding mutation in the coding sequence (CDS) region to generate the SNP-derived protein sequences that differ from the reference proteome. The human A-to-I editing sites (hg38) were downloaded from the REDIPortal database (Mansi et al., 2021). The RNA-editing-derived protein sequences were constructed similarly as SNPs by replacing A with G in coding regions. The novel SARS-CoV-2 peptides were then scanned against the SNP-mutated and RNA editing-mutated proteins.

2.4. Ribo-seq data analysis

Ribo-seq reads were firstly mapped to the host-virus merged genomes, SARS-CoV-2 gRNA (MN996528) with sgRNA junctions plus Vero E6 (Chlorocebus sabaues Ensembl v99) or Calu-3 (human hg38) genome, using STAR (v2.7.2b) program (Dobin et al., 2013) with parameters “-sjdbFileChrStartEnd JSDB.txt -outFilterMultimapNmax 1 -alignSJoverhangMin 6 -outSJfilterOverhangMin 6 6 6 6 -outSJfilterCountUniqueMin 3 3 3 3 -outSJfilterCountTotalMin 3 3 3 3 -outSJfilterDistToOtherSJmin 0 0 0 0 -scoreGap -4 -scoreGapNoncan -4 -scoreGapATAC -4 -alignIntronMax 30,000 -alignMatesGapMax 30,000 -alignSJstitchMismatchNmax -1 -1 -1 -1”. The mapped reads with length between 27 and 33 nt were used in further analysis.

2.5. RNA-protein interactome MS data analysis

The RNA-protein interactome data was from iProX accession PXD024808. The raw data related to SARS-CoV-2 infection were searched by MaxQuant (1.6.12.0) with default label-free parameters. We divided the data into three groups according to the corresponding probes. Only the peptides appearing in those we identified from the above 24 MS datasets were used in further analysis.

2.6. Epitope prediction

Potential epitopes were predicted by netMHCpan (v4.1) program with default parameters (Reynisson et al., 2020) except for setting the % Rank_EL and %Rank_BA cutoffs as 0.02 to be more stringent. The novel epitopes were defined as those derived from noncanonical sgRNA encoded proteins but not appearing in any of SARS-CoV-2 annotated proteins in Uniprot (version 2021/04 containing 16 entries) and NCBI RefSeq (reference NC_045512.2 containing 38 entries).

3. Results

3.1. Overview of *vipep* pipeline and SARS-CoV-2 proteome

To explore the global landscape of SARS-CoV-2 proteome, we downloaded and analyzed 24 mass spectrometry (MS) datasets with SARS-CoV-2 infection from iProX (Ma et al., 2019) and PRIDE (Perez-Riverol et al., 2019) databases (Fig. 1A and Supplementary Table S1). Within *vipep* pipeline, we first built a comprehensive search database for SARS-CoV-2 including its annotated proteins, additional ORFs from gRNA in-silico translation in 6-frames, predicted junction-spanning ORFs from sgRNAs, and host proteins. The SARS-CoV-2 sgRNA junction sites are from our integrated analysis of multiple NGS and nanopore sequencing data (Wang et al., 2021). We identified peptides in collected MS data using MaxQuant software (Tyanova et al., 2016). After quality control and several strict filtering steps, we keep highly authentic novel SARS-CoV-2 peptides for further downstream analyses (Fig. 1A and Supplementary Fig. S1A).

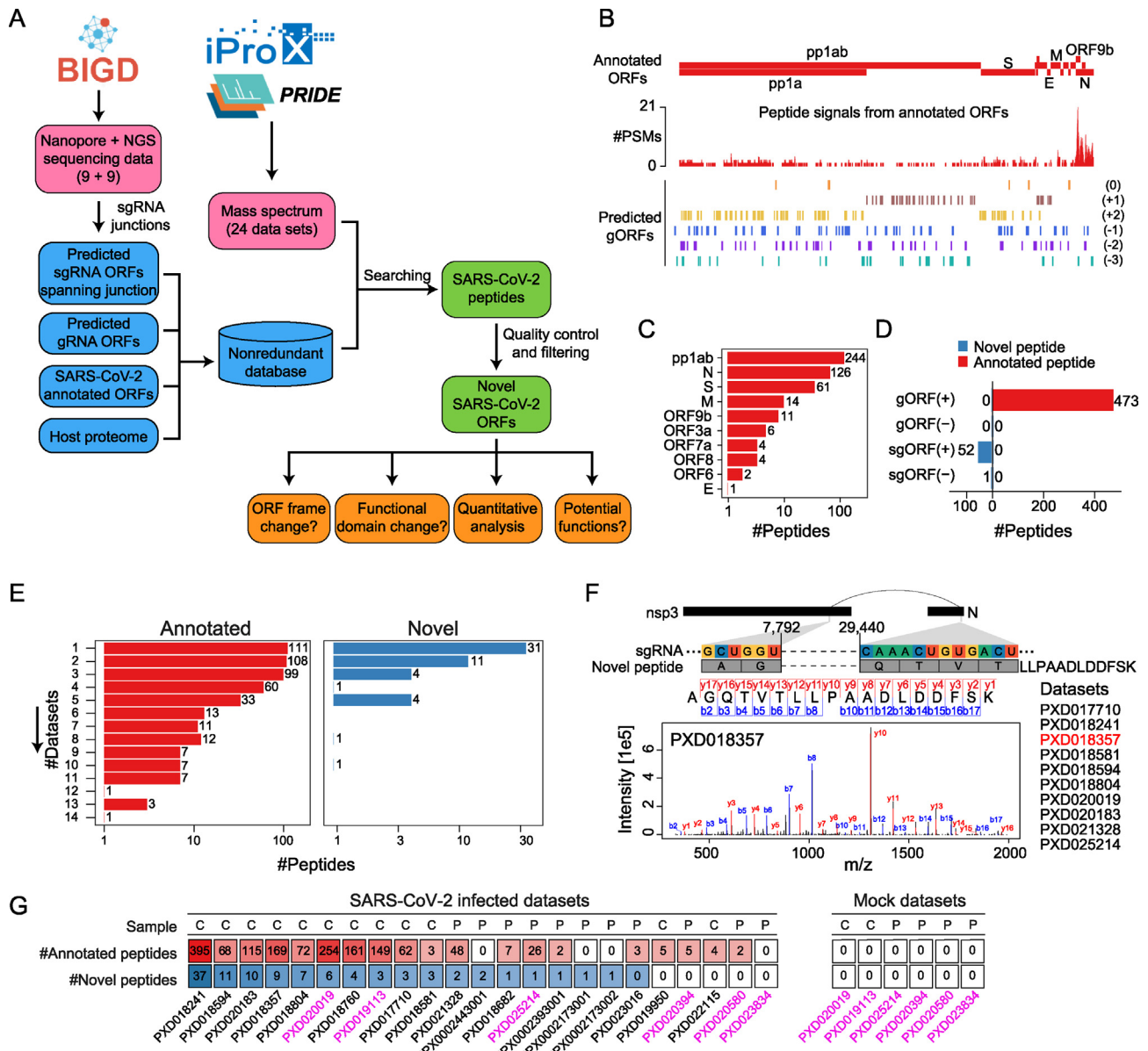


Fig. 1. The vipep pipeline and global landscape of SARS-CoV-2 encoded peptides. **A** Schematic diagram for identifying and annotating novel SARS-CoV-2 encoded peptides. A custom database was built to search SARS-CoV-2 peptides in 24 mass spectrometry datasets currently available in the literature. Pink represents the source of the datasets, blue represents the construction method of the nonredundant mass spectrometry database, green represents the peptides identified by vipep, and orange represents the functional exploration of the novel peptides. **B** Genome browser view of annotated gORFs, numbers of peptide spectrum matches (PSMs) for annotated ORFs, and predicted gORFs by six frames across the SARS-CoV-2 genome. **C** Total counts of peptides detected in all mass spectrometry datasets for the annotated ORFs of SARS-CoV-2. **D** The numbers of annotated and novel peptides derived from genomic or subgenomic ORFs in sense or antisense (\pm) strands. **E** The statistics of peptides supported by different numbers of MS datasets for annotated (left) and novel peptides (right). **F** An example of multi-datasets supported novel peptides spanning a sgRNA junction. The peptide is represented underneath the sgRNA junction spanning two ORFs (top) with the MS/MS spectra found in multiple datasets (bottom). The dataset for the shown spectra is highlighted in red. **G** The numbers of annotated (red) and novel peptides (blue) in all datasets grouped by virus infection and mock samples. The sample origin is labeled for each dataset (C, cell; P, patient). The datasets with mock samples are marked in purple.

We used 194,309 sgRNA junction sites (JSs) found in RNA-seq datasets (Supplementary Fig. S1B) to predict the junction-spanning ORFs that are different from the annotated SARS-CoV-2 proteins. We identified a total of 649 and 440 peptides derived from annotated and novel junction-spanning ORFs, respectively. To minimize false positives, we strictly screened the peptides detected by MS (Supplementary Fig. S1A). First, potential pollutants and reversed peptides were discarded. Second, for a peptide spanning a sgRNA junction, the RNA sequence encoding the peptide was required to have more than 6 nt overhangs at both ends. Third, the leucine and isoleucine content of the peptide should be less than 25% since these two amino acids cannot be

distinguished by molecular weight. Fourth, we set a higher threshold of MaxQuant search score for peptides detected in a single dataset. We found that 92.73% (408/440) of novel peptides only appear in one dataset versus 42.2% of annotated peptides (Supplementary Fig. S1C). Based on the distribution of those search scores for annotated peptides, we determined the threshold as 90 for multiple-sample supported peptides (Supplementary Fig. S1D left). For single-sample supported peptides, we increased the threshold to 120 to reduce potential false positives (Supplementary Fig. S1D right).

Finally, we identified 473 annotated and 53 novel peptides using the vipep workflow (Supplementary Table S2 and Supplementary Table S3).

For annotated ORFs, ORF1ab has the largest number of peptides, followed by N and S proteins (Fig. 1B and C), which may be due to their long lengths and high abundances. Interestingly, no novel peptide was found in any of the predicted gORFs (Fig. 1B) beyond annotated ORFs (Fig. 1D). Although (–)gRNA has many predicted ORFs, none of them has any peptide from MS (Fig. 1D). In principle, there should be no viral peptides in non-infected samples (mock), based on which we evaluated the data quality of all MS datasets. In the MS data IPX0002166001, there are five annotated peptides detected in seven healthy people (Supplementary Fig. S1E). We suspect that some of these healthy people are asymptomatic carriers at the time of taking the samples due to the three healthy ones (H02, H03, and H04) being undifferentiable from the recovered patients (Li et al., 2020), and we have removed this dataset in the following analysis.

All novel peptides derived from sgRNAs are junction-spanning, except two peptides (Supplementary Table S2 and Supplementary Fig. S2A), which are located in a different frame inside the N gene and resulted from different degrees of proteinase digestion (Supplementary Fig. S2A). Expectedly, all but one novel sgRNA-derived peptides are from the sense strand of the viral genome (Fig. 1D and Supplementary Fig. S2B). Considering the limitations of MS technology, not all proteins can be detected by MS depending on the presence of appropriate protease digestion sites. We defined an ORFs as ‘MS theoretical’ if it can produce at least one appropriate enzymolysis peptide fragment within the length of 6–20 aa by trypsin digestion. Of 16 annotated ORFs, 14 can be considered as ‘MS theoretical’ ORFs, and 10 of them can be detected by MS with vipep (Supplementary Fig. S2C left). There are 44 and 94 ‘MS theoretical’ gORFs in sense and negative strands, respectively; however, no novel peptide was detected (Supplementary Fig. S2C middle). There are about 42,000 and 13,000 ‘MS theoretical’ predicted sgORFs, of which there are 732 and 1 sgORFs are detected by MS in sense and negative strands, respectively (Supplementary Fig. S2C right). These proteome-based results suggest that the sense strand of SARS-CoV-2 RNA is indeed more likely to produce proteins, which is consistent with previous works reporting that SARS-CoV-2 showed a high degree of bias and efficiency in generating sense strand RNA (Zhao et al., 2021) and the majority of Ribo-seq reads were derived from sense strand (Puray-Chavez et al., 2022).

We reason that the more frequently observed in different datasets, the higher probability of being a real peptide. Not unexpectedly, about 76.5% of annotated peptides have support from two or more MS datasets (Fig. 1E left). For novel peptides, about 41.5% (22/53) are detected in two or more MS datasets (Fig. 1E right). The low consistency between different datasets could be partially caused by the inclusion of some datasets with very few viral peptides or a specific dataset with very high sensitivity. As a representative example in Fig. 1F, the novel peptide spanning the sgRNA junction (7792–29,440) fusing nsp3 and N protein, is detected in up to ten datasets.

Furthermore, to exclude the possibility that the novel peptides are generated by unannotated host RNAs or newly transcribed RNAs in response to viral infection, we performed 3-frame and 6-frame translations on the host’s transcriptome and genome, respectively. We kept those predicted ORFs with RNA-seq signals and translated them into protein sequences for searching for novel peptides. No novel peptide is identified from those protein sequences of the host. Notably, we searched the novel peptides against the NCBI non-redundant protein database using tblastn, and found that the best hits are all derived from SARS-CoV-2. In addition, we performed the SNPs and RNA editing analysis and confirmed that the novel peptides are neither from the SNP-mutated nor the RNA editing-mutated proteins.

While the mock datasets do not have annotated peptides as controlled in the analysis, they do not contain any novel peptides either (Fig. 1G). Besides, the amount of a detected peptide in the samples derived from viral transfected cells is significantly higher than that in the samples derived from patients. Thus, we believe that the novel peptides we found are reliable based on a set of quality controls,

suggesting that the translation of noncanonical sgRNAs is a widespread phenomenon.

3.2. Global landscape of SARS-CoV-2 sgRNA-derived peptides

To more intuitively display the novel junction-spanning peptides we identified above, we mapped the corresponding sgRNAs in the SARS-CoV-2 genome. We observed that many novel peptides could be derived from multiple sgRNAs (Supplementary Figs. S3A–B). We reasoned that this phenomenon may be caused by the multiple mappings of the coding sequences of the peptides or synonymous codons around the junction sites. For further analysis, we only reserved the 42 single-positional novel peptides (Supplementary Fig. S3C).

Similar to our previous work, we classified the SARS-CoV-2 sgRNAs into four types by the positions of junction sites (JSs): Leader type, ORF1ab/ORF1ab, ORF1ab/S–N, and S–N/S–N (Fig. 2A right) (Wang et al., 2021). The first type is also named canonical sgRNA, while the latter three types are called noncanonical sgRNAs. Although it was reported that the expression levels of canonical sgRNAs are higher than the levels of noncanonical sgRNAs (Kim et al., 2020; Wang et al., 2021), we found that all the novel peptides detected by MS were derived from noncanonical sgRNAs (Fig. 2A left, Fig. 2B), suggesting that canonical sgRNAs do not produce novel proteins. Only two types of noncanonical sgRNAs, ORF1ab/S–N and S–N/S–N, have detected novel peptides (14 and 28 peptides), which are dispersedly observed in different datasets (Fig. 2C). It is worth noting that the dataset PXD018241 (virus-infected cells for 72 h and divided into 20 MS fractions) has significantly higher sensitivity than all other datasets, and 19 novel peptides are only detected in this dataset. In excluding this best dataset, 47.8% (11/23) of novel peptides are repeatedly detectable in two or more datasets.

To further confirm the sgRNA translation for those novel peptides, we downloaded and re-analyzed the Ribo-seq data published by Finkel et al. (2021). Using junction-spanning Ribo-seq reads to verify the translation of junction-spanning peptides, we found that five novel peptides are supported by Ribo-seq signals (Fig. 2D). A representative peptide is shown in Fig. 2E, which is supported by three different Ribo-seq datasets. The results further support the translational capability of novel ORFs at the level of ribosome binding.

3.3. Annotation of sgRNA ORFs containing novel peptides

To explore the characteristics of those translated noncanonical sgRNAs, we classified the novel peptide-originated ORFs into in-frame, frame-shift, and novel types according to whether they can change the coding frame of the corresponding annotated ORFs (Fig. 3A). The in-frame type ORF fuses upstream and downstream annotated ORFs or results from a deletion in one annotated ORF, without frame change. The frame-shift type ORF fuses two parts of annotated ORFs, in which upstream, downstream, or both-end are in a different frame. The novel type ORF contains a sequence that is not annotated as a coding region (Fig. 3A). Two representative ORFs of frame-shift type are shown in Fig. 3B and C, while a novel type ORF is shown in Fig. 3D.

Based on the locations of annotated ORFs, we found that most novel peptides are derived from S–N/S–N type sgRNAs (19 for N–N), followed by two kinds ORF1ab/S–N type sgRNAs (9 for nsp3-N and 5 for nsp2-N, Fig. 3E). The novel peptides derived from in-frame type ORFs were much more than those from the frame-shift type, but no novel peptide was found in both-end frame-shift ORFs (Fig. 3F).

To evaluate the ability of maintaining coding frame of the ORFs containing novel peptides, we counted the RNA-seq junction-spanning (JS) reads from theoretical ORFs and peptide numbers for the novel peptides with both upstream and downstream sub-peptides located in the same annotated ORFs. We collected N–N type peptides for this analysis. We found that a larger proportion of in-frame type novel ORFs can be translated compared to frame-shift type ones ($P = 0.03$, Fig. 3G). This result suggests that noncanonical sgRNAs have a tendency to retain the coding

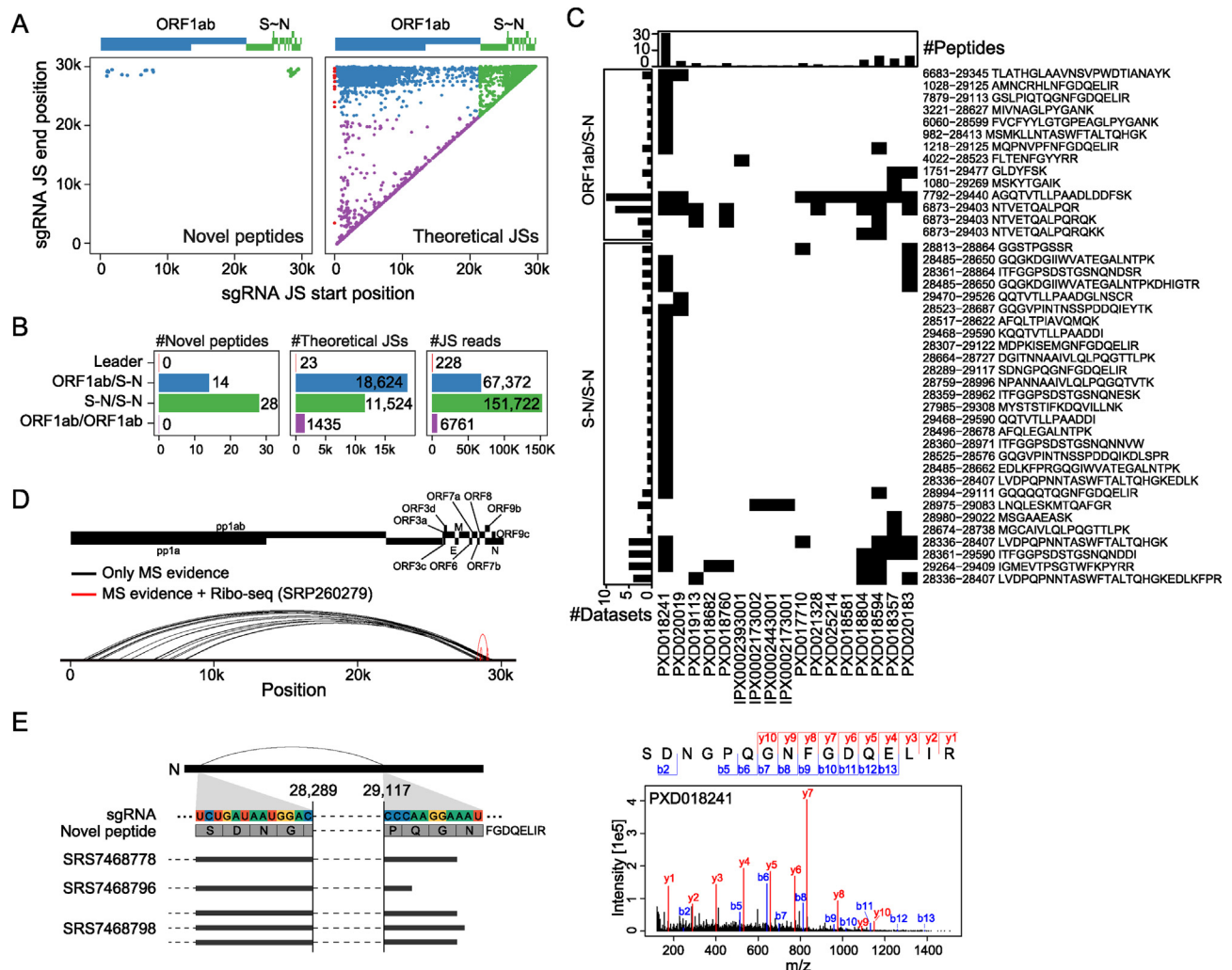


Fig. 2. Global landscape of sgRNA-derived peptides. **A** Global view of sgRNA junctions with mass spectrometry evidence (left) and all theoretical sgRNA junctions (right) in SARS-CoV-2. The canonical Leader-type and 3 noncanonical sgRNA junctions are shown in different colors (Leader, red; ORF1ab/S-N, blue; S-N/S-N, green; ORF1ab/ORF1ab, purple). **B** The statistics of mass spectrometry (MS) peptides, theoretical sgRNA junctions, and the reads spanning theoretical sgRNA junctions by the sgRNA type as in A. **C** The presence (black) or absence (white) heatmap of MS peptides (row) in different datasets (column). The peptides are grouped by sgRNA type as in A. The sgRNA junction positions and peptide sequences are annotated at the right. The total number of peptides in each dataset and the number of supporting datasets for each peptide are shown on the top and on the left side, respectively. **D** Global arc-view of sgRNA junctions with evidence from mass spectrometry only (black) and from both MS and Ribo-seq (red). **E** An exemplar junction-spanning peptide of noncanonical sgRNA with evidence from both Ribo-seq reads. The MS/MS spectra of the peptide are shown on the right.

frame of their corresponding annotated ORFs. It is worth noting that in N-N frame-shift type peptides, five out of seven peptides were downstream frame-shift (Fig. 3E). Interestingly, less than 20% (8 out of 42) of novel peptides used a different upstream frame of corresponding annotated ORFs (Fig. 3F). These results indicate that novel sgRNAs-derived ORFs may have a high probability to share the same start codon with the annotated ORFs.

3.4. A quantitative map of peptides from SARS-CoV-2 sgRNAs

To investigate the dynamics of proteins encoded by noncanonical sgRNAs, we chose a time-course MS dataset (PXD018594) with SARS-CoV-2 infection at two different MOIs (multiplicity of infections). We found that the amounts of both novel peptides from noncanonical sgRNAs and annotated peptides increase along with the infection time (Fig. 4A). By counting the numbers of spectra in specific ORFs, as expected, all annotated proteins increase along with the infection time (Fig. 4B), consistent with a previous report (Bojkova et al., 2020). We also observed the same increasing trend for the two types of noncanonical sgRNA-derived peptides. Interestingly, we found that higher MOI makes it easier to

detect both annotated and novel peptides at an earlier time point, and the levels of both annotated and novel peptides begin to keep stable at the late stage of viral infection, independent of virus MOI (Fig. 4B). These results reveal the reliability of novel peptides and indicate that the noncanonical sgRNAs can be translated and are regulated in the same manner as canonical sgRNAs.

To estimate the relative proportion of noncanonical sgRNAs-derived proteins in the viral proteome, we devised a computational scheme by using junction-spanning peptides to represent novel proteins and using continuous junction-covering peptides to represent annotated proteins (Fig. 4C). The rationale is due to the caveat: the continuous peptides not covering the junctions may be derived from annotated ORFs or non-canonical sgORFs. As shown in Fig. 4D (using dataset PXD018594), the relative abundance of novel peptides reached a maximum of around 13% of viral proteome using ORF9b and N as representative ORFs. Furthermore, we extended the same analysis to the top 5 datasets with the largest number of novel peptides, and found that the average abundances of novel peptides are similar (Fig. 4E). These results imply that the translation of noncanonical sgRNAs was widespread and reached a considerable proportion level.

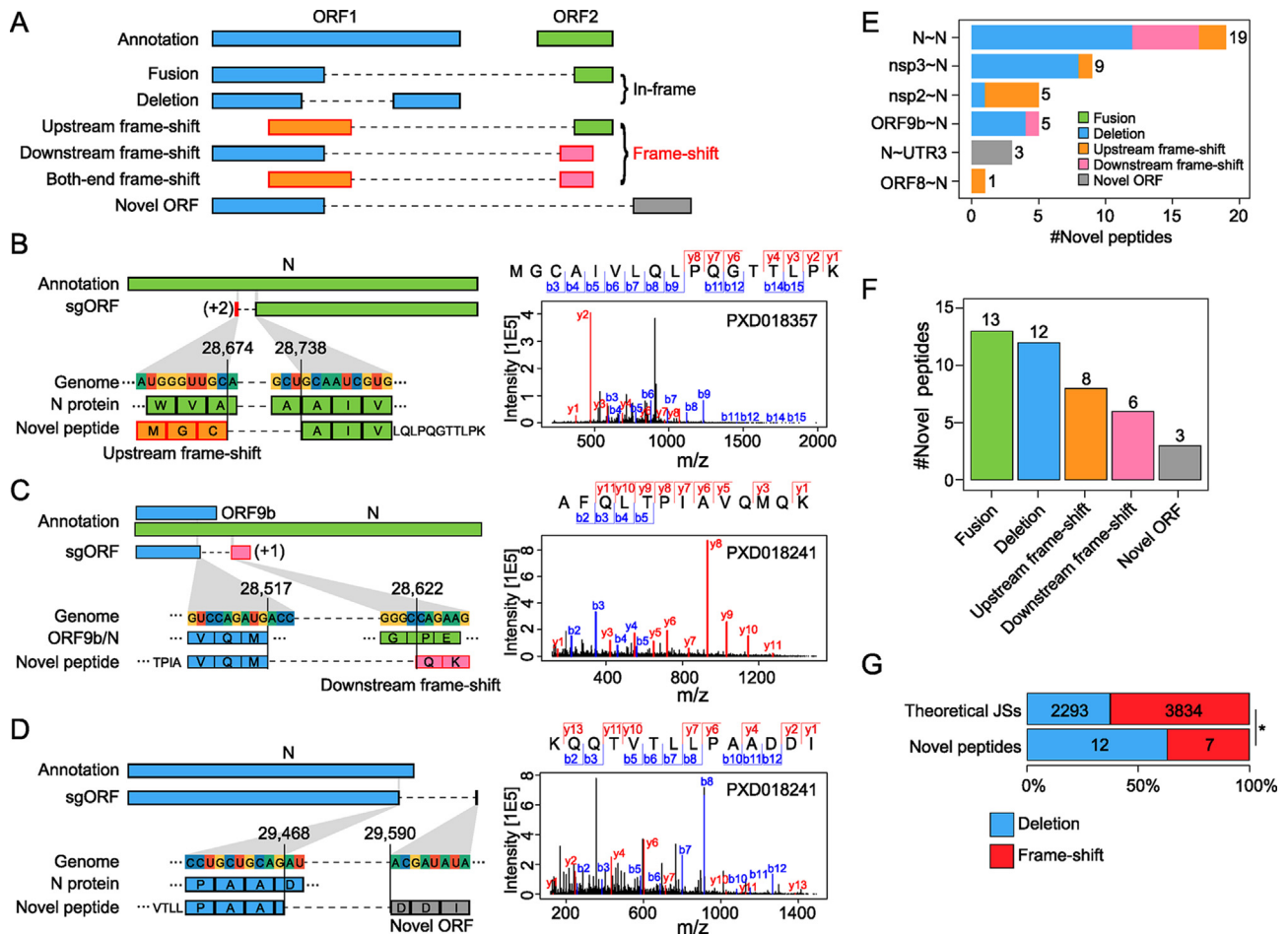


Fig. 3. Characterization of the sgRNA-derived peptides. **A** Graphical illustration of the classification for peptides derived from sgRNAs. ORF1 and ORF2 represent two different ORFs or two terminals of the same protein. Blue represents the upstream non-changed ORF, green represents the downstream non-changed ORF, orange represents the upstream ORF with frame-shift, pink represents the downstream ORF with frame-shift, and grey represents the ORF located in unannotated regions. **B–D** Representative examples for three types of peptides spanning sgRNA junctions (**B**, Upstream frame-shift; **C**, Downstream frame-shift; **D**, Novel). The spanned ORF(s) and the sgRNA ORF are shown at the top. The junction position, RNA and protein sequences of annotated ORF(s), and novel peptide sequence are presented at the bottom. The MS/MS spectra with both y and b ion information for the peptide are shown on the right. **E** The statistics of novel peptides grouped by sgRNA junction spanned ORF(s). **F** The statistics of novel peptides by peptide type as in **A**. **G** The numbers of RNA-seq junction-spanning (JS) reads and detected novel peptides from theoretical noncanonical sgRNA ORFs with deletion and frame-shift inside the same gene (N–N type). The statistical significance is based on two sides Fisher's exact test ($P = 0.03$).

Together, these findings revealed a global and wide translation phenomenon for SARS-CoV-2 noncanonical sgRNAs, and the stable level of novel peptides indicates that the translated proteins may perform certain functions in infected cells.

3.5. Functional potentials of sgRNA-encoded proteins

To further explore the biological functions of noncanonical sgRNAs, we used recently published RNA-protein interactome data to explore the potential RNA binding ability for sgRNA-derived proteins. This dataset used two separate pools of RNA probes targeting the regions of ORF1ab and S–N, respectively (Fig. 5A top). The ORF1ab probes are expected to specifically hybridize with the genomic RNA while S–N probes are expected to enrich both the genomic RNA and sgRNAs (Lee et al., 2021). Using the novel peptides we identified above, we found that 13 novel peptides are immunoprecipitated by the RNA probes. For the 10 single-positional peptides, 5 peptides are detected in both groups of RNA probes, and 5 peptides are detected in only one group of probes (Fig. 5A bottom). Not unexpectedly, many annotated proteins are immunoprecipitated by both groups of RNA probes (Fig. 5B left), and similarly, most

novel peptides (7 out of 13) are shared in the two groups of RNA probes (Fig. 5B). From the detailed analysis of each peptide, we found that the amounts of novel peptides in the two groups of SARS-CoV-2 probes are much higher than that of the rRNA probe control, similar to the annotated peptides (Fig. 5C). Expectedly, the mock control samples do not have any peptides encoded by SARS-CoV-2 RNAs. These results indicate that the proteins encoded by noncanonical sgRNAs have the potential to bind with SARS-CoV-2 RNAs.

Previous studies have shown that T cell responses play an essential role in SARS-CoV-2 immunity and viral clearance (Altmann and Boyton, 2020; Grifoni et al., 2020; Le Bert et al., 2020). In a recent study, Shira et al. found that HLA-I peptides derived not only from canonical ORFs but also from internal out-of-frame ORFs which are not captured by current vaccines (Weingarten-Gabbay et al., 2021). We thus predicted the potential antigenic peptides derived from SARS-CoV-2 proteins. After passing a strict binding score cutoff, a total of 12 novel neo-antigenic epitope clusters from nine noncanonical sgRNAs were identified (Supplementary Fig. S4A and Supplementary Table S4). Together, these data suggest that noncanonical sgRNAs-derived proteins may function in immunophysiology.

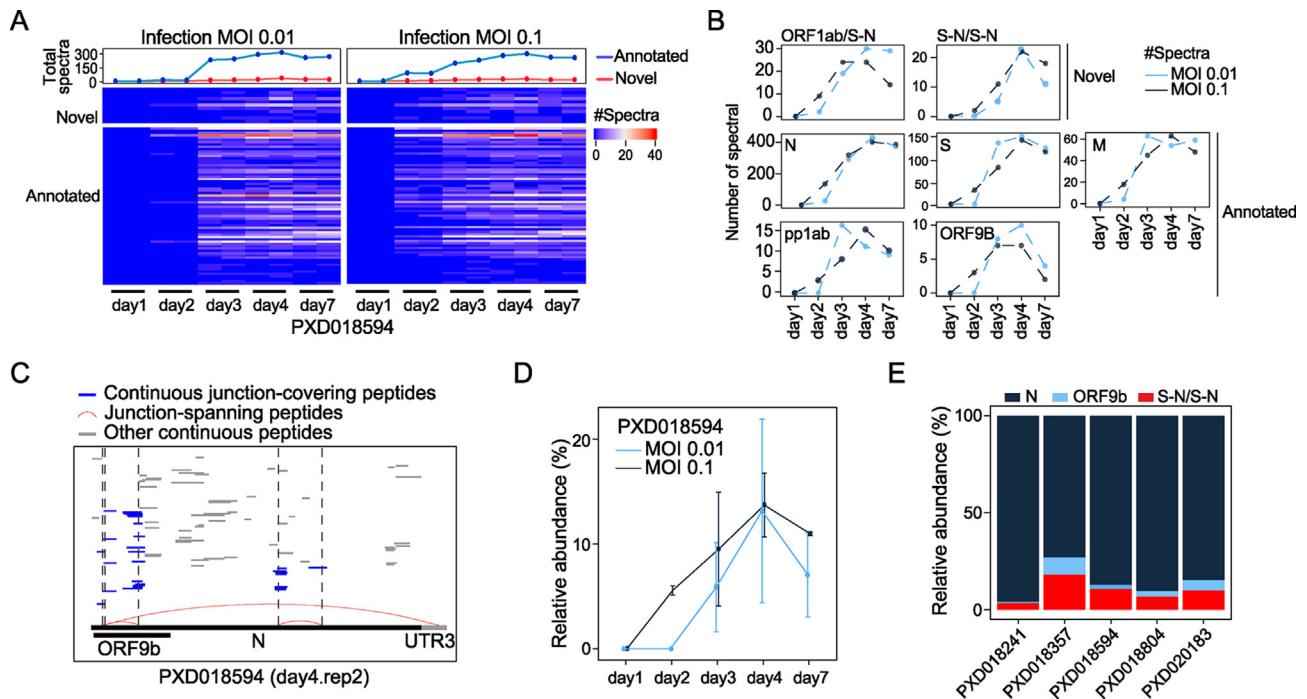


Fig. 4. The increasing dynamics of viral peptides after SARS-CoV-2 infection. **A** Heatmap of the spectra counts for novel and annotated SARS-CoV-2 peptides at different time points in dataset PXD020019. The total spectra counts for each time point are shown on the top. **B** The counts of novel sgRNA-derived peptides and peptides in different annotated ORFs. The counts are summed by spectra with different virus MOI. **C** Graphical illustration of the continuous junction-covering peptides (blue), junction-spanning peptides (red), and other continuous peptides (grey) located in *ORF9b* and *N* genes. All peptides are from the dataset PXD018594 (day4.rep2). **D** Estimated proportions of novel peptides at different time points as in **C**. **E** Estimated proportions of the novel (red) and annotated (blue) peptides by spectra in five different datasets as in **C**.

4. Discussion

Many noncanonical sgRNAs of SARS-CoV-2 have been identified (Kim et al., 2020; Nomburg et al., 2020; Wang et al., 2021), but only a few studies mentioned that they have translation ability (Davidson et al., 2020; Finkel et al., 2021; Wang et al., 2021). This study combined RNA-seq and mass spectrometry data to explore the translation landscape of SARS-CoV-2. We found that many noncanonical sgRNAs could be

translated into proteins under strict criteria. The proteins are stably existing and may play important roles in the same manner as annotated proteins, such as binding to viral RNAs.

The SARS-CoV-2 noncanonical sgRNA-derived proteins can be regarded as structural variations of corresponding annotated proteins due to deletions or frame-shifts. We further evaluated the functional potentials of those proteins at the domain level. Some deletion-type novel proteins lack complete functional domains. For example, the

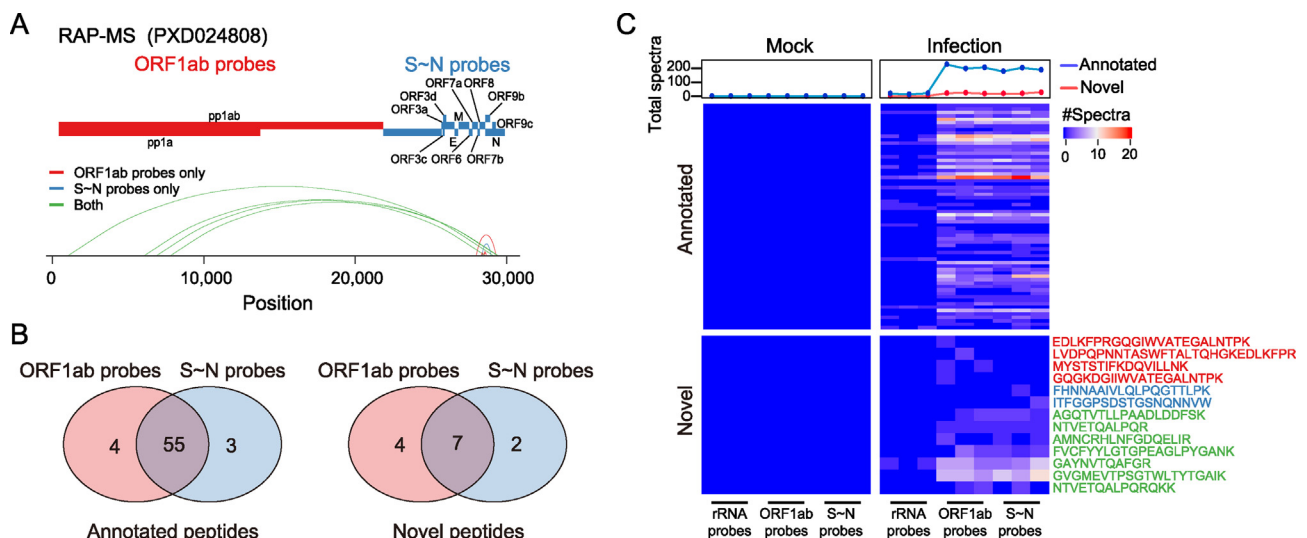


Fig. 5. The functional potentials of noncanonical sgRNA-derived proteins. **A** Global arc-view of unique-positional novel peptides detected in RAP-MS dataset PXD024808. The targeted regions by two groups of probes are marked in red and blue at the top, and the detected sgRNA junction-spanning peptides are shown at the bottom. **B** Venn diagram of detected peptides between the two types of probes in **C** for annotated (left) and novel peptides (right). **C** Spectra counts heatmap of annotated and novel peptides in dataset PXD024808. The total spectra counts for different time points are shown on the top, and the sequences of novel peptides are shown on the right side colored as in **A**.

noncanonical sgRNA (28,359–28962) derived protein lacks an RNA-binding domain relative to its annotated N protein (Supplementary Fig. S4B), which may generate a novel smaller protein losing the RNA binding activity (Banerjee et al., 2020) and cannot participate in the biological function of forming phase separation complex (Lu et al., 2021), whereas its dimerization domain may allow this protein to retain the function of interacting with other proteins (Gordon et al., 2020a; Lu et al., 2021). For fusion-type noncanonical sgRNAs-derived proteins, the combination of functional domains in different annotated proteins may produce new biological functional proteins. For example, the sgRNA (6060–28599) derived protein was a fusion protein combining the macro and peptidase C16 domains in the nsp3 protein with the dimerization domain in the N protein (Supplementary Fig. S4C).

It is worth noting that current mass spectrometry technology has certain limitations, such as low sensitivity to detect proteins of low abundance, and the inability to detect all kinds of proteins due to dependence on the presence of appropriate protease digestion sites. Thus, noncanonical sgRNAs-derived proteins seem difficult to be fully detected.

The COVID-19 caused by SARS-CoV-2 has been in pandemic for more than two years. One reason for unable to end the pandemic is that the virus is continuously mutating (Kaushik, 2021; Sharma et al., 2021; Varahachalam et al., 2021). Some new variants, such as delta and omicron variants, have higher transmissibility and immune evasion than the early original virus strain (Fisman et al., 2020; Funk et al., 2021; Kim et al., 2021; Thomas et al., 2021). Currently, researchers focus on the mutations which change the spike protein of SARS-CoV-2 and affect viral transmissibility, pathogenicity, and immune escape (Harvey et al., 2021; Tian et al., 2021). Nevertheless, we hypothesize that many variants in non-coding regions may change the canonical and noncanonical sgRNAs when they break or gain the RNA-RNA interactions that regulate sgRNA biogenesis (Wang et al., 2021), such that novel proteins encoded by noncanonical sgRNAs may play important roles in regulating viral pathogenicity. We anticipate having more MS data, especially under viral infection of different variant strains, to further investigate noncanonical sgRNAs-encoded proteins.

Currently, the novel peptides identified from the computational analysis of MS data lack additional biochemical experimental data. We propose two possible experimental methods. First, we may use the novel peptides to generate specific antibodies for validation with Western blotting. Second, we may try to tag a GFP to the N gene in a SARS-CoV-2 pseudovirus, and then enrich proteins translated from non-canonical N sgRNAs for validation with mass spectrometry. However, both strategies are challenging and time-consuming, which thus merit a further separate study, to complement our current computational study.

5. Conclusions

We developed a general vipep pipeline for RNA viruses to analyze mass spectrum datasets and constructed the translation landscape of SARS-CoV-2. We identified many peptides translated from noncanonical sgRNAs. Those novel peptides are dynamically regulated in the same manner as annotated proteins, suggesting that they are stably existing and functional. The novel proteins may play important roles by combining different domains or losing domains in annotated proteins. Some of the proteins are found to bind with viral RNAs or are predicted to have immunogenic activity. Our results expand the SARS-CoV-2 translation proteome and indicate that noncanonical sgRNAs-derived proteins are a non-negligible component meriting further studies.

Data availability

All data relevant to the study are included in the article or uploaded as supplementary information.

Ethics statement

This article does not contain any studies with human or animal subjects performed by any of the authors.

Author contributions

Kai Wu: conceptualization, formal analysis, methodology, writing-original draft preparation. Dehe Wang: conceptualization, formal analysis, methodology. Junhao Wang: downloaded and organized the MS datasets. Yu Zhou: conceptualization, data curation, funding acquisition, methodology, supervision, validation, writing – review & editing.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by grants from the Hubei Provincial Natural Science Foundation of China (2020CFA057) and the China NSFC projects (31922039 and 31871316) to Y.Z. Part of the computation in this work was done on the supercomputing system in the Supercomputing Center of Wuhan University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.virs.2022.09.003>.

References

- Altmann, D.M., Boyton, R.J., 2020. SARS-CoV-2 T cell immunity: specificity, function, durability, and role in protection. *Sci. Immunol.* 5, eabd6160.
- Banerjee, A.K., Blanco, M.R., Bruce, E.A., Honson, D.D., Chen, L.M., Chow, A., Bhat, P., Ollikainen, N., Quinodoz, S.A., Loney, C., Thai, J., Miller, Z.D., Lin, A.E., Schmidt, M.M., Stewart, D.G., Goldfarb, D., De Lorenzo, G., Rihn, S.J., Voorhees, R.M., Botten, J.W., Majumdar, D., Guttman, M., 2020. SARS-CoV-2 disrupts splicing, translation, and protein trafficking to suppress host defenses. *Cell* 183, 1325–1339 e21.
- Bojkova, D., Klann, K., Koch, B., Widera, M., Krause, D., Ciesek, S., Cinatl, J., Münch, C., 2020. Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* 583, 469–472.
- Chan, J.F.W., Kok, K.H., Zhu, Z., Chu, H., To, K.K.W., Yuan, S., Yuen, K.Y., 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microb. Infect.* 9, 221–236.
- Davidson, A.D., Williamson, M.K., Lewis, S., Shoemark, D., Carroll, M.W., Heesom, K.J., Zambon, M., Ellis, J., Lewis, P.A., Hiscox, J.A., Matthews, D.A., 2020. Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. *Genome Med.* 12, 68.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* 29, 15–21.
- Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* 20, 533–534.
- Fernandes, J.D., Hinrichs, A.S., Clawson, H., Gonzalez, J.N., Lee, B.T., Nassar, L.R., Raney, B.J., Rosenbloom, K.R., Nerli, S., Rao, A.A., Schmelzer, D., Fyfe, A., Maulding, N., Zweig, A.S., Lowe, T.M., Ares, M., Corbet-Detig, R., Kent, W.J., Haussler, D., Haussler, M., 2020. The UCSC SARS-CoV-2 genome browser. *Nat. Genet.* 52, 991–998.
- Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., Tamir, H., Achdout, H., Stein, D., Israeli, O., Beth-Din, A., Melamed, S., Weiss, S., Israely, T., Paran, N., Schwartz, M., Stern-Ginossar, N., 2021. The coding capacity of SARS-CoV-2. *Nature* 589, 125–130.
- Firth, A.E., 2020. A putative new SARS-CoV protein, 3c, encoded in an ORF overlapping ORF3a. *J. Gen. Virol.* 101, 1085–1089.
- Fisman, D.N., Bogoch, I., Lapointe-Shaw, L., McCready, J., Tuite, A.R., 2020. Risk factors associated with mortality among residents with coronavirus disease 2019 (COVID-19) in long-term care facilities in Ontario, Canada. *JAMA Netw. Open* 3, e2015957.
- Funk, T., Pharris, A., Spiteri, G., Bundle, N., Melidou, A., Carr, M., Gonzalez, G., Garcia-Leon, A., Crispie, F., O'Connor, L., Murphy, N., Mossong, J., Vergison, A., Wienecke-

- Baldacchino, A.K., Abdelrahman, T., Riccardo, F., Stefanelli, P., Di Martino, A., Bella, A., Lo Presti, A., Casaca, P., Moreno, J., Borges, V., Isidro, J., Ferreira, R., Gomes, J.P., Dotsenko, L., Suija, H., Epstein, J., Sadikova, O., Sepp, H., Ikonen, N., Savolainen-Kopra, C., Blomqvist, S., Mötönen, T., Helve, O., Gomes-Dias, J., Adlhoj, C., Covid study groups, 2021. Characteristics of SARS-CoV-2 variants of concern B.1.1.7, B.1.351 or P.1: data from seven EU/EEA countries, weeks 38/2020 to 10/2021. *Euro Surveill.* 26, 2100348.
- Gordon, D.E., Hiatt, J., Bouhaddou, M., Rezelj, V.V., Ulferts, S., Braberg, H., Jureka, A.S., Obernier, K., Guo, J.Z., Batra, J., Kaake, R.M., Weckstein, A.R., Owens, T.W., Gupta, M., Pourmal, S., Titus, E.W., Cakir, M., Soucheray, M., McGregor, M., Cakir, Z., Jang, G., O'Meara, M.J., Tummino, T.A., Zhang, Z., Foussard, H., Rojic, A., Zhou, Y., Kuchenov, D., Hüttenhain, R., Xu, J., Eckhardt, M., Swaney, D.L., Fabius, J.M., Ummadi, M., Tutuncuoglu, B., Rathore, U., Modak, M., Haas, P., Haas, K.M., Naing, Z.Z.C., Pulido, E.H., Shi, Y., Barrio-Hernandez, I., Memon, D., Petsalaki, E., Dunham, A., Marrero, M.C., Burke, D., Koh, C., Vallet, T., Silvas, J.A., Azumaya, C.M., Billesbølle, C., Brilot, A.F., Campbell, M.G., Diallo, A., Dickinson, M.S., Diwanji, D., Herrera, N., Hoppe, N., Kratochvil, H.T., Liu, Y., Merz, G.E., Moritz, M., Nguyen, H.C., Nowotny, C., Puchades, C., Rizo, A.N., Schulze-Gahmen, U., Smith, A.M., Sun, M., Young, I.D., Zhao, J., Asarnow, D., Biel, J., Bowen, A., Braxton, J.R., Chen, J., Chio, C.M., Chio, U.S., Deshpande, I., Doan, L., Faust, B., Flores, S., Jin, M., Kim, K., Lam, V.L., Li, F., Li, J., Li, Y.-L., Li, Y., Liu, X., Lo, M., Lopez, K.E., Melo, A.A., Moss, F.R., Frost, A., Nguyen, P., Paulino, J., Pawar, K.I., Peters, J.K., Pospiech, T.H., Safari, M., Sangwan, S., Schaefer, K., Thomas, P.V., Thwin, A.C., Trenker, R., Tse, E., Tsui, T.K.M., Wang, F., Whitis, N., Yu, Z., Zhang, K., Zhang, Y., Zhou, F., Saltzberg, D., Qcrg Structural Biology Consortium, Hodder, A.J., Shun-Shion, A.S., Williams, D.M., White, K.M., Rosales, R., Kehrer, T., Miorin, L., Moreno, E., Patel, A.H., Rihn, S., Khalid, M.M., Vallejo-Gracia, A., Fozoumi, P., Simoneau, C.R., Roth, T.L., Wu, D., Karim, M.A., Ghousaini, M., Dunham, I., Berardi, F., Weigang, S., Chazal, M., Park, J., Logue, J., McGrath, M., Weston, S., Haupt, R., Hastie, C.J., Elliott, M., Brown, F., Burness, K.A., Reid, E., Dorward, M., Johnson, C., Wilkinson, S.G., Geyer, A., Giesel, D.M., Baillie, C., Raggatt, S., Leech, H., Toth, R., Goodman, N., Keough, K.C., Lind, A.L., Zoonomia Consortium, Klesh, R.J., Hemphill, K.R., Carlson-Stevermer, J., Oki, J., Holden, K., Maures, T., Pollard, K.S., Sali, A., Agard, D.A., Cheng, Y., Fraser, J.S., Frost, A., Jura, N., Kortemme, T., Manglik, A., Southworth, D.R., Stroud, R.M., Alessi, D.R., Davies, P., Frieman, M.B., Ideker, T., Abate, C., Jouvenet, N., Kochs, G., Shoichet, B., Ott, M., Palmarini, M., Shokat, K.M., Garcia-Sastre, A., Rassen, J.A., Grosche, R., Rosenberg, O.S., Verba, K.A., Basler, C.F., Vignuzzi, M., Peden, A.A., Beltrao, P., Krogan, N.J., 2020a. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* 370, eabe9403.
- Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O'Meara, M.J., Rezelj, V.V., Guo, J.Z., Swaney, D.L., Tummino, T.A., Hüttenhain, R., Kaake, R.M., Richards, A.L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., Kim, M., Haas, P., Polacco, B.J., Braberg, H., Fabius, J.M., Eckhardt, M., Soucheray, M., Bennett, M.J., Cakir, M., McGregor, M.J., Li, Q., Meyer, Z.B., Roesch, F., Vallet, T., Mac Kain, A., Miorin, L., Moreno, E., Naing, Z.Z.C., Zhou, Y., Peng, S., Shi, Y., Zhang, Z., Shen, W., Kirby, I.T., Melnyk, J.E., Chhorba, J.S., Lou, K., Dai, S.A., Barrio-Hernandez, I., Memon, D., Hernandez-Armenta, C., Lyu, J., Mathy, C.J.P., Perica, T., Pilla, K.B., Ganesan, S.J., Saltzberg, D.J., Rakesh, R., Liu, X., Rosenthal, S.B., Calviello, L., Venkataramanan, S., Liboy-Lugo, J., Lin, Y., Huang, X.-P., Liu, Y., Wankowicz, S.A., Bohn, M., Safari, M., Ugur, F.S., Koh, C., Saver, N.S., Tran, Q.D., Shengjuler, D., Fletcher, S.J., O'Neal, M.C., Cai, Y., Chang, J.C.J., Broadhurst, D.J., Klippstein, S., Sharp, P.P., Wenzell, N.A., Kuzuoglu-Ozturk, D., Wang, H.-Y., Trenker, R., Young, J.M., Caverro, D.A., Hiatt, J., Roth, T.L., Rathore, U., Subramanian, A., Noack, J., Hubert, M., Stroud, R.M., Frankel, A.D., Rosenberg, O.S., Verba, K.A., Agard, D.A., Ott, M., Emerman, M., Jura, N., von Zastrow, M., Verdine, E., Ashworth, A., Schwartz, O., d'Enfert, C., Mukherjee, S., Jacobson, M., Malik, H.S., Fujimori, D.G., Ideker, T., Craik, C.S., Floor, S.N., Fraser, J.S., Gross, J.D., Sali, A., Roth, B.L., Ruggiero, D., Taunton, J., Kortemme, T., Beltrao, P., Vignuzzi, M., Garcia-Sastre, A., Shokat, K.M., Shoichet, B.K., Krogan, N.J., 2020b. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 583, 459–468.
- Graham, F., 2020. Daily briefing: WHO describes coronavirus as a pandemic to spur countries to action. *Nature*. <https://doi.org/10.1038/d41586-020-00732-y>.
- Grifoni, A., Weiskopf, D., Ramirez, S.L., Mateus, J., Dan, J.M., Moderbacher, C.R., Rawlings, S.A., Sutherland, A., Premkumar, L., Jhadi, R.S., Maramba, D., de Silva, A.M., Frazier, A., Carlin, A.F., Greenbaum, J.A., Peters, B., Krammer, F., Smith, D.M., Crotty, S., Sette, A., 2020. Targets of T Cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell* 181, 1489–1501 e15.
- Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., , Covid-19 Genomics UK (COG-UK) Consortium, Peacock, S.J., Robertson, D.L., 2021. SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* 19, 409–424.
- Jungreis, I., Sealfon, R., Kellis, M., 2021. SARS-CoV-2 gene content and COVID-19 mutation impact by comparing 44 Sarbecovirus genomes. *Nat. Commun.* 12, 2642.
- Kaushik, A., 2021. Manipulative magnetic nanomedicine: the future of COVID-19 pandemic/epidemic therapy. *Expet Opin. Drug Deliv.* 18, 531–534.
- Kim, D., Lee, J.Y., Yang, J.S., Kim, J.W., Kim, V.N., Chang, H., 2020. The architecture of SARS-CoV-2 transcriptome. *Cell* 181, 914–921 e10.
- Kim, Sinae, Nguyen, T.T., Taitt, A.S., Jhun, H., Park, H.Y., Kim, S.H., Kim, Y.G., Song, E.Y., Lee, Y., Yum, H., Shin, K.-C., Choi, Y.K., Song, C.-S., Yeom, S.C., Kim, B., Netea, M., Kim, Soohyun, 2021. SARS-CoV-2 omicron mutation is faster than the chase: multiple mutations on spike/ACE2 interaction residues. *Immune Netw* 21, e38.
- Le Bert, N., Tan, A.T., Kunasegaran, K., Tham, C.Y.L., Hafezi, M., Chia, A., Chng, M.H.Y., Lin, M., Tan, N., Linster, M., Chia, W.N., Chen, M.I.C., Wang, L.F., Ooi, E.E., Kalimuddin, S., Tambayah, P.A., Low, J.G.H., Tan, Y.J., Bertoletti, A., 2020. SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature* 584, 457–462.
- Lee, S., Lee, Y.-S., Choi, Y., Son, A., Park, Y., Lee, K.M., Kim, J., Kim, J.S., Kim, V.N., 2021. The SARS-CoV-2 RNA interactome. *Mol. Cell* 81, 2838–2850 e6.
- Li, Y., Wang, Y., Liu, H., Sun, W., Ding, B., Zhao, Y., Chen, P., Zhu, L., Li, Z., Li, N., Chang, L., Wang, H., Bai, C., Xu, P., 2020. Urine proteome of COVID-19 patients. *URINE* 2, 1–8.
- Liu, W., Zhang, Q., Chen, J., Xiang, R., Song, H., Shu, S., Chen, L., Liang, L., Zhou, J., You, L., Wu, P., Zhang, B., Lu, Y., Xia, L., Huang, L., Yang, Y., Liu, F., Semple, M.G., Cowling, B.J., Lan, K., Sun, Z., Yu, H., Liu, Y., 2020. Detection of covid-19 in children in early january 2020 in wuhan, China. *N. Engl. J. Med.* 382, 1370–1371.
- Lu, S., Ye, Q., Singh, D., Cao, Y., Diedrich, J.K., Yates, J.R., Villa, E., Cleveland, D.W., Corbett, K.D., 2021. The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat. Commun.* 12, 502.
- Ma, J., Chen, T., Wu, S., Yang, C., Bai, M., Shu, K., Li, K., Zhang, G., Jin, Z., He, F., Hermjakob, H., Zhu, Y., 2019. iProX: an integrated proteome resource. *Nucleic Acids Res.* 47, D1211–D1217.
- Mansi, L., Tangaro, M.A., Lo Giudice, C., Flati, T., Kopel, E., Schaffer, A.A., Castrignano, T., Chillemi, G., Pesole, G., Picardi, E., 2021. REDportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments. *Nucleic Acids Res.* 49, D1012–D1019.
- Nelson, C.W., Ardern, Z., Goldberg, T.L., Meng, C., Kuo, C.H., Ludwig, C., Kolokotronis, S.O., Wei, X., 2020. Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic. *Life* 9, e59633.
- Nomburg, J., Meyerson, M., DeCaprio, J.A., 2020. Pervasive generation of non-canonical subgenomic RNAs by SARS-CoV-2. *Genome Med.* 12, 108.
- Parker, M.D., Lindsey, B.B., Leary, S., Gaudieri, S., Chopra, A., Wyles, M., Angyal, A., Green, L.R., Parsons, P., Tucker, R.M., Brown, R., Groves, D., Johnson, K., Carrilero, L., Heffer, J., Partridge, D.G., Evans, C., Raza, M., Keeley, A.J., Smith, N., Filipe, A.D.S., Shepherd, J.G., Davis, C., Bennett, S., Sreenu, V.B., Kohl, A., Aranday-Cortes, E., Tong, L., Nichols, J., Thomson, E.C., , Covid-19 Genomics UK (COG-UK) Consortium, Wang, D., Mallal, S., de Silva, T.I., 2021. Subgenomic RNA identification in SARS-CoV-2 genomic sequencing data. *Genome Res.* 31, 645–658.
- Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Linares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Pérez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yilmaz, S., Tiwary, S., Cox, J., Audain, E., Walzer, M., Jarnczak, A.F., Ternent, T., Brazma, A., Vizcaíno, J.A., 2019. The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* 47, D442–D450.
- Pearlman, S., Netland, J., 2009. Coronaviruses post-SARS: update on replication and pathogenesis. *Nat. Rev. Microbiol.* 7, 439–450.
- Puray-Chavez, M., Lee, N., Tenneti, K., Wang, Y., Vuong, H.R., Liu, Y., Horani, A., Huang, T., Gunsten, S.P., Case, J.B., Yang, W., Diamond, M.S., Brody, S.L., Dougherty, J., Kutluay, S.B., 2022. The translational landscape of SARS-CoV-2-infected cells reveals suppression of innate immune genes. *mBio* 13, e0081522.
- Reynisson, B., Alvarez, B., Paul, S., Peters, B., Nielsen, M., 2020. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* 48, W449–W454.
- Sharma, P.K., Kim, E.S., Mishra, S., Ganbold, E., Seong, R.S., Kaushik, A.K., Kim, N.Y., 2021. Ultrasensitive and reusable graphene oxide-modified double-interdigitated capacitive (DIDC) sensing chip for detecting SARS-CoV-2. *ACS Sens.* 6, 3468–3476.
- Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Skitkin, K., 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311.
- Stern, D.F., Kennedy, S.I., 1980. Coronavirus multiplication strategy. I. Identification and characterization of virus-specified RNA. *J. Virol.* 34, 665–674.
- Thomas, E., Delabot, S., Carattini, Y.L., Andrews, D.M., 2021. SARS-CoV-2 and variant diagnostic testing approaches in the United States. *Viruses* 13, 2492.
- Tian, D., Sun, Y., Zhou, J., Ye, Q., 2021. The global epidemic of the SARS-CoV-2 delta variant, key spike mutations and immune escape. *Front. Immunol.* 12, 751778.
- Tyanova, S., Temu, T., Cox, J., 2016. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* 11, 2301–2319.
- UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49, D480–D489.
- Varahachalam, S.P., Lahooti, B., Chamaneh, M., Bagchi, S., Chhibber, T., Morris, K., Bolanos, J.F., Kim, N.Y., Kaushik, A., 2021. Nanomedicine for the SARS-CoV-2: state-of-the-art and future prospects. *Int. J. Nanomed.* 16, 539–560.
- Wang, D., Jiang, A., Feng, J., Li, G., Guo, D., Sajid, M., Wu, K., Zhang, Q., Ponty, Y., Will, S., Liu, F., Yu, X., Li, S., Liu, Q., Yang, X.L., Guo, M., Li, X., Chen, M., Shi, Z.L., Lan, K., Chen, Y., Zhou, Y., 2021. The SARS-CoV-2 subgenome landscape and its novel regulatory features. *Mol. Cell* 81, 2135–2147 e5.
- Weingarten-Gabbay, S., Klaeger, S., Sarkizova, S., Pearlman, L.R., Chen, D.Y., Gallagher, K.M.E., Bauer, M.R., Taylor, H.B., Dunn, W.A., Tarr, C., Sidney, J., Rachimi, S., Conway, H.L., Katsis, K., Wang, Y., Leistriz-Edwards, D., Durkin, M.R., Tomkins-Tinch, C.H., Finkel, Y., Nachson, A., Gentili, M., Rivera, K.D., Carulli, I.P., Chea, V.A., Chandrashekar, A., Bozkus, C.C., Carrington, M., Bhardwaj, N., Barouch, D.H., Sette, A., Maus, M.V., Rice, C.M., Clauser, K.R., Keskin, D.B., Pregibon, D.C., Hacohen, N., Carr, S.A., Abelin, J.G., Saeed, M., Sabeti, P.C., MGH COVID-19 Collection, Processing Team, 2021. Profiling SARS-CoV-2 HLA-I peptide reveals T cell epitopes from out-of-frame ORFs. *Cell* 184, 3962–3980 e17.

- Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C., Zhang, Y.Z., 2020. A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269.
- Zhao, Y., Sun, J., Li, Y., Li, Z., Xie, Y., Feng, R., Zhao, J., Hu, Y., 2021. The strand-biased transcription of SARS-CoV-2 and unbalanced inhibition by remdesivir. *iScience* 24, 102857.
- Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., Chen, H.D., Chen, J., Luo, Y., Guo, H., Jiang, R.D., Liu, M.Q., Chen, Y., Shen, X.R., Wang, X., Zheng, X.S., Zhao, K., Chen, Q.J., Deng, F., Liu, L.L., Yan, B., Zhan, F.X., Wang, Y.Y., Xiao, G.F., Shi, Z.L., 2020. Addendum: a pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 588, E6.
- Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G.F., Tan, W., China Novel Coronavirus Investigating and Research Team, 2020. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* 382, 727–733.