A negative feedback model to explain regulation of SARS-CoV-2 replication and transcription

Shan Gao[1$], Zhi Cheng[1$], Xiufeng Jin[1], Fang Wang[3], Yibo Xuan[4]

Hao Zhou[1], Chang Liu[1], Jishou Ruan[5], Guangyou Duan[6]*, Xin Li[1]*

[1] College of Life Sciences, Nankai University, Tianjin, Tianjin 300071, P.R.China;
[2] State key laboratory of medicinal chemical biology, Nankai University, Tianjin, Tianjin 300071, P.R.China;
[3] The Second Hospital of Tianjin Medical University, Tianjin 300211, P. R. China;
[4] College of Life Sciences, Hebei Normal University, Shijiazhuang, Hebei 050024, P.R.China;
[5] School of Mathematical Sciences, Nankai University, Tianjin, Tianjin 300071, P.R.China;
[6] School of Life Sciences, Qilu Normal University, Jinan, Shandong 250200, P.R.China

[$] These authors contributed equally to this paper.

* The corresponding authors.

XL：lix1980@nankai.edu.cn
GD：guangyou.duan@qlnu.edu.cn

Abstract

Background: Coronavirus disease 19 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Although the replication and transcription mechanisms of SARS-CoV-2 have been preliminarily revealed recently, their regulation is still unclear.

Results: By reanalysis of public data, we proposed a negative feedback model to explain the regulation of replication and transcription of, but not limited to coronavirus (CoVs). The key step leading to new discoveries is that we identified the cleavage sites of nps15, an RNA uridylate-specific endoribonuclease. According to this model, nsp15 regulates the synthesis of subgenomic RNAs (sgRNAs) or genomic RNAs (gRNAs) by cleavage of transcription regulatory sequences in the body. The expression level of nsp15 determines the relative proportions of sgRNAs and gRNAs. The increase of nsp15 results in the decrease of itself to reach equilibrium between the replication and transcription of CoVs.

Conclusions: The replication and transcription of CoVs are regulated by a negative feedback strategy. This is a basic mechanism to influence the persistence of CoVs in hosts. Our study indicates that nsp15 is an important and ideal target for drug development against CoVs.

Keyword: coronavirus; transcription; replication; regulation model; nsp15

Introduction

Coronavirus disease 19 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). CoVs carry the largest genomes (26–32 kb) among all RNA virus families. As an enveloped virus with a positive-sense, single-stranded RNA, SARS-CoV-2 has a genome of ~30 kb [1]. Besides ORF1a and 1b (**Figure 1A**), SARS-CoV-2 encodes conserved four structural proteins (spike protein [S], envelope protein [E], membrane protein [M] and nucleocapsid protein [N]) and six accessory proteins (ORF3a, 6, 7a, 7b, 8 and 10) that have not yet been experimentally verified.

The genomic RNA, noted as gRNA(+), is used as the template for translation of ORF1a and ORF1b, replication and transcription of SARS-CoV-2. However, the translation of 10 other proteins (S, E, M, N, ORF3a, 6, 7a, 7b, 8 and 10) has not been revealed until a recent study [2]. That study has directly proved the prevailing "leader-to-body fusion" model using Nanopore RNA-seq (an RNA direct sequencing method [3]). In that model (**Figure 1A**), replication and transcription need the use of gRNA(+) as the template for the synthesis of antisense genomic RNAs (gRNAs(-)) and antisense subgenomic RNAs (sgRNAs(-)) by RNA-dependent RNA polymerase (RdRP). If RdRP pauses when it crosses a transcription regulatory sequence in the body (TRS-B) and switches the template to the TRS in the leader (TRS-L), resulting in sgRNAs(-). Otherwise, RdRP reads through gRNAs(+) without interruption, resulting in gRNAs(-). Then, gRNAs(-) and sgRNAs(-) are used as templates to synthesize gRNAs(+) and sgRNAs(+), respectively. sgRNAs(+) are used as the templates for translation of 10 other proteins.

TRS-L is usually the first 60~70 nt of 5' UTR in a CoV genome. TRS-B is located immediately upstream of ORFs except ORF1a and 1b (**Figure 1A**). During negative-strand synthesis, the discontinuous transcription (named as polymerase jumping or template switch) by RdRP is described by the "leader-to-body fusion" model. However, the underlying mechanisms are still unknown. In the present study,

we aimed to determine the molecular basis of the “leader-body fusion” model and construct a model to explain the regulation of SARS-CoV-2 replication and transcription.

Results

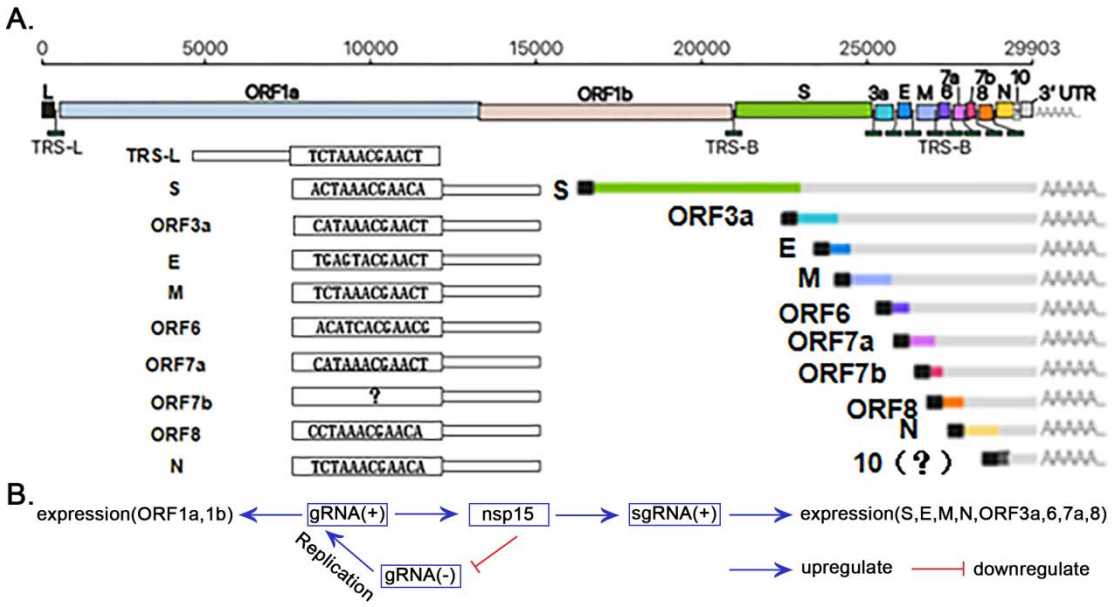

Figure 1 Replication and transcription of SARS-CoV-2

The elements used to represent the SARS-CoV-2 genome (GenBank: MN908947.3) were originally used in the previous study [2]. **A**. TRS-L: transcription regulatory sequence in the body; TRS-B: transcription regulatory sequence in the leader. **B**. gRNA(+):genomic RNAs; gRNAs(-): antisense genomic RNAs.

Reanalysis of leader-body junction regions

Using public Nanopore RNA-seq data (**Materials and Methods**), 575,106 out of 879,679 reads from a sample of SARS-CoV-2 infected Vero cells were aligned to the SARS-CoV-2 reference genome. Among all aligned reads, 575,106 positive reads represented gRNAs(+) or sgRNAs(+), while 30 negative reads represented gRNAs(-) or sgRNAs(-). This super high ratio between positive and negative reads could be explained by a reason that there was significant differences in the degradation efficiencies between gRNAs(+)/sgRNAs(+) and gRNAs(-)/sgRNAs(-), but not

between gRNAs(+)/gRNAs(-) and sgRNAs(+)/sgRNAs(-). Another reason to explain the high ratio is that gRNAs(+)/sgRNAs(+) are protected by binding to the N proteins. This phenomenon, however, not noticed in the previous study [2].

By reanalysis of junction reads, we found that TRS-Bs share ~12 nt junction regions with TRS-Ls in SARS-CoV-2 (**Figure 1A**). Most of junction regions of eight protein-coding genes (S, E, M, N, ORF3a, 6, 7a and 8) exhibit high similarity to those of TRS-Ls and are defined as canonical junction regions [2]. Accordingly, the sgRNAs(+) and sgRNAs(-) containing these canonical junction regions are defined as canonical sgRNAs(+) and sgRNAs(-), respectively. The junction regions of ORF7b exhibit diversity of their sequences. If the junction regions of ORF10 exist is still unclear [2]. By the analysis of canonical junction regions, we determined that the core sequence (CS) of them is ACGAAC (**Figure 1A**).

Further investigation of "leader-to-body fusion" model

Table1 The genomic sites of motif GTTCGT

Motif	Position	Region	Type
GTTCGT	16015	ORF1b (nsp12)*	ICS
	28199	ORF8	ICS
	28234	ORF8	ICS
ACGAAC	71	5' UTR	TRS-L
	21557	S	TRS-B
	25386	ORF3a	TRS-B
	26238	E	TRS-B
	26474	M	TRS-B
	27042	ORF6	TRS-B
	27389	ORF7a	TRS-B
	27889	ORF8	TRS-B
	28261	N	TRS-B

The motifs were mapped to the SARS-CoV-2 genome (GenBank: MN908947.3) using positions of the first nucleotides. ACGAAC indicated the motif GTTCGT, read on the negative strand of the SARS-CoV-2 genome. * this site is located in the CDS of nsp12 (RdRP). TRS-L: transcription

regulatory sequence in the body; TRS-B: transcription regulatory sequence in the leader; ICS: internal cleavage site; RdRP: RNA-dependent RNA polymerase.

Further analysis of 1265 betacoronaviruses showed that ACGAAC is highly conserved in their genomes. This suggested that canonical junction regions contain a specific motif for enzyme reorganization. Since all CoVs follows the "leader-to-body fusion" model [2], this or these enzymes are supposed to be encoded by ORF1a or 1b genes for their priority to be translated. After searching of 16 non-structural proteins (nsp1-16) encoded by ORF1a and 1b genes, we determined that nsp15 (a nidoviral RNA uridylate-specific endoribonuclease, NendoU [4]) is mostly like to function on these junction regions, since a homolog of nsp15 has cleavage sites containing "GU" [5]. Thus, the cleavage site of nsp15 was identified to follow the motif "GTTCGT|N" (a vertical line indicates the breakpoint and N indicates any nucleotide type), read from the negative stands of CoV genomes. Our finding does not rule out other possible breakpoints (e.g. "GTTCGTT|N" or "GTTCGTTT|N") in the junction regions.

Searching of "GTTCGT" in the genomes of betacoronavirus subgroup B, the occurrence of "GTTCGT" on the negative strand is more than 1.6 times than that on the positive strand. Particularly, "GTTCGT" occurred 3 and 9 times (**Table 1**) on the positive and negative strands of the SARS-CoV-2 genome, respectively. These findings suggested that the basic function of nsp15 is responsible for the degradation of gRNAs and gsRNAs, and the super high ratio between positive and negative reads (**See above**) resulted from much more cleavage of gRNAs(-)/gsRNAs(-) than that of gRNAs(+)/gsRNAs(+). Among three sites containing "GTTCGT" on the positive strand the SARS-CoV-2 genome (named internal cleavage sites, ICSs), one is located in the CDS of nsp12 (RdRP), while the other two of them are located in the gene ORF8 (**See below**). These two ICSs are also located in ORF8 of most SARS-CoV-2, SARS-2 like CoV, SARS-like CoV genomes, however, they are absent in SARS-CoV

genomes from humans (GenBank: AY274119 and AY278489) and SARS-like CoV genomes from civets (GenBank: AY304486, AY515512 and AY572034). One of two ICSs is present in the genome of SARS-like CoV strain WIV1 from bats (GenBank: KF367457), which was the ancestor of SARS-CoV. These findings can be used to explain the pandemic of SARS-CoV caused by an ORF8 recombinant event (**Conclusion and Discussion**). In addition, we found that an ICS of nsp15 is present in the S1 subunits of SARS-CoV genomes from humans and civets, but is absent in all other genomes of betacoronavirus subgroup B. An ICS containing “GTTCG[T/C]” of nsp15 is present in the S2 subunits of the SARS-like CoV strain WIV1 and SARS-CoV genomes from humans and civets, but is absent in all other genomes of betacoronavirus subgroup B. These findings suggested that ICSs of nsp15 enhance degradation of S at the RNA level, thus inhibiting the cell entry ability of SARS-CoV.

Based on the above analyses, we modified the “leader-body fusion” model and proposed the molecular basis of it. In our model, nsp15 cleaves synthesized gRNAs(-) and sgRNAs(-). The cleavage occurs at TRS-Bs(-) that are synthesized using TRS-Bs as templates. And then, the free 3' ends (~6 nt) of TRS-Bs(-) hybridize the junction regions of TRS-Ls for template switch. An alternative step is that the cleaved TRS-Ls(-) are used as templates to synthesize TRS-Ls, which then hybridize the junction regions of the cleaved sgRNAs(-) without TRS-Ls(-) for template switch and gRNAs(+)/sgRNAs(+) synthesis.

Some non-canonical sgRNAs(+) and sgRNAs(-) are synthesized due to occasional hybrid between the free 3' ends of TRS-Bs(-) and high similar sequences of TRS-L junction regions. The sgRNAs(-) without TRS-Ls are also synthesized due to missing hybrid. In the previous study, sgRNAs(-) without TRS-Ls, non-canonical sgRNAs(+) and sgRNAs(-) were reported [2], which supported our model. In addition, the cleavage also occurs at ICSs on the positive strands. And then, the free 3' ends of the cleaved 5' parts hybridize the negative strands to synthesize recombinant

sgRNAs(+) or even gRNAs(+). This could be used to explain the multiple recombinant events in the betacoronavirus genomes.

Proposal of a negative feedback model

Based on the above analyses, we proposed a negative feedback model (**Figure 1B**) to explain the regulation of replication and transcription of, but not limited to CoVs. In this model (**Figure 1**), nsp15 regulates the synthesis of subgenomic RNAs (sgRNAs) or genomic RNAs (gRNAs) by cleavage of TRS-Bs(-). The expression level of nsp15 determines the relative proportions of sgRNAs and gRNAs. The increase of nsp15 expression causes less gRNAs(-) and more gsRNAs(-), and then less gRNAs(+) and more gsRNAs(+) are synthesized, respectively. The decrease of gRNAs(+) results in the decrease of nsp15 expression, since nsp15 is synthesized using gRNA(+) as the template. Furthermore, the ICS in the CDS of SARS-CoV-2 nsp12 (**Table 1**) enhances the negative feedback. By this negative feedback strategy, CoVs reach equilibrium between the replication and transcription (**Figure 1B**). So, this is a basic mechanism to influence the persistence of CoVs in hosts.

Negative feedback regulation is necessary

In our previous study [6], we proposed that first hairpins (immediately upstream of ORF1a) have an important role in the functions (e.g. regulation of translational initiation) of ribosome binding sites (RBSs) in 5' UTRs of the SARS-CoV-2 genomes. SARS-CoV and SARS-CoV-2 have an identical first hairpin, which may enhance the translation of downstream genes. To indirectly prove that the negative feedback is a basic mechanism, we designed preliminary experiments to show that first hairpins from SARS-CoV-2 enhance the translation of the downstream enhanced green fluorescent proteins (EGFPs), however, over-expression of EGFP without negative feedback regulation will cause cell death.

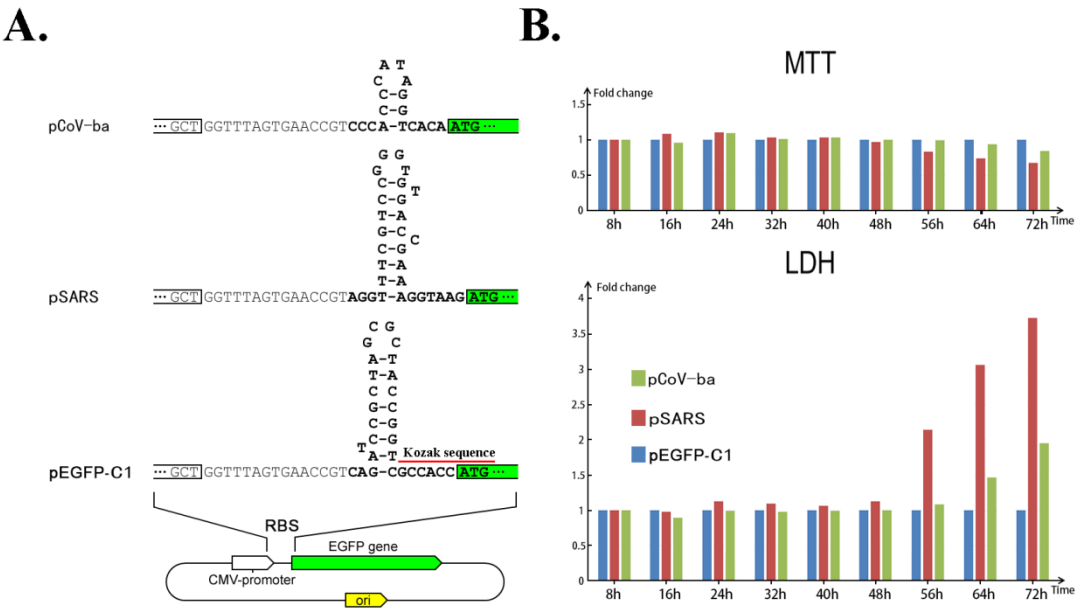

Figure 2 Regulation of EGFP expression by three RBSs

A. The plasmid pEGFP-C1 was used as control, since it contains 17-nt sequences, encoding the first hairpin from Cytomegalovirus (CMV). Two other types of plasmids pSARS and pCoV-ba were proceeded by replacement of first hairpins from SARS-CoV-2 and betacoronavirus subgroup A, respectively (**Materials and Methods**). **B**. HEK293 cells were transfected with three types of plasmids. At each time after transfection, cell viability and cell death were meausured using MTT and LDH, respectively.

In total, three types of plasmids containing the EGFP reporter genes, named pEGFP-C1, pSARS and pCoV-ba (betacoronavirus subgroup A), were used in the experiments (**Figure 2A**). The plasmid pEGFP-C1 was used as control, since it contains 17-nt sequences, encoding the first hairpin from Cytomegalovirus (CMV). Two types of plasmids proceeded by 30- and 29-nt insert replacement were used to evaluate their influence on translation (**Materials and Methods**). These two insert encoded two first hairpins from SARS-CoV-2 and the subgroup A of betacoronaviruses, respectively. Compared the fluorescent brightness of cells transfected with three types of plasmids, expression of EGFP in pSARS was obviously higher than that in pEGFP-C1 and pCoV-ba (**Supplementary 1**). In addition, pSARS caused cell death at the 48 hours after the plasmid transfection. Then, we performed 3-(4,5-dimethyl-2-thiazolyl)-2,5-diphenyl-2-H-tetrazolium bromide

(MTT) and lactate dehydrogenase (LDH) experiments to evaluate the influence (**Materials and Methods**). Both MTT and LDH experiments consistently suggested that pSARS caused significantly more dead HEK293T and Hela cells at the 48 hours after the plasmid transfection, than pEGFP-C1 and pCoV-ba. However, this phenomenon appeared in HEK293 cells at the 56 hours after the plasmid transfection (**Figure 2B**). Since the only difference between three types of plasmids is that they contain 17-, 30- and 29-nt sequences encoding different hairpins, we concluded that these hairpins determined the efficiency in the translation of the downstream EGFPs. The hairpin in pSARS resulted in the over-expression of EGFP, which caused more cell death. To determine that the factor is at the translation level, we performed experiments to rule out other possible factors which may have influence at the replication or transcription level: (1) using HEK293 cells to rule out the possible influence by the differences of plasmid copy numbers, since all three types of plasmids containing the SV40 origins can be replicated to a copy number of between 400~1000 plasmids per cell within HEK293T; and (2) using qPCR to rule out the possible influence by differential transcription (**Supplementary 1**).

Conclusion and Discussion

In the present study, we determined the molecular basis of the “leader-body fusion” model and construct a negative feedback model to explain the regulation of SARS-CoV-2 replication and transcription. The key step leading to new discoveries is that we identified the cleavage sites of nps15, an RNA uridylate-specific endoribonuclease (NendoU). Our model does not rule out other RNAs or proteins involved in these two models. However, current studies of the NendoU protein is still contradictory in fundamental questions [4]: (1) the NendoU proteins are conserved among coronaviruses, arteriviruses and toroviruses, are they present in nonvertebrate-infecting representatives of the nidoviruses order: mesoniviruses and roniviruses? (2) is nsp15 indispensable in viral replication and living? (3) is nsp15

responsible for the protein interference with the innate immune response? and (4) does nsp15 degrade viral RNA to hide it from the host defenses. Cleavage sites of nsp15 in relevant genes are new clues to answer these questions.

The ORF8 gene of SARS-CoV was considered to have played a significant role in adaptation to human hosts following interspecies transmission [7] via the modification of viral replication [8]. A 29-nt deletion in SARS-CoV (GenBank: AY274119) was reported and considered to be associated with attenuation during the early stage of human-to-human transmission [8]. A 382-nt deletion during the early evolution of SARS-CoV-2 was also reported [9]. Based on these reports, SARS-CoV enhanced the viral replication by acquiring a recombinant ORF8, while the 382-nt deletion of ORF8 in SARS-CoV-2 was considered to be associated with attenuation without changes in its replication [10]. The loss of two ICSs of nsp15 in ORF8 of SARS-CoV is the key clue to reveal the functions of ORF8 and the pandemic of SARS-CoV.

The kozak consensus sequence GCCACCAUGG (**Figure 1B**) is not necessary for the protein translation in eukaryotes. The hairpins immediately upstream of ORF1a (first hairpins) have an important role in the functions of CoV RBSs. Hairpin with proper structures may enhance the translation of their downstream proteins. These findings can be used to design vaccine, drug and expression vectors. Based on preliminary experiments, we concluded that the hairpin in pSARS resulted in the over-expression of EGFP, which caused cell death. These facts supported that the negative feedback is necessary to inhibit the over-expression of viral genes.

Materials and Methods

Nanopore RNA-seq data was downloaded from the website (https://osf.io/8f6n9/files/) for the reanalysis of leader-body junction regions. The results were confirmed using Illumina RNA-seq data from the NCBI SRA database under the accession number SRP251618. Data cleaning and quality control were

performed using Fastq_clean [12]. Statistics and plotting were conducted using the software R v2.15.3 with the package ggplot2 [15]. The 5' and 3' ends of gRNAs(+) and sgRNAs(+) were observed and double-checked using the software Tablet v1.15.09.01 [16].

In the present study, three types of plasmids pEGFP-C1 (maintained in the lab of Hao Zhou), pSARS and pCoV-ba, and three types (Hela, HEK293 and HEK293T) of cell maintained in our lab were used for transfection. To construct pSARS, pEGFP-C1 was PCR amplified, using primers fVR and rRBS2 (**Supplementary 1**). Then the linear PCR product was self-ligated into a plasmid by homologous recombination technology using ClonExpress II One Step Cloning Kit (Vazyme Biotech, China). Following the same procedure, primers fVR and rRBS3 (**Supplementary 1**) were used to construct pCoV-ba. For each experiment, 10,000 cells were seeded into one well of a 6-well plate for plasmid transfection. After 12 hours, transfection of 1 μg of plasmid into one well was performed using 3 μL PolyJet (SignaGen Laboratories, USA), following the manufacturer's instructions. The medium was changed every 12 hours after plasmid transfection. MTT (20 μl, 0.5 mg/ml) was added to each well and cultured at 37 °C for 4 h. Then, the cells in each well were removed of medium, washed with PBS, then mixed with 1000 μl DMSO to dissolve the formazan product. Finally, formazan absorbance was measured by a microplate reader with a wavelength of 492 nm (Thermo Labsystems Helsinki, Finland). LDH experiments were performed using LDH cytotoxicity assay detection kit (Beyotime, China), following the manufacturer's instructions.

Supplementary information

Additional file 1: Figure S1. Comparison of fluorescent brightness. Table S1. Primers for plasmid construction.

Abbreviations

COVID-19: Coronavirus disease 19; CoVs: coronavirus; SARS-CoV: Severe Acute Respiratory Syndrome Coronavirus; SARS-CoV-2: SARS Coronavirus 2; SARS-like CoV: SARS-CoV like Coronavirus; SARS-2 like CoV: SARS-CoV-2 like Coronavirus; gRNA: genomic RNA; sgRNA: subgenomic RNA; RBS: ribosome binding site; ORF: open reading frame; CDS: coding sequence; TRS-L: transcription regulatory sequence in the leader; TRS-B: transcription regulatory sequence in the body; ICS: internal cleavage site; RdRP: RNA-dependent RNA polymerase; nsp: non-structural protein; NendoU: nidoviral RNA uridylate-specific endoribonuclease; CMV: Cytomegalovirus; MTT: 3-(4,5-dimethyl-2-thiazolyl)-2,5-diphenyl-2-H-tetrazolium bromide; LDH: lactate dehydrogenase.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent to publish

Not applicable.

Availability of data and materials

All data used in the present study was download from the public data sources.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Natural Science Foundation of Tianjin (18JCYBJC28000) to Fang Wang and Tianjin Key Research and Development

Program of China (19YFZCSY00500) to Shan Gao. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

Shan Gao conceived the project. Xin Li and Guangyou Duan supervised this study. Guangyou Duan and Xiufeng Jin conducted programming. Zhi Cheng, Hao Zhou and Fang Wang conduced the experiments. Shan Gao and Yibo Xuan downloaded, managed and processed the data. Shan Gao drafted the main manuscript text. Shan Gao, Chang Liu and Jishou Ruan revised the manuscript.

Acknowledgments

We are grateful for the help from the following faculty members of College of Life Sciences at Nankai University: Xuetao Cao, Deling Kong, Quan Chen, Wenjun Bu, Ting Ma, Tao Zhang, Dawei Huang, Mingqiang Qiao, Yanqiang Liu, Qiang Zhao, Bingjun He and Zhen Ye.

REFERENCES

1. R.L. Graham and R.S. Baric, *Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission.* Journal of Virology, 2010. **84**(7): p. 3134-3146.
2. D. Kim, J.-Y. Lee, J.-S. Yang, J.W. Kim, V.N. Kim and H. Chang, *The Architecture of SARS-CoV-2 Transcriptome.* Cell, 2020. **181**(4): p. 914-921.
3. X. Xu, H. Ji, X. Jin, Z. Cheng, X. Yao, Y. Liu, Q. Zhao, T. Zhang, J. Ruan, W. Bu, Z. Chen and S. Gao, *Using pan RNA-seq analysis to reveal the ubiquitous existence of 5' and 3' end small RNAs.* Frontiers in Genetics, 2019. **10**: p. 1-11.
4. Y. Kim, R. Jedrzejczak, N.I. Maltseva, M. Wilamowski, M. Endres, A. Godzik, K. Michalska and A. Joachimiak, *Crystal structure of Nsp15 endoribonuclease NendoU from SARS-CoV-2.* Protein Science, 2020. **29**(7): p. 1596-1605.

5. D.D. Nedialkova, R. Ulferts, E. van den Born, C. Lauber, A.E. Gorbalenya, J. Ziebuhr and E.J. Snijder, *Biochemical Characterization of Arterivirus Nonstructural Protein 11 Reveals the Nidovirus-Wide Conservation of a Replicative Endoribonuclease.* Journal of Virology, 2009. **83**(11): p. 5671-5682.
6. G. Duan, J. Shi, Y. Xuan, J. Chen, C. Liu, J. Ruan, S. Gao and X. Li, *5' UTR Barcode of the 2019 Novel Coronavirus Leads to Insights into Its Virulence.* Chinese Journal of Virology (In Chinese), 2020. **36**(3): p. 365-369.
7. S.K.P. Lau, Y. Feng, H. Chen, H.K.H. Luk, W.H. Yang, K.S.M. Li, Y.Z. Zhang, Y. Huang, Z.Z. Song and W.N. Chow, *SARS coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination.* Journal of Virology, 2015: p. JVI.01048-15.
8. D. Muth, V.M. Corman, H. Roth, T. Binger, R. Dijkman, L.T. Gottula, F. Gloza-Rausch, A. Balboni, M. Battilani and D. Rihtarič, *Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission.* Scientific Reports, 2018. **8**(1).
9. Y.C. Su, D.E. Anderson, B.E. Young, F. Zhu, M. Linster, S. Kalimuddin, J.G. Low, Z. Yan, J. Jayakumar, L. Sun, G.Z. Yan, I.H. Mendenhall, Y.-S. Leo, D.C. Lye, L.-F. Wang and G.J. Smith, *Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2.* bioRxiv, 2020. **1**(1): p. 1-23.
10. B.E. Young, S.-W. Fong, Y.-H. Chan, T.-M. Mak, L.W. Ang, D.E. Anderson, C.Y.-P. Lee, S.N. Amrun, B. Lee, Y.S. Goh, Y.C.F. Su, W.E. Wei, S. Kalimuddin, L.Y.A. Chai, S. Pada, S.Y. Tan, L. Sun, P. Parthasarathy, Y.Y.C. Chen, T. Barkham, R.T.P. Lin, S. Maurer-Stroh, Y.-S. Leo, L.-F. Wang, L. Renia, V.J. Lee, G.J.D. Smith, D.C. Lye and L.F.P. Ng, *Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study.* The Lancet, 2020. **396**(10251): p. 1-9.
11. F. Wang, Y. Sun, J. Ruan, R. Chen, X. Chen, C. Chen, J.F. Kreuze, Z. Fei, X. Zhu and S. Gao, *Using small RNA deep sequencing data to detect human viruses.* BioMed Research International, 2016. **2016**(2016): p. 1-9.
12. M. Zhang, F. Zhan, H. Sun, X. Gong, Z. Fei and S. Gao. *Fastq_clean: An optimized pipeline to clean the Illumina sequencing data with quality control*. in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*. 2014. IEEE.
13. Y. Wang, Z. Wang, X. Chen, H. Zhang, F. Guo, K. Zhang, H. Feng, W. Gu, C. Wu and L. Ma, *The Complete Genome of Brucella Suis 019 Provides Insights on Cross-Species Infection.* Genes, 2016. **7**(2): p. 1-12.
14. Q. Cao, A. Li, J. Chen, Y. Sun, J. Tang, A. Zhang, Z. Zhou, D. Zhao, D. Ma and S. Gao, *Transcriptome Sequencing of the Sweet Potato Progenitor*

(Ipomoea Trifida (HBK) G. Don.) and Discovery of Drought Tolerance Genes. Tropical Plant Biology, 2016: p. 1-10.

15. H. Wickham, *ggplot2: elegant graphics for data analysis*. 2009: Springer Science & Business Media.

16. I. Milne, G. Stephen, M. Bayer, P.J. Cock, L. Pritchard, L. Cardle, P.D. Shaw and D. Marshall, *Using Tablet for visual exploration of second-generation sequencing data.* Briefings in bioinformatics, 2012: p. bbs012.