Achilles' heel of SARS-CoV-2: transcription regulatory sequence and leader sequence in 5' untranslated region have unique evolutionary patterns and are vital for virus replication in infected human cells

Manijeh Mohammadi-Dehcheshmeh[1,2ϕ], Sadrollah Molaei Moghbeli[3ϕ], Samira Rahimirad[4], Ibrahim O. Alanazi[5], Zafer Saad Al Shehri[6], Esmaeil Ebrahimie[1,2,7*]

[1] La Trobe Genomics Research Platform, School of Life Sciences, College of Science, Health and Engineering, La Trobe University, Melbourne, Victoria, 3086, Australia. [2]School of Animal and Veterinary Sciences, The University of Adelaide, South Australia, 5371, Australia. [3]Department of Animal Science, University of Wisconsin, Madison, USA. [4]Department of Medical Genetics, National Institute of Genetic Engineering and Biotechnology, Tehran, Iran. [5]National Center for Biotechnology, Life Science and Environment Research Institute, King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia. [6]Clinical Laboratory Department, College of Applied Medical Sciences, Shaqra University, KSA, Al dawadmi, Saudi Arabia. [7]School of BioSciences, The University of Melbourne, Victoria Australia

[ϕ]Authors with equal contributions

[*]Corresponding Author
College of Science, Health and Engineering,
La Trobe University
Melbourne, 3086
Victoria, 3086
Email: E.Ebrahimie@latrobe.edu.au
Phone: 0061449121357

Abstract

Background SARS-CoV-2 has generated a life-treating pandemic and is the main challenge of this century. Some untranslated regions (UTRs) in SARS-CoV-2 genome, specifically leader sequence and transcription regulatory sequence (TRS) in 5'UTR, can be considered as Achilles' heel of virus. Leader sequence are found at the 5' ends of all encoded transcripts that highlights its importance. TRS can explain the host range and pathogenicity of coronavirus. However, our knowledge on the evolution and the role of UTRs in SARS-CoV-2 pathogenicity is very limited. This study is a pioneering attempt to unravel the evolution of key regions in 5' UTR of SARS-CoV-2 and discover the inhibitory microRNAs against 5' UTR of virus.

Methods Evolution of TRS and leader sequence was compared between human pathogenic (SARS-CoV-2, SARS, and MERS) and non-pathogenic (bovine) coronaviruses. Profiling of microRNAs that can inactive the key UTR regions of coronaviruses, UTR-inhibitory microRNAs, was carried out.

Findings We found a distinguished pattern of evolution in leader sequence and TRS of SARS-CoV-2, compared to the other coronaviruses. Mining all available microRNA families against leader sequences of coronaviruses resulted in discovery of 39 microRNAs with an acceptable thermodynamic binding energy against SARS-COV-2, SARS, MERS, Bat Coronavirus, or Bovine Coronavirus. Multivariate analysis demonstrated a distinguished pattern of binding of leader sequence of SARS-CoV-2 against microRNAs, with a lower binding stability.

hsa-MIR-5004-3p was the only human microRNA that can target leader sequence of SARS and SARS-CoV-2. However, its binding stability remarkably decreased in SARS-COV-2 (-19.4 kcal/mol), compared to SARS-COV-2 (-25.9 kcal/mol). We found an insertion-type mutation in leader sequence of SARS-COV-2 that results in lower binding stability and escaping of viral leader sequence from hsa-MIR-5004-3p. Altogether, we suggest lack of innate human inhibitory microRNAs to bind to leader sequence and TRS of SARS-CoV-2 contributes to its high replication in infected human cells.

On the other hand, mining of two hundred million deposited human genomic variants led us to discovery of 6 splice-disrupt mutations in genomic structure of hsa-MIR-5004-3p. These mutations can negatively affect hsa-MIR-5004-3p transcript in preventing SARS-CoV-2 replication.

Interpretation This study unravels the evolution of key regions in 5'UTR of SARS-CoV-2. Inducing microRNAs to bind to the leader sequence and TRS regions by drugs or food supplements can reduce virus replication. Enhancing the microRNA defence machinery against TRS and leader of virus has a potential to prevent SARS-CoV-2 infection at the first place. The mentioned strategy is rapidly achievable against COVID-19. Variation in genomic sequence of 5'UTR inhibitory microRNAs, such as hsa-MIR-5004-3p, can be considered as risk factor of COVID-19.

Keywords

COVID-19, Drug repurposing, microRNA, SARS-CoV-2 replication, Variant discovery

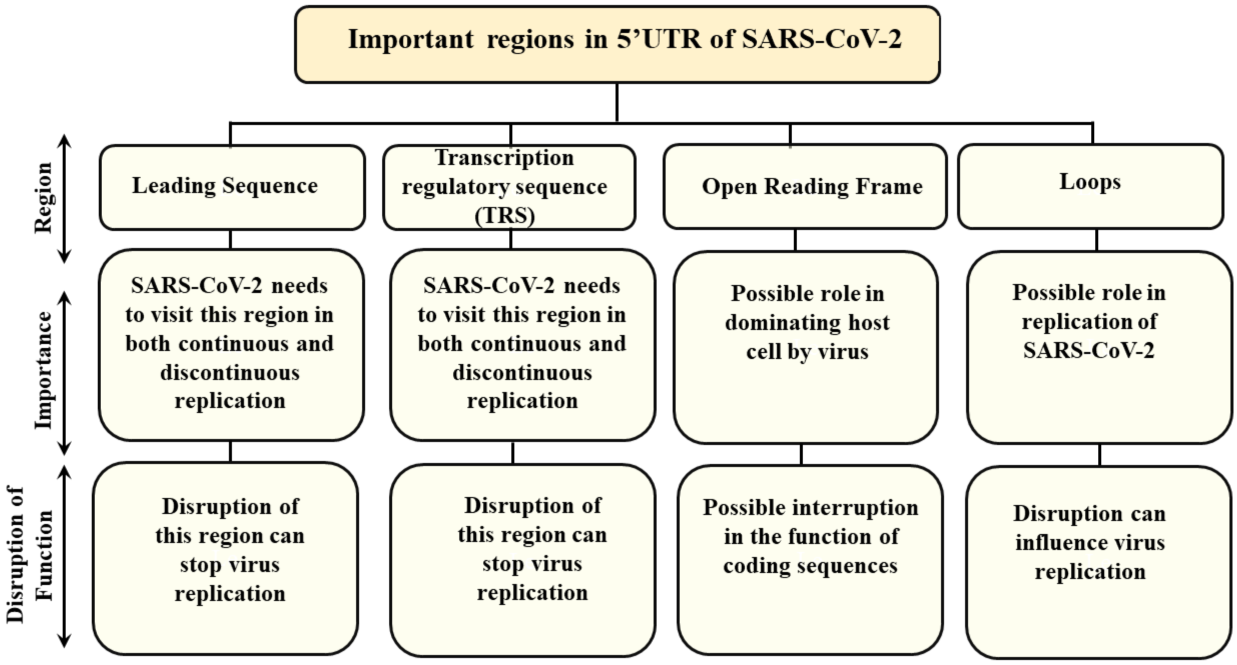
doi:10.20944/preprints202009.0602.v1

Background

Coronavirus genome is a single-stranded mRNA, containing both coding and untranslated regions (UTRs). The 5'UTR and 3'UTR are crucial for coronavirus RNA replication, transcription, and dominating host systems biology (Yang & Leibowitz, 2015). However, their exact roles and their evolutions, specifically in new SARS-CoV-2, are mainly unknown.

Notably, there are some regions in 5'UTR that are vital for virus replication (Figure 1). There is a specific region of 70 nucleotides at the 5' end of the genome, referred as the 'leader' sequence, has been found also at the 5' ends of all encoded transcripts that highlights its importance. A conserved cis-acting element, named transcription regulatory sequence (TRS), immediately follows the leader sequence, representing a unique feature of coronaviruses. Notably, in both continuous and discontinuous replication patterns, SARS-CoV-2 virus segments need to visit TRS and leader sequence on 5'UTR (Figure 2).

Whereas UTR is a non-coding region, another unique structure of coronaviruses is existence of a short reading frame (ORF) on one of the 5'UTR stem loops. In bovine coronavirus, maintenance of the short ORF is positively correlated with viral RNA accumulation (Raman *et al*, 2003). The 3'-UTR folds into a unique "stem-loop two motif "secondary structure that is required for virus viability.



Important regions in 5'UTR of SARS-CoV-2
Region
Importance
Disruption of Function
Leading Sequence
Transcription regulatory sequence (TRS)
Open Reading Frame
Loops
SARS-CoV-2 needs to visit this region in both continuous and discontinuous replication
SARS-CoV-2 needs to visit this region in both continuous and discontinuous replication
Possible role in dominating host cell by virus
Possible role in replication of SARS-CoV-2
Disruption of this region can stop virus replication
Disruption of this region can stop virus replication
Possible interruption in the function of coding sequences
Disruption can influence virus replication

Figure 1. 5'UTR is Achilles' heel of SARS-CoV-2. Disruption of specific regions 5'UTR can stop the virus replication.

UTRs are also important in the context of host microRNA interaction. MicroRNAs play their negative regulatory roles via sequence-specific interactions with the 3′ or 5′UTRs (Alanazi *et al*, 2019). Consequently, sequence variation in UTR regions of virus can prevent the binding of human microRNAs, resulting in lack of microRNA-based immunity. The problem is more serious in the case of SARS-CoV-2 where the virus is originated from bat, and human microRNAs are not evolved enough to control the UTRs of SARS-CoV-2.

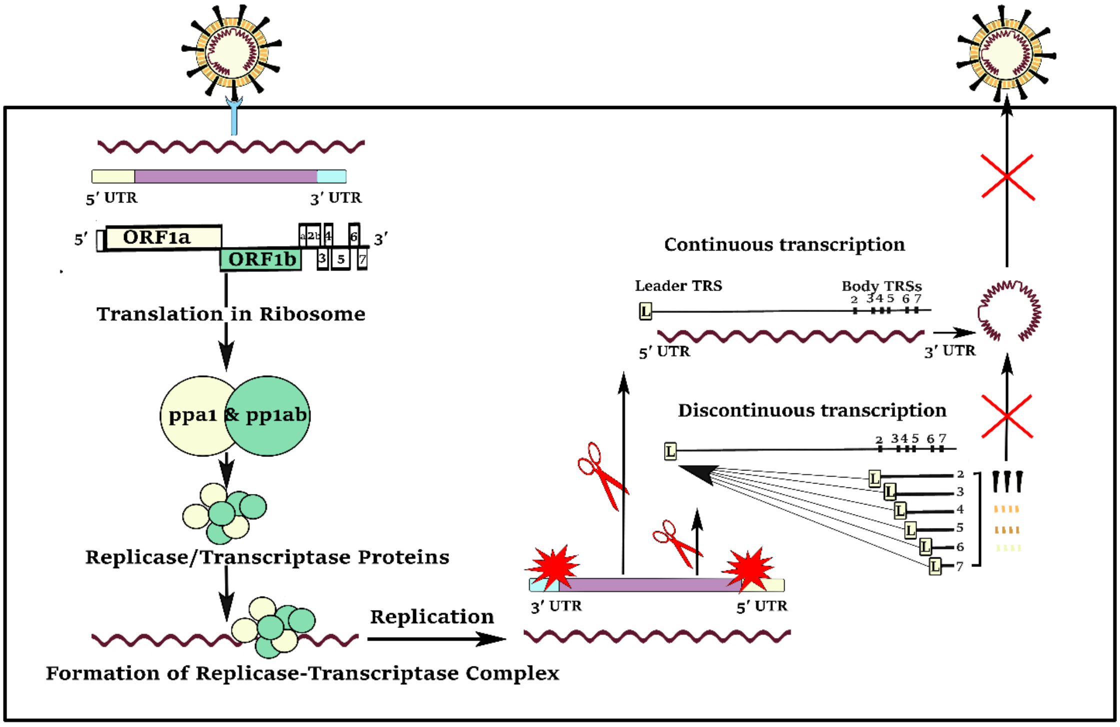


Figure 2. In both continuous and discontinuous replication patterns, SARS-CoV-2 virus segments need to visit 5’UTR, particularly TRS (transcription regulatory sequence) and leader sequence (L). 5’UTR is a good target to lower virus load in host cell.

We

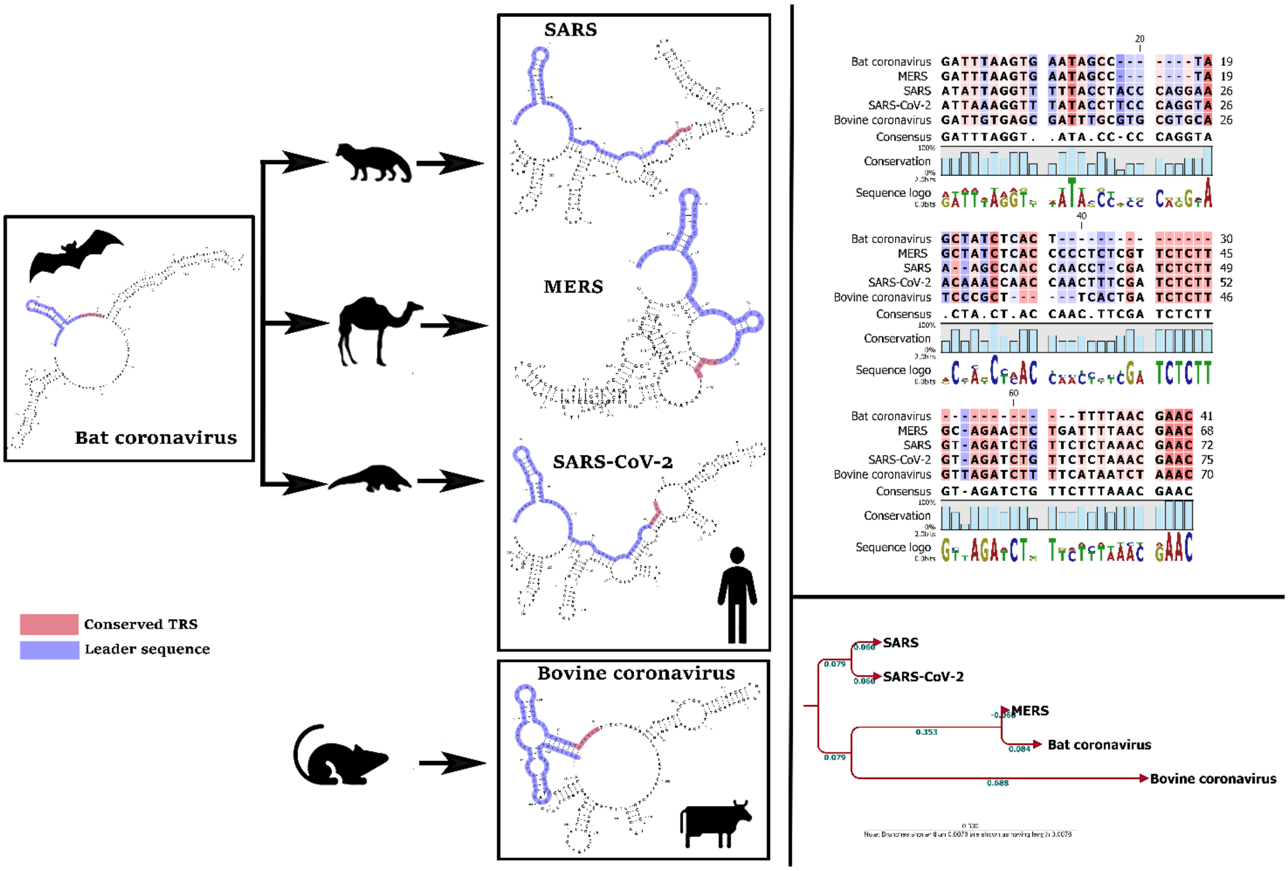
Here, we hypothesize that TRS and leader sequence of 5’ UTRs are crucial for coronavirus RNA replication. Also, microRNAs can bind to TRS and leader sequence of 5’UTR to lower SARS-CoV-2 replication. The first aim of this study was to unravel the evolution of TRS and leader sequence in 5’UTR regions of SARS-CoV-2. The second aim of this study was to identify the inhibitory microRNAs, originated from human and other organisms, that can bind to key UTR regions of SARS-CoV-2 sequences (inhibitory microRNAs).

doi:10.20944/preprints202009.0602.v1

Results

Comparative analysis of 5’UTR of human pathogen and non-pathogen coronaviruses

Our primary analysis on reference sequences demonstrated a distinguished pattern of evolution in leader sequence and TRS of SARS-CoV-2, in comparison to MERS and bovine coronavirus (Figure 3). Interestingly, TRS sequence is identical in all coronaviruses that infect human (SARS-CoV-2, SARS, and MERS), but TRS sequence alters in non-human pathogens, such as bovine coronavirus (Figure 3). In other words, TRS can explain the host range of coronavirus (Figure 3).



```
                                                          20
Bat coronavirus     GATTTAAGTG  AATAGCC---  -----TA  19
MERS                GATTTAAGTG  AATAGCC---  -----TA  19
SARS                ATATTAGGTT  TTTACCTACC  CAGGAA  26
SARS-CoV-2          ATTAAAGGTT  TATACCTTCC  CAGGTA  26
Bovine coronavirus  GATTGTGAGC  GATTTGCGTG  CGTGCA  26
Consensus           GATTTAGGT.  .ATA.CC-CC  CAGGTA
Conservation
Sequence logo
                                   40
Bat coronavirus     GCTATCTCAC  T---------  ------  30
MERS                GCTATCTCAC  CCCCTCTCGT  TCTCTT  45
SARS                A--AGCCAAC  CAACCT-CGA  TCTCTT  49
SARS-CoV-2          ACAAACCAAC  CAACTTTCGA  TCTCTT  52
Bovine coronavirus  TCCCGCT---  ---TCACTGA  TCTCTT  46
Consensus           .CTA.CT.AC  CAAC.TTCGA  TCTCTT
Conservation
Sequence logo
                          60
Bat coronavirus     ----------  ---TTTTAAC  GAAC  41
MERS                GC-AGAACTC  TGATTTTAAC  GAAC  68
SARS                GT-AGATCTG  TTCTCTAAAC  GAAC  72
SARS-CoV-2          GT-AGATCTG  TTCTCTAAAC  GAAC  75
Bovine coronavirus  GTTAGATCTT  TTCATAATCT  AAAC  70
Consensus           GT-AGATCTG  TTCTTTAAAC  GAAC
Conservation
Sequence logo
```

Figure 3. Evolution of “leader sequence” and TRS (transcription regulatory sequence) in 5’UTR of human pathogen (SARS, MERS, and SARS-CoV-2) and non-pathogen (bovine) coronaviruses. We found significant evolution pattern in SARS-CoV-2 compared to the other coronaviruses.

Identifying the microRNAs that can bind to leader sequence and TRS of SARS-CoV-2 (5’UTR inhibitory microRNAs)

Mining all available microRNA families against leader sequence of SARS-CoV-2 resulted in discovery of 39 microRNAs with an acceptable thermodynamic binding energy (less than –15 kcal/mol). Table 1 presents a list of microRNAs, their organisms, and their thermodynamic binding energy (kcal/mol) that can bind to at least one type of coronaviruses (SARS-COV-2, SARS, MERS, Bat Coronavirus, or Bovine Coronavirus). ptc-MiR474b, ptc- MiR474a, csa-let-7d, cin-let-7d-5p, and gga-MiR -6608-3p were microRNAs with stable thermodynamic binding energy (lower than –25 kcal/mol) against leader sequence of SARS-CoV-2.

hsa-MIR-5004-3p was the only human microRNA that can target leader sequence of SARS and SARS-CoV-2. However, its binding stability remarkably decreased in SARS-COV-2 (-19.4

kcal/mol), compared to SARS-COV-2 (-25.9 kcal/mol) (Table 1 and Figure 4). Notably, our analysis showed that leader sequence of SARS-COV-2 is mutated (insertion type mutation, CA) to escape microRNA-RNA hybridization (Figure 5).

Table 1. MicroRNAs and that can bind to the leader sequence of 5' untranslated region (5'UTR) of coronavirus and their binding energy. Lower binding energy demonstrates higher binding stability between microRNA and leader sequence of coronavirus.

MicroRNA	**Organism**	**Thermodynamic binding energy against leader sequence (kcal/mol)**				
		SARS-COV-2	**SARS**	**MERS**	**Bat Coronavirus**	**Bovine Coronavirus**
ptc-miR474b	*Populus trichocarpa*	-27.3	-24.5	-21.6	-21.5	-17
ptc-miR474a	*Populus trichocarpa*	-27.3	-22.2	-22.8	-22.2	-18.1
csa-let-7d	*Ciona savignyi*	-25.1	-22.7	-24.6	-24.6	-19
cin-let-7d-5p	*Ciona intestinalis*	-25.1	-22.7	-24.6	-24.6	-19
gga-miR-6608-3p	*Gallus gallus*	-25	-23.6	-30.1	-14.6	-17.1
eca-miR-9080	*Equus caballus*	-23.4	-27.7	-15.9	-15.9	-16.5
csi-miR3953	*Citrus sinensis*	-22.5	-22.4	-27.2	-14.1	-16.8
ame-miR-3741	*Apis mellifera*	-21.9	-20.9	-35.2	-16	-21.7
cel-miR-8207-3p	*Caenorhabditis elegans*	-20.5	-22	-25.6	-12.2	-22.6
ppy-miR-1273a	*Pongo pygmaeus*	-20.1	-21.1	-23.4	-18.7	-19.5
hsa-miR-5004-3p	*Homo sapiens*	-19.4	-25.9	-17.7	-17.7	-14.1
bta-miR-2284ab	*Bos taurus*	-19.3	-21.6	-16.8	-19.9	-13.8
oan-miR-1395-5p	*Ornithorhynchus anatinus*	-19.3	-27.8	-17.5	-15.3	-13.4
mdo-miR-137b-5p	*Monodelphis domestica*	-17.7	-26.8	-19.4	-15.1	-16.8
dme-miR-4949-3p	*Drosophila melanogaster*	-17.7	-17.4	-13.2	-12.6	-24.4
ssc-miR-9833-5p	*Sus scrofa*	-17.1	-16.3	-15.9	-15.9	-15.7
ptc-miR6464	*Populus trichocarpa*	-16.2	-15.4	-13.7	-13.7	-13.1
mtr-miR2629g	*Medicago truncatula*	-15.1	-21.7	-13.1	-13.1	-14.1
mtr-miR2629f	*Medicago truncatula*	-15.1	-21.7	-13.1	-13.1	-14.1
mtr-miR2629e	*Medicago truncatula*	-15.1	-21.7	-13.1	-13.1	-14.1
mtr-miR2629d	*Medicago truncatula*	-15.1	-21.7	-13.1	-13.1	-14.1
mtr-miR2629c	*Medicago truncatula*	-15.1	-21.7	-13.1	-13.1	-14.1
mtr-miR2629b	*Medicago truncatula*	-15.1	-21.7	-13.1	-13.1	-14.1
mtr-miR2629a	*Medicago truncatula*	-15.1	-21.7	-13.1	-13.1	-14.1
bmo-miR-3293	*Bombyx mori*	-13.9	-14.8	-15.7	-12.7	-16
dsi-miR-986-3p	*Drosophila simulans*	-12.4	-16.5	-25.1	-25.1	-19.7
dme-miR-986-3p	*Drosophila melanogaster*	-12.4	-16.5	-25.1	-25.1	-19.7
dsi-miR-986-3p	*Drosophila simulans*	-12.4	-16.5	-25.1	-25.1	-19.7
dme-miR-986-3p	*Drosophila melanogaster*	-12.4	-16.5	-25.1	-25.1	-19.7
mmu-miR-6957-3p	*Mus musculus*	-11.6	-13.5	-12.2	-12.2	-12.1
ppc-miR-83-5p	*Pristionchus pacificus*	-11.4	-14.7	-10.9	-11.4	-17.6
cme-miR1863	*Cucumis melo*	-11.3	-10.5	-12.5	-12.5	-15.2
cel-miR-2211-5p	*Caenorhabditis elegans*	-10.7	-10.6	-10.5	-10.3	-12.2
ath-miR5638a	*Arabidopsis thaliana*	-9.8	-7.8	-10.9	-10.9	-16.6
bdi-miR5065	*Brachypodium distachyon*	-9.7	-11.8	-23.6	-23.6	-16.9
bdi-miR5065	*Brachypodium distachyon*	-9.7	-11.8	-23.6	-23.6	-16.9

oan-miR-1421l-2-3p	*Ornithorhynchus anatinus*	-9.7	-10.8	-12.6	-11.1	-21.2
mghv-miR-M1-2-3p	*Mouse gammaherpesvirus 68*	-9.4	-8.8	-8.3	-5.7	-7.5
dps-miR-2535-3p	*Drosophila pseudoobscura*	-9.3	-10.6	-17.1	-17.1	-19.6
	Average	-16.33	-18.58	-18.34	-16.35	-16.61

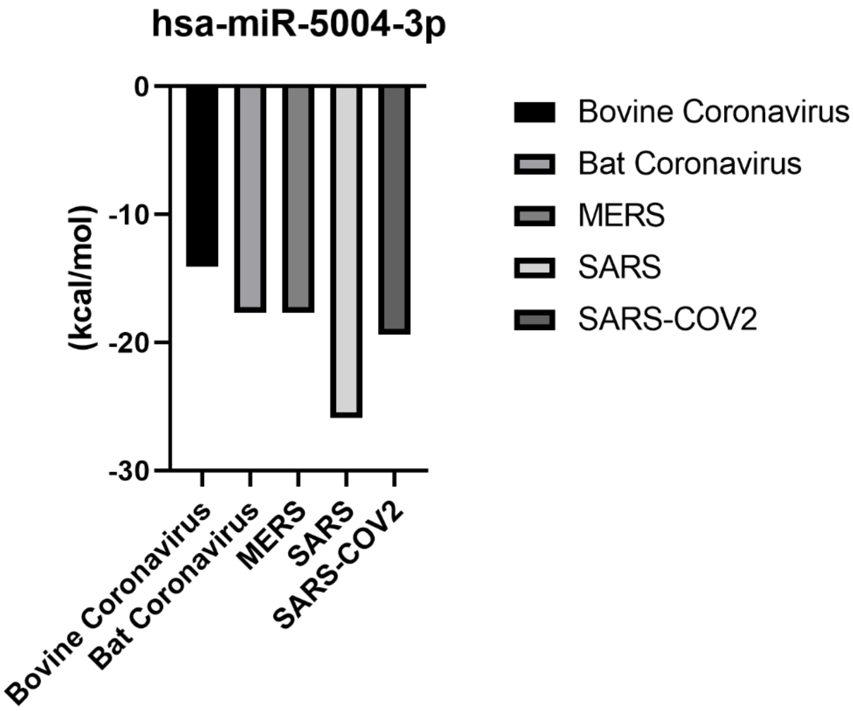


Figure 4. hsa-miR-5004-3p is the only human microRNA that can target leader sequence of SARS-CoV-2 and SARS. Lower binding energy demonstrates higher binding stability between microRNA and leader sequence of coronavirus.

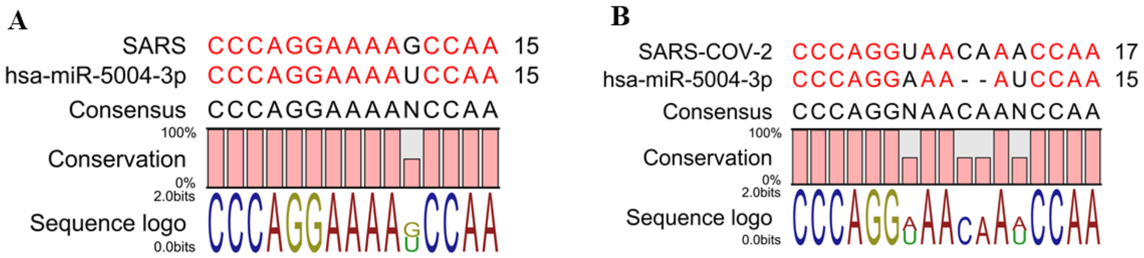

A
SARS CCCAGGAAAAGCCAA 15
hsa-miR-5004-3p CCCAGGAAAAUCCAA 15
Consensus CCCAGGAAAANCCAA
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits
B
SARS-COV-2 CCCAGGUAACAAACCAA 17
hsa-miR-5004-3p CCCAGGAAA--AUCCAA 15
Consensus CCCAGGNAACAANCCAA
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits

Figure 5. A two-nucleotide insertion-type mutation in leader sequence of SARS-COV-2 results in lower binding between viral leader sequence and seed region of hsa-MIR-5004-3p. (A) Alignment of hsa-miR-5004-3p with SARS. (B) Alignment of hsa-miR-5004-3p with SARS-CoV-2.

bta-MiR-2284, bta-MiR-2957, mmu-MiR-6414, mmu-MiR-6970-5p, mmu-MiR-873b, and hsa-MiR-148a were the microRNAs that can target TRS of 5'UTR. hsa-MiR-148a was the only human microRNA that can target TRS of SARS-COV-2 (Figure 6).

Lack of innate human inhibitory microRNAs to bind to leader sequence and TRS of SARS-CoV-2 contributes to the high replication of SARS-CoV-2 in infected human cells.

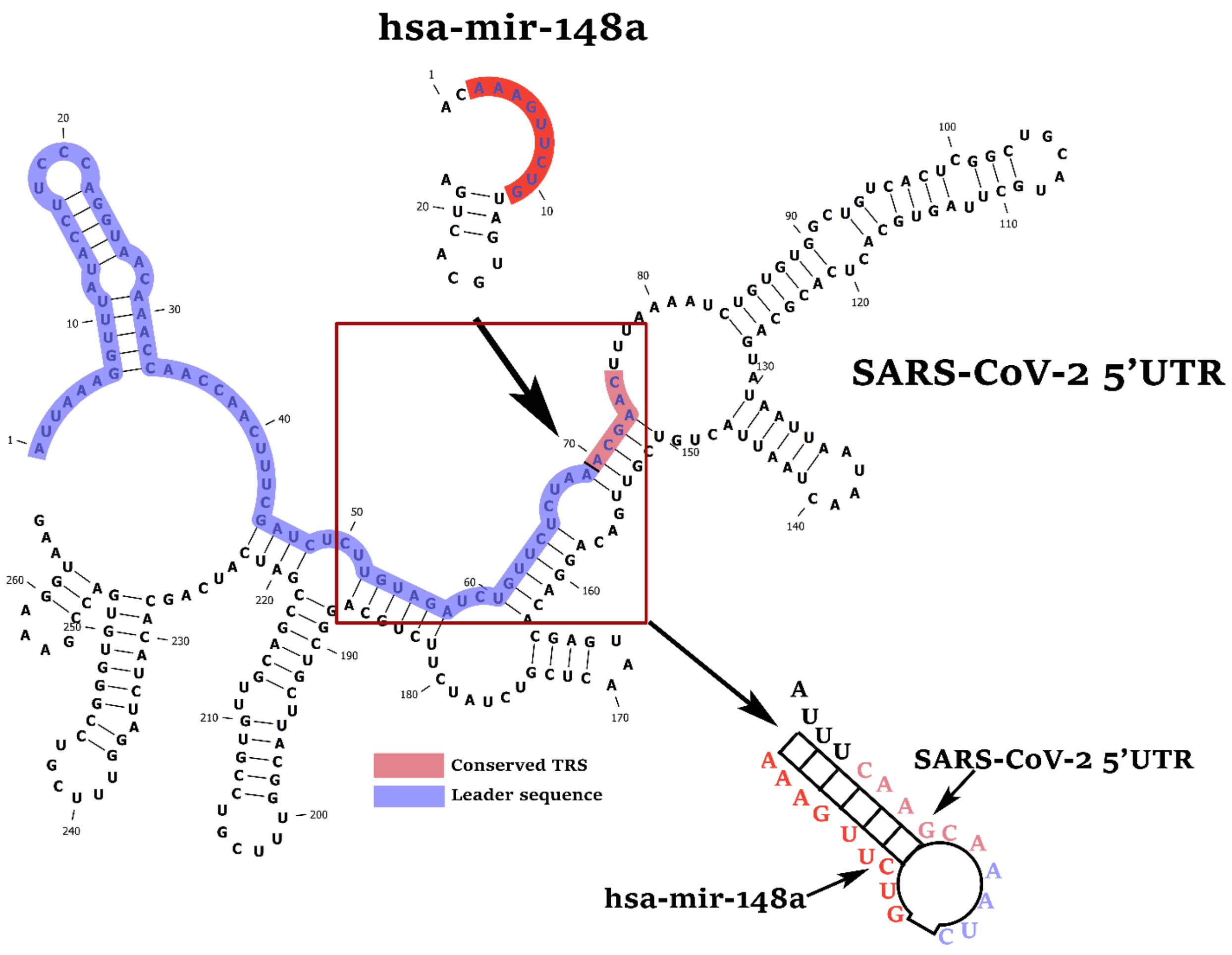


Figure 6. Hsa-MiR-148a is the only human microRNA that can target transcription regulatory sequence (TRS) and leader sequence of SARS-CoV-2.

Leader sequence of SARS-CoV-2 has a distinguished pattern of microRNA binding, compared to SARS, MERS, bat, and bovine coronaviruses

Multivariate analysis of thermodynamic binding energy values of leader sequence of coronavirus against mined inhibitory microRNAs (Table 1) demonstrates a distinguished pattern of SARS-CoV-2 evolution (Figure 7).

PCA efficiently discriminated SARS-CoV-2 from the other types of coronaviruses where PCA1 and PCA2 described 71.9% of variation in data (Supplementary 1). SARS-CoV-2 had negative values of both PCA1 and PCA2 (Figure 7). Interestingly, hsa-MIR-5004-3p is one of the top 5 important microRNAs in PCA1 with absolute coefficient > 0.2 (Supplementary 1). ptc-miR474b, cme-miR1863, bta-miR-2284ab, and bdi-miR5065 were top microRNAs in PCA2. In line with this finding, in comparison with the other human pathogens (SARS and MERS), leader sequence of SARS-COV-2 has remarkably higher binding energy against microRNAs, -16.33 kcal/mol against 18.58 kcal/mol and -18.34 kcal/mol, respectively (Table 1). Higher binding energy results in lower stability of microRNA-5'UTR binding and provides this opportunity for SARS-COV-2 to escape the inhibitory microRNAs.

Clustering shows that SARS-CoV-2 has only 41.6% similarity with SARS in pattern of binding to microRNAs (Supplementary 1 and Figure 7).

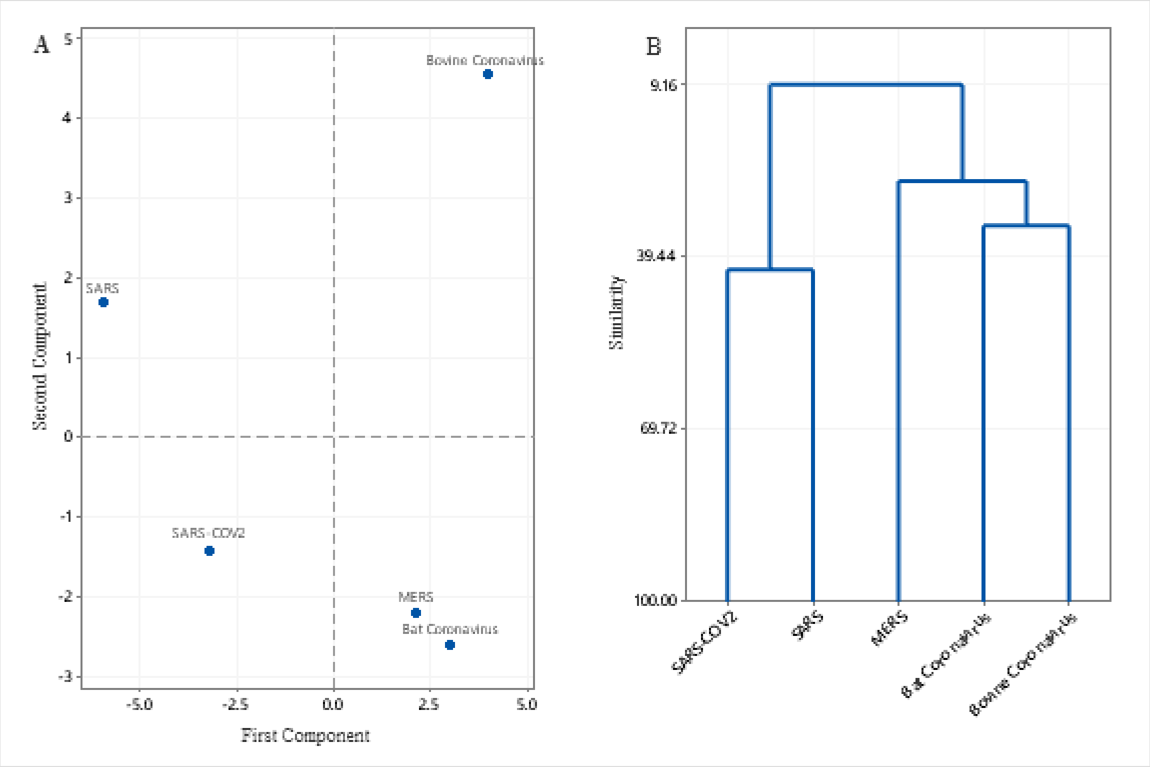


Figure 7. Multivariate analysis of thermodynamic binding energy values of leader sequence of coronavirus against mined inhibitory microRNAs (Table 1) demonstrates a distinguished pattern of SARS-CoV-2 evolution. (A) Principle Component Analysis discriminates SARS-CoV-2 from the rest of coronaviruses. (2) Clustering shows that SARS-CoV-2 has only 41.6% similarity with SARS in pattern of binding to microRNAs.

Drug repurposing to induce 5'UTR inhibitory microRNA

As presented in Figure 8, in this study, we developed a literature mining-based drug repurposing to induce inhibitory microRNAs against leader sequence and TRS of SARS-CoV-2. To this end, more than 1 million drug and small molecule and 12 million relations (binding, biomarker, expression, chemical reaction, promoter binding, microRNA effect, etc) were mined by Natural Language Processing (NLP).

Literature mining-based drug repurposing could not find any drug, small molecule, or food supplement with direct interaction with hsa-MIR-5004-3p (Figure 8). However, downregulation of *KLF4*, by heme and calcium, can indirectly upregulate hsa-MIR-5004-3p, as an inhibitory microRNA against leader sequence of SARS-CoV-2. Heme is an iron-containing tetradentate ligand. Relation of network and their references are provided in Supplementary 2.

Our drug repurposing analysis showed that bufalin is a potential drug to activate MiR148 against TRS region in 5'UTR of SARS-CoV-2 (Figure 9). Bufalin is an anticancer drug that upregulates MiR148a (Chang *et al*, 2015). Bufalin is used in traditional Chinese medicine for many years and can be rapidly utilised for treatment of COVID-19 (Wang *et al*, 2017). However, further analysis showed that there is a risk that MiR148 downregulates *OAS2*, a gene involved in innate immune response (Gottwein *et al*, 2011; Skalsky *et al*, 2012). Consequently, upregulation of MiR148 against SARS-CoV-2 needs further investigation. Relation of network and their references are provided in Supplementary 2.

doi:10.20944/preprints202009.0602.v1

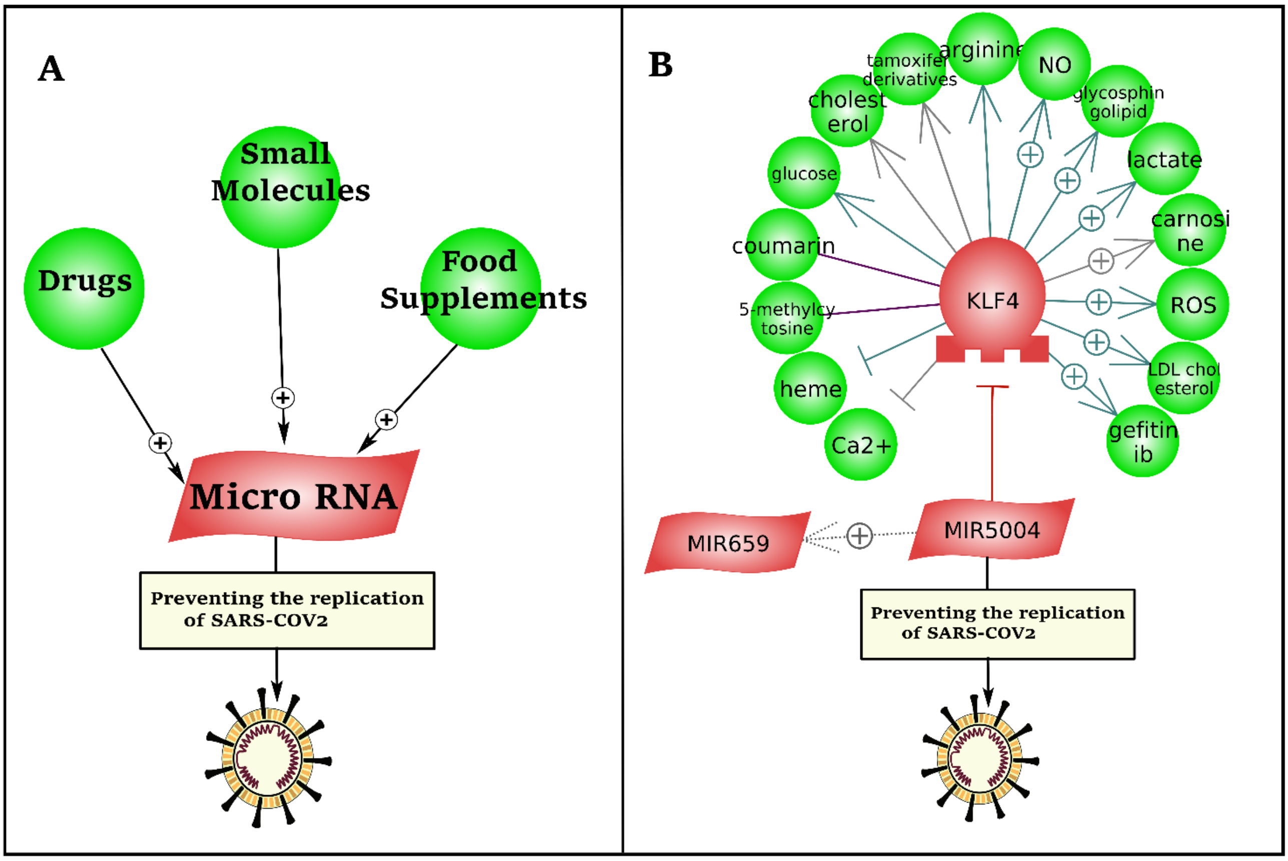


Figure 8. Drug repurposing to induce inhibitory microRNAs against leader sequence and TRS of 5'UTR region of SARS-CoV-2 in this study. (A) The presented pipeline of literature mining-based drug discovery. (B) Upregulation of hsa-MIR-5004-3p has potential to prevent the SARS-CoV-2 replication via binding to leader sequence of 5'UTR of virus.

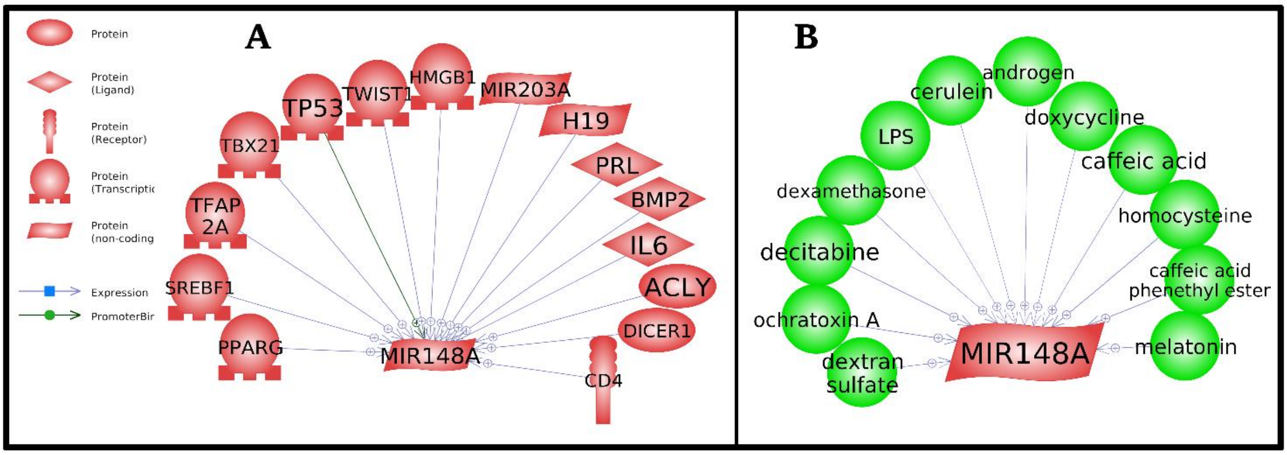


Figure 9. Molecular network of hsa-MiR-148a (A) Proteins that can upregulate MiR-148a expression. (2) Chemicals that can upregulate MiR-148a expression. Bufalin is a drug that can upregulate the MiR-148a expression.

hsa-MIR-5004-3p genomic variation as risk factor of COVID-19 infection

Impaired 5'UTR inhibitory microRNAs in human genome can contribute to high rate of virus replication in host cells. Consequently, mutations in genomic sequences of 5'UTR inhibitory microRNAs can be considered as risk factor of COVID-19 infection. Mining more than two hundred million deposited variants in dbSNP, using Pathway Studio tool (Elsevier), resulted in discovery of 9 variants in hsa-MIR-5004 (Table 2). Six of thesis variants were splice-disrupt mutations with possible regulatory functions. Table 3 presents the probably damaging variants according to GERP++ conservation score. GERP is an evolutionary conservation score which have a good correspondence with clinical significance and pathogenicity level. GERP++ conservation score varies from -12.3 to 6.17. Bigger value of GERP++ conservation score demonstrates higher conservation. Splice-disrupt mutations in MIR5004 have high GERP++ score, suggesting their high clinical importance.

Table 2. hsa-MIR-5004-3p genomic variation as risk factor of COVID-19 infection. High GERP++ conservation score is associated with higher functional Impact.

rsId	Chr.	Location	ref	Alt.	Gene	Gene region	Translational Impact	GERP++ Score
rs369274154	6	33406128	T	C	MIR5004	5UTR		
rs371304188	6	33406147	C	T	MIR5004	5UTR		
rs375913209	6	33406168	C	T	MIR5004	5UTR		
Not assigned	6	33406194	A	C	MIR5004	5UTR	splice-disrupt	4.77
Not assigned	6	33406194	A	G	MIR5004	5UTR	splice-disrupt	4.77
Not assigned	6	33406194	A	T	MIR5004	5UTR	splice-disrupt	4.77
Not assigned	6	33406195	G	A	MIR5004	5UTR	splice-disrupt	4.77
Not assigned	6	33406195	G	C	MIR5004	5UTR	splice-disrupt	4.77
Not assigned	6	33406195	G	T	MIR5004	5UTR	splice-disrupt	4.77

Discussion

Leader sequence and TRS, in 5' non-coding part of SARS-CoV-2, can be considered as Achilles' heel of SARS-CoV-2. Leader sequence has been found at the 5' ends of all encoded transcripts that highlights its significance. TRS can explain the host range and pathogenicity of coronavirus. This study is a pioneering attempt to unravel the evolution of TRS and leader sequence in 5'UTR of SARS-CoV-2 and discover the inhibitory microRNAs for lowering virus load in infected cell.

UTRs, particularly 5'UTR, are of high translational importance. UTRs are potential sites for antiviral drugs to bind and inhibit virus replication. There is no report on disruption of 5'UTR in SARS-CoV-2. However, in bovine coronavirus, it has been found that disruption of stem-loop III and stem-loop IV of 5'UTR stops virus RNA replication, postulating that these regions function as cis-acting element (Raman *et al.*, 2003; Raman & Brian, 2005). On the other hand, microRNAs that bind to SARS-CoV-2 UTRs can be induced by drugs or food supplements to lower virus replication. Enhancing host microRNA defence machinery against 5'UTR region of virus can prevent SARS-CoV-2 infection. The above-mentioned strategies are rapidly achievable treatment strategies against COVID-19. In this study, we presented a model of literature mining-based drug discovery to induce inhibitory microRNAs against leader sequence and TRS of 5UTR of SARS-CoV-2.

hsa-miR-5004-3p was the unique human microRNAs with the ability to target the leader sequence of SARS and SARS-CoV-2. A decreased level of hsa-miR-5004-3p is reported during dengue virus (DENV) infection in blood of patients. Interestingly, hsa-miR-5004-3p was undetectable in early DENV infection, but expression was high in a majority of the healthy controls and recovered patients dengue. This finding demonstrated suppression of hsa-miR-5004-3p during the early phases of DENV infection and its importance in patient recovery from the viral infection (Tambyah *et al*, 2016). Dengue is the most common arboviral illness worldwide (Tambyah *et al.*, 2016) with high viral transmission success and immune evasion, like COVID-19. Due to the high stability of microRNAs in blood (Alanazi *et al*, 2018), we suggest that hsa-miR-5004-3p can be considered as a biomarker for sever COVID-19 infection in future studies.

We found a significant trend in 5'UTR of SARS-CoV-2 to escape from binding of hsa-MIR-5004-3p to its leader sequence by inducing insertion-type mutation. This mutation decreases microRNA-5'UTR binding stability and allows SARS-CoV-2 to escape the human microRNA immunity system. We suggest lack of innate human inhibitory microRNAs to bind to leader sequence and TRS of SARS-CoV-2 contributes to its high replication in infected human cells. On the other hand, mining of two hundred million deposited human genomic variants led us to discovery of splice-disrupt mutations in genomic structure of hsa-MIR-5004-3p. These mutations can negatively affect hsa-MIR-5004-3p function in preventing SARS-CoV-2 replication and can be considered as COVID-19 risk factors in future studies.

It should be noted that genetic signature of the pathogenesis severity in the non-coding regions of SARS-CoV-2 is unknown. Sequence variation in the non-coding regions of virus can predispose people to developing sever disease. Availability of 75,574 (at Sep 23, 2020) with full genome and high coverage in GISAID (https://www.gisaid.org/) (Shu & McCauley, 2017) and NCBI provides the chance of pattern recognition in 5'UTR sequences of SARS-CoV-2, particularly against host microRNA inhibitory machinery, by machine learning models. Models and statistics such as decision tree, classification based on association rule mining and deep learning (Ebrahimi *et al*, 2020; Hosseinpour *et al*, 2013; Kargarfard *et al*, 2015; Kargarfard *et al*, 2019; Kargarfard *et al*, 2016; Mohammadi-Dehcheshmeh *et al*, 2018) that has been used for eukaryotic promoter and UTR analysis can be examined for UTR analysis of SARS-CoV-2 as well. It should be noted that some of the sequences that have been deposited in However GISAID as full genomes have incomplete or low-quality sequencing in 3'UTR and 5'UTR regions.

Conclusion

Non-coding regions are crucial for SARS-CoV-2 replication, transcription, and dominating host systems biology. Unravelling the underpinning reasons of SARS-CoV-2 success in dominating human cell and its high transmission rate is crucial. This study is a significant step to unravel the evolution of 5'UTR in SARS-CoV-2, discover the key regions, and utilize the UTRs for lowering virus load in infected cells. We showed that hsa-MIR-5004-3p is a good human microRNA candidate to target the leader sequence of SARS-CoV-2 at 5'UTR region. Furthermore, we developed a literature mining-based drug repurposing strategy to induce inhibitory microRNAs against leader sequence of SARS-CoV-2. Activation of inhibitory machinery microRNAs by drug repurposing and food supplements are rapidly achievable treatment strategies against COVID-19.

In this study, an evolutionary pattern in 5'UTR of SARS-CoV-2 from SARS was discovered where SARS-CoV-2 tries to escape from hsa-MIR-5004-3p binding by generation of insertion-type mutation. Lack of human inhibitory microRNAs to target the 5' UTR of SARS-CoV-2 is one of the underpinning reasons of immunity evasion and dominating human cell by SARS-CoV-2. This study adds a new dimension to our knowledge on COVID-19 pathology and opens a new avenue for drug repurposing and activating innate microRNA inhibitory machinery against 5'UTR of SARS-CoV-2.

Conflicts of interest

The authors declare that there is no conflict of interest.

Acknowledgement

We would like to greatly thank Genomics Research Platform, College of Science, Health and Engineering of La Trobe University as well as School of Animal and Veterinary Sciences of The University of Adelaide for providing the computational resources to perform this study.

Methods

Data collection

Complete genomes of SARS-COV-2, SARS, MERS, Bat Coronavirus, and Bovine Coronavirus, including complete sequences of non-coding regions, were downloaded from NCBI. The sequences were: Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 (NC_045512), SARS coronavirus Tor2 (NC_004718), Tylonycteris bat coronavirus HKU4 isolate CZ01 (MH002337), Middle East respiratory syndrome-related coronavirus isolate NL13892 (MG987420), and Bovine coronavirus (NC_003045).

Inhibitory microRNA prediction

All available microRNA information of human and other organisms, deposited in TargetScan , were downloaded (Agarwal *et al*, 2015). In total, 9994 miRNA families, including sequence of seed region and mature sequence were used (Supplementary 4). Then the reverse complement sequence of UTR regions of SARS-COV-2, SARS, MERS, and bovine coronavirus were searched against the known seed sequences to find the miRNAs with affinity binding to the specific regions in 5'UTR, TRS and leader sequence.

The thermodynamic binding energy of –15 kcal/mol was used as the minimum cut-off, and binding energy of –25 kcal/mol pointed as the the stable binding (Ghoshal *et al*, 2015). All mature sequences of miRNAs were retrieved from miRBase database (Kozomara & Griffiths-Jones, 2014). Using RNAhybrid v2.2 database the binding probability of miRNAs were evaluated through calculating of a minimum free energy hybridization (Rehmsmeier *et al*, 2004).

Sequence alignment between 5'UTR of coronavirus and microRNA seed regions and phylogenetic trees

Multiple Sequence Alignment was performed using CLUSTALW algorithm by CLC Genomics Workbench 20 (QIGEN). Phylogenetic trees, based on Maximum Likelihood Phylogeny approach, were constructed using UPGMA method, Nucleotide substitution model of Jukes Cantor, Gamma distribution parameter of 1.0, and bootstrap of 1000.

Literature-mining based Drug repurposing

Pathway Studio Database (Elsevier) were employed to find the drugs, food supplements, and chemicals (small molecules) with promoting effects on 5' UTR inhibitory microRNAs, as described recently (Alanazi *et al.*, 2018). To perform literature mining, MedScan tool was used that employs NLP algorithm to mine extract relations from biomedical texts, mainly PubMed (Novichkova *et al*, 2003). In addition to the mined sentences of relations, Medscan also records the title of literature, authors, the year of publication, Medline (PubMed) reference number, and type of relation. The results are deposited in Mammalian + ChemEffect + DiseaseFx database of Pathway Studio which is enriched by a range of extra systems biology information like the subcellular location and protein class (such as receptor, ligand, transcription factor, small RNA, small molecule, etc.), from Gene Ontology Consortium, as well as KEGG pathways. The database updates using cloud technology by addition of new mined relationships and entities from recent publications. Statistics of Mammalian + ChemEffect + DiseaseFx database used for literature mining is presented at Table 4.

In short, a highly enriched database with more than million chemicals, 138000 proteins, and 12 million protein interactions were employed for drug repurposing.

Table 3. Statistics of Mammalian + ChemEffect + DiseaseFx database used for literature mining in this study (Sep 2020).

Entities	Number	Relations	Number

Small molecules (including drugs)	1053259	Binding	1123702
Protein	138106	Biomarker	120448
Cell process	9771	Cell expression	1213035
Cell Object	607	Chemical reaction	58888
Cells	4155	Clinical trial	109386
Clinical parameters	5126	Direct regulation	766707
Complex	998	Expression	832784
Diseases	20855	Functional associations	1775463
Functional class	5489	Genetic change	379261
Genetic Variant	127872	Molsynthesis	160178
Organ	3839	Moltransport	251347
Treatments	78	Promoter binding	44619
Tissue	574	Protein modification	73859
Total number of entities	**1370729**	Quantitative change	421884
		Regulation	5193796
		State change	128112
		MicroRNA effects	57743
		Total number of relations	

Multivariate analysis

Thermodynamic binding energy values (kcal/mol) between microRNAs and 5'UTR regions of coronaviruses were used as input for PCA and clustering analysis. Analysis was performed in MINITAB 18 (https://www.minitab.com). Graphs were visualized by GraphPad Prisim 7. Correlation matrix were used for PCA analysis. Clustering was performed based on Euclidean distance matrix and Average Linkage method.

Variant discovery on genomic sequence of hsa-MIR-5004-3p, 5'UTR inhibitory microRNAs, as COVID-19 risk factors

Human genetic variation database of NCBI (dbSNP database) (Sherry *et al*, 2001) was employed as the main resource for gathering of variants. Pathway Studio tool was used for retrieving hsa-MIR-5004-3p genomic variants by mining more than two hundred million deposited variants in dbSNP and 1000 Genomes project. Genomic locations of variants were recorded (CDs, 3′UTR, 5′UTR, intergenic, or intronic variants). Then, translational impact of the identified hsa-MIR-5004-3p variants including missense, splice disrupt, CDs indel, nonsense, misstart, and non-stop were determined. Finally, the proportion of each type of variant was calculated.

To evaluate the possible functional impact of the identified hsa-MIR-5004-3p variants, GERP++ conservation score was used. GERP is an evolutionary conservation score which have a good correspondence with clinical significance and pathogenicity level (Davydov *et al*, 2010). GERP++ demonstrates the constrained elements in multiple alignments by quantifying substitution deficits. These deficits identify substitutions that would have happened if the element were neutral DNA but did not happen as the element has been experienced functional constraint. Low values of GERP++ score stand for low level of conservation and high values for high level of conservation.

References

Agarwal V, Bell GW, Nam J-W, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *elife* 4: e05005

Alanazi IO, Al Shehri ZS, Ebrahimie E, Giahi H, Mohammadi-Dehcheshmeh M (2019) Non-coding and coding genomic variants distinguish prostate cancer, castration-resistant prostate cancer, familial prostate cancer, and metastatic castration-resistant prostate cancer from each other. *Molecular carcinogenesis* 58: 862-874

Alanazi IO, AlYahya SA, Ebrahimie E, Mohammadi-Dehcheshmeh M (2018) Computational systems biology analysis of biomarkers in lung cancer; unravelling genomic regions which frequently encode biomarkers, enriched pathways, and new candidates. *Gene* 659: 29-36

Chang Y, Zhao Y, Gu W, Cao Y, Wang S, Pang J, Shi Y (2015) Bufalin inhibits the differentiation and proliferation of cancer stem cells derived from primary osteosarcoma cells through Mir-148a. *Cellular Physiology and Biochemistry* 36: 1186-1196

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology* 6: e1001025

Ebrahimi M, Novikov B, Ebrahimie E, Spilman A, Ahsan R, Tahsili MR, Najafi M, Navvabi S, Shariaty F (2020) The first report of the most important sequential differences between COVID-19 and MERS viruses by attribute weighting models, the importance of Nucleocapsid (N) protein.

Ghoshal A, Shankar R, Bagchi S, Grama A, Chaterji S (2015) MicroRNA target prediction using thermodynamic and sequence curves. *BMC genomics* 16: 999

Gottwein E, Corcoran DL, Mukherjee N, Skalsky RL, Hafner M, Nusbaum JD, Shamulailatpam P, Love CL, Dave SS, Tuschl T (2011) Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. *Cell host & microbe* 10: 515-526

Hosseinpour B, Bakhtiarizadeh MR, Khosravi P, Ebrahimie E (2013) Predicting distinct organization of transcription factor binding sites on the promoter regions: a new genome-based approach to expand human embryonic stem cell regulatory network. *Gene* 531: 212-219

Kargarfard F, Sami A, Ebrahimie E (2015) Knowledge discovery and sequence-based prediction of pandemic influenza using an integrated classification and association rule mining (CBA) algorithm. *Journal of biomedical informatics* 57: 181-188

Kargarfard F, Sami A, Hemmatzadeh F, Ebrahimie E (2019) Identifying mutation positions in all segments of influenza genome enables better differentiation between pandemic and seasonal strains. *Gene* 697: 78-85

Kargarfard F, Sami A, Mohammadi-Dehcheshmeh M, Ebrahimie E (2016) Novel approach for identification of influenza virus host range and zoonotic transmissible sequences by determination of host-related associative positions in viral genome segments. *BMC genomics* 17: 1-10

Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* 42: D68-D73

Mohammadi-Dehcheshmeh M, Niazi A, Ebrahimi M, Tahsili M, Nurollah Z, Ebrahimi Khaksefid R, Ebrahimi M, Ebrahimie E (2018) Unified transcriptomic signature of arbuscular mycorrhiza colonization in roots of Medicago truncatula by integration of machine learning, promoter analysis, and direct merging meta-analysis. *Frontiers in plant science* 9: 1550

Novichkova S, Egorov S, Daraselia N (2003) MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* 19: 1699-1706

Raman S, Bouma P, Williams GD, Brian DA (2003) Stem-loop III in the 5′ untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. *Journal of virology* 77: 6720-6730

Raman S, Brian DA (2005) Stem-loop IV in the 5′ untranslated region is a cis-acting element in bovine coronavirus defective interfering RNA replication. *Journal of virology* 79: 12434-12446

Rehmsmeier M, Steffen P, Höchsmann M, Giegerich R (2004) Fast and effective prediction of microRNA/target duplexes. *Rna* 10: 1507-1517

doi:10.20944/preprints202009.0602.v1

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. *Nucleic acids research* 29: 308-311

Shu Y, McCauley J (2017) GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* 22: 30494

Skalsky RL, Corcoran DL, Gottwein E, Frank CL, Kang D, Hafner M, Nusbaum JD, Feederle R, Delecluse H-J, Luftig MA (2012) The viral and cellular microRNA targetome in lymphoblastoid cell lines. *PLoS Pathog* 8: e1002484

Tambyah PA, Ching CS, Sepramaniam S, Ali JM, Armugam A, Jeyaseelan K (2016) microRNA expression in blood of dengue patients. *Annals of clinical biochemistry* 53: 466-476

Wang Y, Feng L, Piao B, Zhang P (2017) Review on Research about Traditional Chinese Medicine in Cancer Stem Cell. *Evidence-Based Complementary and Alternative Medicine* 2017

Yang D, Leibowitz JL (2015) The structure and functions of coronavirus genomic 3′ and 5′ ends. *Virus research* 206: 120-133