The Analysis of the 5’ Untranslated region mRNA Secondary Structure of SARS-COV-2 and its related genomic strains

Meeraj Amin and Chun Wu
Chemistry and Biochemistry Department, Rowan University, Glassboro, NJ 08028
Emails: ma1512@scarletmail.rutgers.edu; wuc@rowan.edu

This work was supported by Rowan Startup and SEED grant, and the National Science Foundation under Grant NSF ACI-1429467 and XSEDE MCB177088.

Abstract

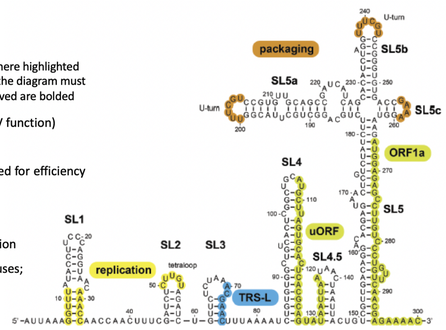
Over 4.5 million people have died as a result of SARS-CoV-2 worldwide since its first outbreak. To know the origin of SARS-CoV-2 is critical for finding solutions to end the pandemic. After the outbreak, the genomes of ~20 SARS-COV2-like viruses including RacCS203 have been reported to support the natural origin of SARS-COV-2. However, twitter scholar @Daoyu15 has pointed out that the 5’ untranslated region (UTR) of RacCS203 (91.5% sequence identity) likely adopts an abnormal secondary structure, leading to loss of function for a live virus. To examine this issue, in this study we observed the secondary structure of the 5’ UTR for this set of genomes with reference to a set of bat SARS-COV-2-like genomes (~50) collected before the outbreak. Using RNA secondary structure prediction software, the set of pre-pandemic genomes supported a functional 5’ UTR (six parts) which provides instructions for successful viral infection. Emphasis is placed at SL2 which is vital for forming the replication complex and SL3 which contains the leader transcriptional regulatory sequence (TRS-L) that is needed for synthesis of sub-genomic RNA (Miao, Tidu et al. 2021). Without a functional TRS-L, necessary proteins (S,E,M,N…etc) are not translated hence creating a non-living virus (Nomburg, Meyerson et al. 2020). In contrast, within the post-pandemic set, five 5’ UTR secondary RNA structures (RacCS2xx genomes) showed an overlap between SL1, SL2, and SL3 in a singular hairpin structure. This hairpin structure does not allow for a living virus to exist due to a non-accessible 5’ UTR. A deeper examination into viral genome functionality must be taken into account since these post-pandemic genomes were used to support a natural origin. These results indicate that fabrication might have taken place in releasing these genome sequences hence making them unusable in proving SARS-CoV-2’s origin.

Background

5’ Untranslated Region

Regions where highlighted portions in the diagram must be conserved are bolded

- ❖ Unconventional translation initiation
- ❖ Encodes ORF1a/b & sgRNA (vital genes for SARS-2-COV function)
  - **SL1** - loop not conserved in variants; aids in replication but loop & bulge not required for efficiency
  - **SL2** - Involved in replication complex formation
  - **SL3** – Leader TRS (TRS-L); needs to be accessible for successful replication
  - SL4 – Start codon of uORF present in all coronaviruses; may be involved sgRNA synthesis
  - SL4.5 - Rest of SL4 uORF; uORF another method of translation regulation
  - SL5 – Four way junction present in all coronaviruses; involved in RNA packaging



Six regions of 5’ UTR SARS-COV-2

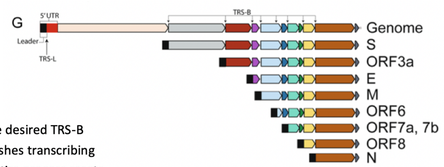
Subgenomic RNA

ORFs = open reading frames (code for NSPs)
NSP = non-structural protein (facilitate transcription & replication)

- ❖ Needed due to:
  - large viral genome size
  - downstream ORFs far from 5’ end of genome
- ❖ 66% canonical (dependent on TRS-L) / 33% non-canonical sgRNAs present
- ❖ 8 major sgRNAs generated using TRS & TRS-L sequences

TRS-L

- ❖ Transcriptional regulatory sequence leader
- ❖ Core Sequence – ACGAAC
- ❖ How sgRNA are made using TRS-L:
  - RDRP progresses 3’ to 5’ until it reaches the desired TRS-B
  - RDRP jumps to the TRS-L sequence and finishes transcribing
  - The antisense strand is released and the entire process resets
  - The antisense strand is transcribed to create a (+) strand which is then translated



8 major SgRNAs

Genome Sequence

- ❖ Examined ~62 genomes’ 5’ UTR secondary Structure (~300 nt)
- ❖ Wuhan-Hu-1 (MN908947) served as reference (also known as SARS-2-COV)
- ❖ Outliers in particular include RacCs2xx variants (bat CoV; 91.5% identity)
- ❖ RaCS2xx variants used to support a natural origin theory

@Daoyu15

- ❖ First highlighted the secondary structure abnormality (Mar ‘21)
- ❖ S-gene unable to bind to any ACE2 receptors (how virus harms body)

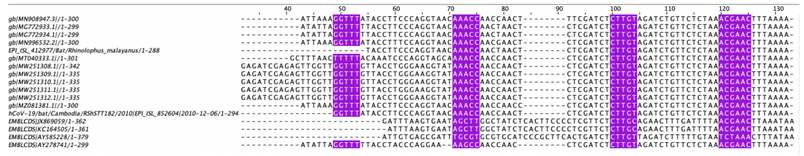
Purpose

1. To examine 5’ UTR secondary structure abnormalities in a larger data set
2. To highlight how inaccessible structures make a virus functionless
3. To counter the validity of genomes being used to support a natural origin

Current Research

Methods

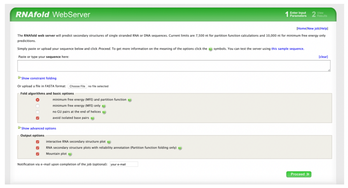
- ✓ Compile complete genome sequences of ~62 CoV genomes
  - Consisting of:
    - ✓ SARS-2-COV, MERS, SARS, bat species, pangolin species, etc…
    - ✓ Genomes released before and after COVID-19 outbreak
- ✓ Edit the genomes to include only 5’ UTR
  - about 300 nt; SARS-2-COV 5’ UTR used as reference size

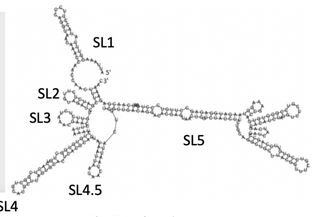


Required nt sequences for specific regions highlighted
*Only the most important sequences shown

SL1 SL1 SL2 SL3

- ✓ Calculate secondary structures using an online prediction software
  - In this case, http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi
- ✓ Compile images for all structures; label 3’ & 5’ ends, 6 parts



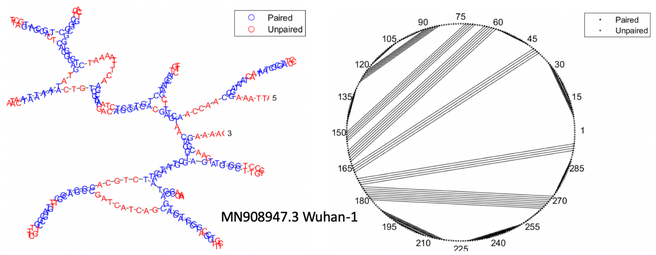


5’ UTR of Wuhan-Hu-1

This particular software was chosen due to its ability to form consistent results with multiple high quality image outputs available. Other softwares were used to check structures for precision and for any structural bias due to software choice.

MATLAB simulations serve as an alternate method to produce the same results and drawings to ensure accuracy. This redundancy provided more evidence for portraying the structural abnormality present in certain strains.

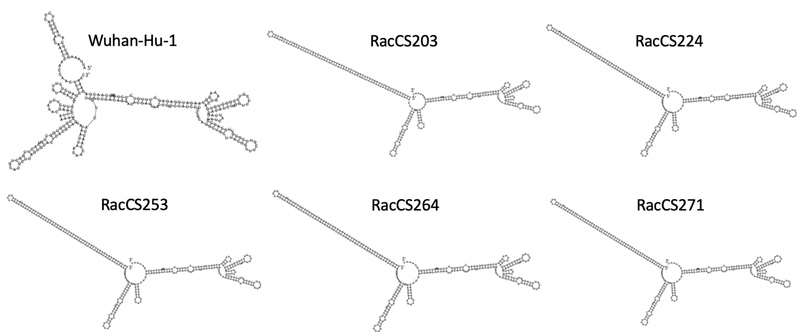
- ✓ Other forms of displaying secondary structures were obtained using MATLAB
  - Results were output in various diagrams and figures
  - Examples below:



MN908947.3 Wuhan-1

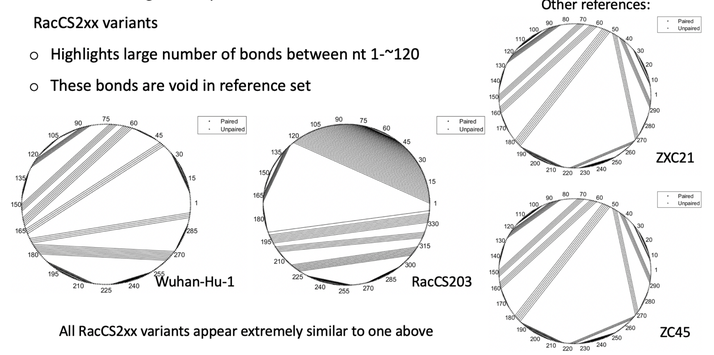
Results

❖Secondary structures of five RacCS20XX variants exhibit a long hairpin



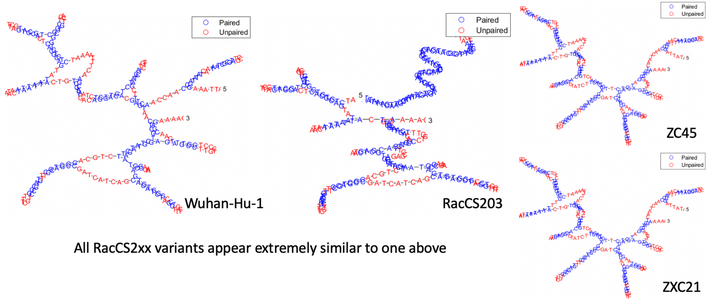
- ❖ MATLAB circle figures emphasize difference between reference and 5 RacCS2xx variants
  - Highlights large number of bonds between nt 1-~120
  - These bonds are void in reference set

Other references:



All RacCS2xx variants appear extremely similar to one above

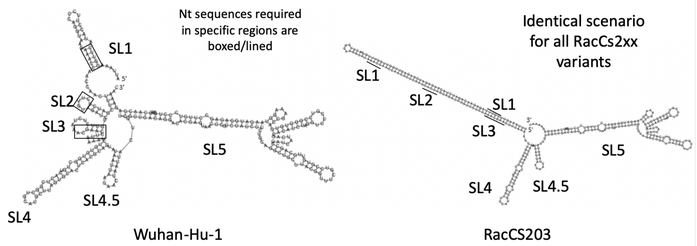
- ❖ MATLAB diagrams emphasize difference between reference and 5 RacCS2xx variants
  - Large nt clustering seen in RacCS2xx variants (top right)
  - Reference set shows more stems/loops exhibiting an uninhibited structure



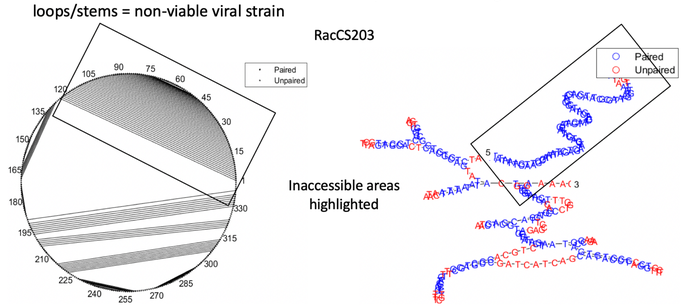
All RacCS2xx variants appear extremely similar to one above

Conclusion

- ❖ Hairpin structure verifies that SL1, SL2, & SL3 are not accessible for transcription
- ❖ If these regions are not accessible, no sgRNAs can be made
- ❖ Renders the virus useless since vital proteins cannot be made



- ❖ Supplementary figures show a lack of accessibility due to clustering or non-existential loops/stems = non-viable viral strain



Overall, RacCs2xx strains presented their secondary structure with a hairpin containing regions SL1, SL2, SL3. These strains were used to prove a natural lab origin, but in reality they are non-viable. This disputes the natural origin theory leading to the possibility of a lab origin. In a follow up study, a larger dataset should be included while creating tertiary structures to study mRNA interactions. In a perfect world, anyone would be able to access/sequence all the genomes in person to find any possible sequencing errors. Email for works cited and full data set.