An Integrated Computational Framework for Replication Site and Regulatory Motif Discovery in Coronavirus Genomes

Pushpa Susant Mahapatro[1], Basant Kumar[2],

[1] Lincoln University College, 47301, Petaling Jaya, Selangor Darul Ehsan, Malaysia;

[2] Modern College of Business and Science, Oman;

[1] pdf.pushpa@lincoln.edu.my, [2] basant@mcbs.edu.om

Abstract: Coronavirus genomes contain regulatory regions that play an important role in genome replication, transcription, and other stages of the viral life cycle. Identifying these regions computationally is difficult because coronavirus sequences show considerable variation, while regulatory motifs are often short, weakly conserved, and susceptible to nucleotide changes. Although several computational approaches have been used to study regulatory sequences in coronaviruses, there is still limited comparative evaluation of motif discovery methods when the target motifs contain mutations or non-contiguous patterns. This study proposes a computational framework for locating candidate replication-associated regions and identifying regulatory motifs in coronavirus genomes. Complete coronavirus genome sequences will be collected from publicly available databases and subjected to quality filtering, alignment, and comparative analysis. Genome-wide sequence characteristics and conservation patterns will be examined to identify regions that may be associated with replication. Four motif discovery techniques—Greedy Motif Search, Greedy Motif Search with Pseudocounts, Randomized Motif Search, and Gibbs Sampling—will be evaluated using a common experimental framework. The analysis will specifically consider motifs containing one or two nucleotide mutations, as well as non-contiguous patterns that may be overlooked by conventional exact motif searches. The performance of the four approaches will be assessed in terms of motif identification accuracy, consistency, tolerance to mutations, and computational efficiency. Candidate motifs will further be assessed against conserved genomic regions, existing annotations, and available biological and structural evidence. The framework is intended to provide a reproducible approach for prioritizing potentially important coronavirus genomic regions for further experimental investigation. Such computationally prioritized regions may subsequently support studies of antiviral targets and contribute to genomic surveillance of emerging coronavirus variants.

Keywords: Coronavirus, genome replication, regulatory motifs, motif discovery, Greedy Motif Search, Gibbs Sampling.

Introduction

Coronaviruses are positive-sense single-stranded RNA viruses with large genomes, typically 26-32 kb in size [1]. Their genomes contain coding regions as well as untranslated and regulatory regions that work together to support viral replication, transcription, translation, and genome packaging. Replication is carried out by a viral replication–transcription complex, in which the RNA-dependent RNA polymerase

and other non-structural proteins interact with specific regions of the viral RNA [2]. Understanding where these functionally important regions occur within the genome is therefore important for studying coronavirus biology and for identifying genomic features that may have therapeutic relevance.

Coronavirus gene expression differs from that of many other RNA viruses because it involves a discontinuous transcription mechanism. During this process, the viral polymerase interacts with transcription regulatory sequences (TRSs) while producing a nested set of subgenomic RNAs. Most people notice that TRSs usually sit near the leader region or just before the accessory genes. The position and sequence characteristics of transcription regulatory sequences (TRSs) make them useful targets for computational analysis.

Besides TRSs coronavirus RNA has parts and shapes that change how the virus copies itself makes proteins and packs its genome. When scientists look at SARS-CoV-2 they see RNA shapes and patterns that stay the same throughout the whole genome. This tells me that the virus uses both the order of its letters and the way those letters fold to send messages. Because of this if we only look at the parts we already know about, we might miss important patterns that help the virus work.

Motif discovery is one way to look at these sequence patterns in a way. The job is hard with viral genomes because the regulatory motifs are usually small and might not be the same in different versions of the virus. Changes, in the sequence can change one or more parts of a motif. Not always stop it from doing its job. Because of these techniques that rely on matching patterns or looking for very similar sequences might miss some important motifs.

There are different ways to solve this problem. Greedy Motif Search works by building a motif one step at a time from pattern ideas. Another way is to use pseudocounts, which stops the math from hitting zero when making position- profiles. Randomized Motif Search uses a bit of luck to start. Then makes things better over time. Then there is Gibbs Sampling, which uses a math-based search to look for motif shapes. Even though people use these algorithms all the time in biology no one has really tested which algorithm is best, at finding coronavirus motifs when mutation are present. People have not yet put these algorithms into one test to see how they handle coronavirus motifs.

This study therefore focuses on finding parts of coronavirus genomes that are linked to replication and on identifying patterns in the genetic code. The idea is to use computer methods to look at the genome and find these patterns. The plan uses four ways to find these patterns: Greedy Motif Search, Greedy Motif Search with Pseudocounts, Randomized Motif Search and Gibbs Sampling. Special care is taken to look for patterns that're not next to each other and patterns that have one or two changes in the letters. I will check the patterns found using numbers that show how well they work and compare these numbers with what we know about the genetic structure of the virus and how the virus behaves.

Related work

2.1 Coronavirus Genome Organization and Replication

Coronavirus genomes are among the largest RNA virus genomes and contain different regions that contribute to various stages of the viral life cycle. Along with protein-coding sequences, they include untranslated regions and cis-acting RNA elements involved in processes such as replication, transcription, translation, and genome packaging. During RNA synthesis, the viral replication–transcription complex interacts with these genomic regions. Therefore, examining the genome sequence can provide useful information about the mechanisms involved in coronavirus replication.

Studies conducted in recent years have provided increasing evidence that coronavirus replication is associated with particular sequence and structural features located at different positions within the genome. Research on SARS-CoV and SARS-CoV-2 has reported conserved RNA structures near the genomic termini, while studies of MERS-CoV have shown that higher-order RNA structures within the nsp1 coding region can influence transcription and replication. These observations suggest that replication-associated information may be distributed across the genome rather than being restricted to a single promoter-like sequence. This makes genome-wide computational analysis important for identifying sequence features that may be associated with coronavirus replication.
More recent experimental work has also identified promoter elements associated with the SARS-CoV-2 RNA-dependent RNA polymerase complex and demonstrated their role in RNA synthesis. Such findings are relevant to computational studies because they provide experimentally supported regions against which computationally identified candidate replication-associated regions can eventually be evaluated.

2.2 Transcription Regulatory Sequences and Coronavirus Regulatory Elements
Coronavirus gene expression involves the production of genomic and subgenomic RNAs. Transcription regulatory sequences (TRSs) play a central role in this process by directing discontinuous transcription. These sequences are located in the leader region and near the beginning of downstream open reading frames, although their sequence characteristics can differ among coronavirus species and genomic locations.
Computational approaches have been developed to identify TRSs and distinguish potential regulatory regions from other genomic sequences. One notable approach formulated coronavirus TRS identification using sequence information together with constraints derived from genome organization. Such approaches show the importance of adding traits into computer-based sequence analysis instead of looking at motif finding just as a general pattern-finding task.
At the time research on SARS-CoV-2 has shown that control signals can change over time. New TRS-like sequences have been seen in virus groups. Some of them create extra subgenomic RNAs. This finding is very important for the study because it shows how changes, in the virus can make control motifs or change existing ones while the virus keeps reproducing well.

2.3 RNA Secondary Structure and Cis-Acting Elements
Just looking at the sequence information is not enough to understand all the information hidden in coronavirus RNA. RNA molecules do not stay in a line. Instead, RNA molecules fold into shapes and higher-order structures. These shapes are important because these structures can decide how viral proteins and other RNA molecules interact with the genome.

When scientists look at the genome of SARS-CoV-2 they find structured regions. These structured regions include parts, near the 5’ and 3’ ends of the genome that stay the same. Some of these structures are also found in coronaviruses. This makes me think that the fact these structures stay the same over time means they serve a purpose. Other studies have also found RNA structures, inside the coding regions. These RNA structures can change how the virus handles transcription or replication.

Research on MERS-CoV provides particularly relevant evidence. Experimental alteration of specific stem-loop structures in the nsp1 region affected viral transcription and genomic RNA synthesis, demonstrating that RNA structure can have a functional role even when the corresponding region is located within a protein-coding sequence. These observations support the use of sequence conservation and structural information as complementary evidence when evaluating computationally identified motifs.

2.4 Computational Motif Discovery

Motif discovery is a classical problem in computational biology. The goal is to find short sequence motifs that are repeated within a set of biological sequences and that may represent functional signals. The problem becomes considerably more difficult when the motif is short, degenerate, weakly conserved, or contains mutations.

Several algorithmic strategies have been developed to address this problem. Greedy Motif Search uses an iterative profile-building strategy to identify a set of sequences that produces a high-scoring motif. The use of pseudocounts modifies the profile calculation and reduces the effect of zero probabilities, which can be important when relatively small datasets contain nucleotide variation.

2.5 Mutation and Motif Conservation

RNA viruses accumulate nucleotide changes over time, and coronavirus populations contain substantial sequence diversity. A regulatory motif may therefore appear in several related forms rather than as an identical sequence across all genomes [3].

This creates a difficulty for conventional motif identification. If a nucleotide substitution occurs at a position that contributes strongly to the motif score, the resulting sequence may no longer appear sufficiently similar to the consensus pattern. An algorithm designed to identify highly conserved motifs may consequently miss a potentially meaningful variant of the original motif.

The problem is particularly important when the objective is genomic surveillance. A motif that is conserved in most genomes but contains a substitution in a particular lineage may provide information about evolutionary change. Conversely, a newly occurring motif may potentially influence viral gene expression or fitness. Recent research on SARS-CoV-2 TRS evolution provides an example of how changes in regulatory sequence patterns can be associated with the emergence of novel subgenomic RNAs.

2.6 Non-Contiguous Motifs

Most conventional motif discovery approaches are formulated around contiguous sequence patterns. In biological systems, however, functional information does not always depend on every nucleotide occurring consecutively. Important positions within a motif may be separated by variable nucleotides, structural elements, or positions that can tolerate mutations [4]. This makes motif identification more

difficult because a computational method may detect short conserved regions without recognizing that these regions are part of a larger functional pattern [5].

Particular attention is given to motifs containing one or two nucleotide changes and to non-contiguous patterns in which important positions may be separated by variable nucleotides. The objective is not to assign biological function solely on the basis of computational scores [6]. Instead, the predicted motifs are intended to be treated as candidate regions that require further assessment using sequence conservation and other relevant biological evidence [7]. Instead, the identified candidates will be compared with how the sequence is conserved, with genomic annotations with structural details and with coronavirus regulatory elements that were already known. This approach aims to create a trustworthy list of candidate regions, for future experiments and to explore how they might help with finding antiviral targets and monitoring coronavirus genomes.

Research Gap

Based on the reviewed literature, the following research gaps have been identified:

1. Limited genome-wide identification of replication-associated regions

Although coronavirus replication mechanisms and several cis-acting regulatory elements have been studied extensively, there is still limited work on systematically analyzing complete coronavirus genomes to identify candidate replication-associated regions using computational sequence characteristics [8].

2. Limited comparative evaluation of motif discovery algorithms

Existing studies have used different computational approaches for identifying coronavirus regulatory sequences. However, a systematic comparison of Greedy Motif Search, Greedy Motif Search with Pseudocounts, Randomized Motif Search, and Gibbs Sampling under the same experimental conditions remains limited [9].

3. Difficulty in identifying mutation-containing motifs

Coronavirus genomes continuously accumulate mutations. A nucleotide change within a short regulatory motif can reduce its similarity to the consensus sequence, potentially causing conventional motif discovery methods to overlook the motif. The ability of different algorithms to recover motifs containing one or two mutations therefore requires further investigation.

4. Insufficient investigation of non-contiguous motifs

Most conventional motif discovery approaches are designed to identify contiguous sequence patterns. Regulatory information, however, may be distributed across positions separated by variable or mutated nucleotides. The computational identification of such non-contiguous motifs in coronavirus genomes remains insufficiently explored.

5. Lack of an integrated computational framework

Replication-associated region analysis, motif discovery, mutation analysis, algorithm comparison, and biological validation are commonly addressed as separate problems. There is a need for an integrated and reproducible framework that brings these components together [10].

6. Limited systematic validation of computationally identified motifs

A computationally significant motif does not necessarily represent a biologically functional regulatory element. Thorough checking using how genes are kept the same across species what parts of the DNA do how RNA folds and known parts that control gene activity is required.

Even though there has been a lot of progress, in learning about how coronaviruses copy themselves and how their RNA parts control things there is still not a way to use computers to find parts of the genome that are involved in copying. There is also not work on comparing different ways to find patterns especially when those patterns have changes or are not altogether.

Method, Experiments and Results

4.1 Research Design

The study uses a computer-based and comparative research design to analyse coronavirus genome sequences and to find candidate replication-associated regions and regulatory motifs. The workflow includes genome sequence collection, quality control, sequence preprocessing, genome-level analysis, motif discovery, mutation-aware motif analysis, comparison of motif discovery algorithms and validation of the identified candidates. The methodology is designed so that the four chosen motif discovery algorithms are evaluated with input data and performance measures.

4.2 Genome Sequence Dataset

Complete coronavirus genome sequences will be collected from publicly available genomic databases. NCBI GenBank will be used as the primary source, while GISAID may also be considered for obtaining SARS-CoV-2 sequences and information related to circulating variants [11, 12].

4.3 Inclusion and Exclusion Criteria

Inclusion criteria:

- Full coronavirus genome sequences [13, 14].
- Sequences that have enough nucleotide quality for us to run computer analysis.
- Genomes that do not have confusing or unclear nucleotides.
- Sequences that have annotation when it is possible to find it.
- Representative sequences that cover the coronavirus groups or coronavirus variants.

Exclusion criteria:

- Incomplete or fragmented genome sequences.
- Sequences containing excessive ambiguous nucleotides.
- Poor-quality sequences.
- Duplicate or highly redundant sequences.
- Sequences that do not belong to the selected coronavirus groups.

4.4 Randomized Motif Search

Randomized Motif Search will begin with randomly selected k-mers and iteratively refine the motif using a profile-based scoring procedure [15]. Because the result can depend on the initial random selection, multiple independent runs will be performed. The best-scoring motifs across repeated runs will be retained for subsequent analysis.

4.5 Gibbs Sampling

Gibbs Sampling will use a probabilistic iterative procedure for motif discovery. At each iteration, one sequence will temporarily be excluded from the motif model, and a new candidate position will be selected according to the probability generated from the remaining sequences [16]. Repeated runs will be used to examine the stability of the motifs obtained through the stochastic sampling process [17].

4.6 Mutation-Aware Motif Identification

A specific component of the methodology will examine the ability of the algorithms to identify motifs containing one or two nucleotide mutations. Known or computationally generated motif variants will be analyzed by introducing controlled nucleotide substitutions at selected motif positions. The resulting sequences will then be supplied to the motif discovery algorithms under comparable conditions [18]. The study will examine whether the original motif pattern can still be identified when one or two nucleotide positions are altered. For each mutation scenario, the position of the mutation, the resulting motif score, and the ability of each algorithm to recover the underlying motif will be recorded. This controlled experiment will provide a basis for comparing the tolerance of the four motif discovery algorithms to nucleotide mutations.

4.7 Non-Contiguous Motif Identification

A further limitation is the difficulty of identifying motifs that contain nucleotide mutations or have non-contiguous positions. Such patterns may be missed when computational methods primarily depend on continuous and highly conserved sequence patterns. A high computational score does not, by itself, establish that a predicted motif functions as a regulatory element; therefore, computational predictions need to be supported by additional biological evidence. To be confident, about a motif you really need to examine conservation, RNA structure, genomic context, RNA–protein interaction or experimental results [19, 20].

4.8 Comparative Evaluation of Algorithms

The four algorithms will be tested against the datasets. When possible, I will make sure that I use the same motif length and parameters in my analysis for all of them. It will be analyzed with the following measures:

- Motif identification accuracy
- Motif score
- Conservation across sequences
- Mutation tolerance
- Ability to identify motifs
- Reproducibility across repeated runs
- Computational execution time

4.9 Validation of Identified Motifs

The computer discovered patterns will go through checking. Possible patterns will be looked at alongside:

- Known coronavirus regulatory elements.
- Conserved genomic regions.
- Available genomic annotations.
- Experimentally characterized regulatory regions reported in the literature.

- Predicted or experimentally determined RNA secondary structures, where available.
- Conservation patterns across related coronavirus genomes.

Just because a motif gets a score from a computer does not mean that the motif is actually working in a biological way. I believe we need to look at different pieces of proof before we decide which motifs are worth spending time on, for more study.

4.10 Statistical Analysis

The outcomes from running the algorithm times will be combined using simple numbers that show what occurred. For algorithms that have parts we will examine how the results change when the algorithm is run over and over again. This will help us determine if the results are consistent or not. Numbers that show how each algorithm performs under changes will be compared. This will help us understand how adding one or two changes affects each algorithm. If the numbers meet the conditions, we will use tests to see if the differences between algorithms are real and not just due, to chance.

4.11 Reproducibility

Reproducibility is the goal. The study will record genome accession numbers, sequence-selection criteria, preprocessing procedures, motif length, algorithm parameters, number of iterations number of runs, mutation conditions and evaluation criteria. Reproducibility also demands that the computational implementation be shared in a repository and that repository must follow the terms of the databases that were used. By doing this reproducibility will let other researchers reproduce the experiments and apply the methodology to coronavirus genomes.

4.12 Overall Methodological Workflow

The whole process can be summed up like this:

Coronavirus Genome Collection → Quality Control → Sequence Preprocessing → Multiple Sequence Alignment → Genome-Level Analysis → Candidate Replication-Associated Region Identification → Motif Discovery → Mutation-Aware Motif Analysis → Motif Analysis → Algorithm Comparison → Biological/Structural Validation → Candidate Prioritization

This methodology will help you reach your research goals. This methodology also makes sure there is a link, between the candidates found by computers and the biological functions that are proven in a lab.

Discussions

- The current study focuses on searching for candidate regions that are connected to replication and regulatory elements in coronavirus genomes using a computer-based approach. The approach combines a search with motif hunting to solve a problem in tools that only rely on known regulatory sequences. By looking at genome data, the study provides an opportunity to examine pattern clues in both unknown sections of coronavirus genomes. Candidate regions and motifs can subsequently be compared with known regulatory elements, genomic annotations, conserved regions, and available RNA structural information. Agreement between multiple sources of evidence would provide stronger support than relying on the score produced by a single computational method.

The findings may also have relevance to genomic surveillance. Monitoring mutations occurring within candidate regulatory motifs could help identify recurring patterns as well as newly emerging sequence variations. However, the presence of a mutation within a predicted motif should not, by itself, be interpreted as evidence of altered pathogenicity, transmissibility, or viral fitness.

From an antiviral research perspective, the proposed framework can be used to prioritize candidate regulatory regions for subsequent experimental investigation. Conserved regions that are repeatedly identified across genetically diverse coronavirus genomes may be of particular interest. Similarly, motifs associated with replication or transcription could provide potential targets for future studies aimed at disrupting viral RNA functions.

Conclusions

Overall, the proposed framework may contribute to more systematic analysis of coronavirus regulatory sequences and provide a computational basis for future studies of replication-associated regions, genomic surveillance, and potential antiviral targets. The study also highlights that computational identification should be seen as a step for generating and selecting candidates not as proof of biological function. Checking through sequence conservation genomic information, RNA structure details and known regulatory elements can help increase trust in the candidates that are found.

The framework that is suggested could be useful in tracking coronavirus genomes and working on treatments especially by helping to focus on parts of the genome that are the same or might have a role in function. More experiments will be needed to prove the role of the parts that are found through computation. In total the study offers a way to combine analysis of regions linked to replication finding patterns that take mutations into account finding patterns that're not next, to each other comparing different methods and checking the results with real biology. This approach can be used with coronavirus genomes and might also be changed to look at other fast-changing RNA viruses.

Future work will focus on benchmarking the proposed computational framework, with a large and diverse coronavirus genome dataset. The computational framework will be evaluated against the coronavirus genome dataset. Extending motif detection to multiple mutation positions and variable-length non-contiguous motifs can be performed.

References

1. A. R. Fehr and S. Perlman, “Coronaviruses: An overview of their replication and pathogenesis,” Methods Mol. Biol., vol. 1282, pp. 1–23, 2015, doi: 10.1007/978-1-4939-2438-7_1.
2. M. M. Ba Abduallah and M. G. Hemida, “Comparative analysis of the genome structure and organization of the Middle East respiratory syndrome coronavirus (MERS-CoV) 2012 to 2019 revealing evidence for virus strain barcoding, zoonotic transmission, and selection pressure,” Rev. Med. Virol., vol. 31, no. 1, pp. 1–12, 2021, doi: 10.1002/rmv.2150.

3. S. Duffy, “Why are RNA virus mutation rates so damn high?,” PLoS Biol., vol. 16, no. 8, Art. no. e3000003, 2018, doi: 10.1371/journal.pbio.3000003.
4. J. Hu, B. Li, and D. Kihara, “Limitations and potentials of current motif discovery algorithms,” Nucleic Acids Res., vol. 33, no. 15, pp. 4899–4913, 2005, doi: 10.1093/nar/gki791.
5. G. Z. Hertz and G. D. Stormo, “Identifying DNA and protein patterns with statistically significant alignments of multiple sequences,” Bioinformatics, vol. 15, nos. 7–8, pp. 563–577, 1999, doi: 10.1093/bioinformatics/15.7.563.
6. C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, “Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment,” Science, vol. 262, no. 5131, pp. 208–214, 1993, doi: 10.1126/science.8211139.
7. T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, “MEME SUITE: Tools for motif discovery and searching,” Nucleic Acids Res., vol. 37, no. Web Server issue, pp. W202–W208, 2009, doi: 10.1093/nar/gkp335.
8. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” J. Mol. Biol., vol. 215, no. 3, pp. 403–410, 1990, doi: 10.1016/S0022-2836(05)80360-2.
9. A. R. Panchenko and S. H. Bryant, “A comparison of position-specific score matrices based on sequence and structure alignments,” Protein Sci., vol. 11, no. 2, pp. 361–370, 2002, doi: 10.1110/ps.19902.
10. B. P. Steil and D. J. Barton, “Cis-active RNA elements (CREs) and picornavirus RNA replication,” Virus Res., vol. 139, no. 2, pp. 240–252, 2009, doi: 10.1016/j.virusres.2008.07.027.
11. Y. Shu and J. McCauley, “GISAID: Global initiative on sharing all influenza data—from vision to reality,” Euro Surveill., vol. 22, no. 13, Art. no. 30494, 2017, doi: 10.2807/1560-7917.ES.2017.22.13.30494.
12. D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank,” Nucleic Acids Res., vol. 41, no. D1, pp. D36–D42, 2013, doi: 10.1093/nar/gks1195.
13. The UniProt Consortium, “UniProt: The universal protein knowledgebase in 2023,” Nucleic Acids Res., vol. 51, no. D1, pp. D523–D531, 2023, doi: 10.1093/nar/gkac1052.
14. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The Protein Data Bank,” Nucleic Acids Res., vol. 28, no. 1, pp. 235–242, 2000, doi: 10.1093/nar/28.1.235.
15. H. Zhang, S. Li, L. Zhang, D. H. Mathews, and L. Huang, “LazySampling and LinearSampling: Fast stochastic sampling of RNA secondary structure with applications to SARS-CoV-2,” Nucleic Acids Res., vol. 51, no. 2, Art. no. e7, 2023, doi: 10.1093/nar/gkac1029.
16. T. Ling-Hu, E. Rios-Guzman, R. Lorenzo-Redondo, E. A. Ozer, and J. F. Hultquist, “Challenges and opportunities for global genomic surveillance strategies in the COVID-19 era,” Viruses, vol. 14, no. 11, Art. no. 2532, 2022, doi: 10.3390/v14112532.
17. C. Zhang, P. Sashittal, M. Xiang, Y. Zhang, A. Kazi, and M. El-Kebir, “Accurate identification of transcription regulatory sequences and genes in coronaviruses,” Mol. Biol. Evol., vol. 39, no. 7, Art. no. msac133, 2022, doi: 10.1093/molbev/msac133.
18. N. C. Huston, H. Wan, M. S. Strine, R. de Cesaris Araujo Tavares, C. B. Wilen, and A. M. Pyle, “Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory

motifs and mechanisms," Mol. Cell, vol. 81, no. 3, pp. 584–598.e5, 2021, doi: 10.1016/j.molcel.2020.12.041.
19. T. C. T. Lan, M. F. Allan, L. E. Malsick, et al., "Secondary structural ensembles of the SARS-CoV-2 RNA genome in infected cells," Nat. Commun., vol. 13, Art. no. 1128, 2022, doi: 10.1038/s41467-022-28603-2.
20. M. Soszynska-Jozwiak, A. Ruszkowska, R. Kierzek, C. A. O'Leary, W. N. Moss, and E. Kierzek, "Secondary structure of subgenomic RNA M of SARS-CoV-2," Viruses, vol. 14, no. 2, Art. no. 322, 2022, doi: 10.3390/v14020322.