New Short RNA Motifs Potentially Relevant in the Severe Acute Respiratory Syndrome Coronavirus 2 Genome

Miguel Angel Fuertes (mafuertes@cbm.csic.es)  
Spanish National Research Council

Carlos Alonso  
Spanish National Research Council

Research Article

Keywords: COVID-19, SARS-CoV-2, evolution, triplet composon, coronavirus

Posted Date: October 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-952337/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
New short RNA motifs potentially relevant in the severe acute respiratory syndrome coronavirus 2 genome

Miguel Angel Fuertes¹, Carlos Alonso¹†

¹Miguel Angel Fuertes (✉) and ¹Carlos Alonso; Centro de Biología Molecular “Severo Ochoa” (CSIC-UAM), Universidad Autónoma de Madrid, c/Nicolás Cabrera 1, 28049 Madrid, Spain. E-mails: mafuertes@cbm.csic.es.

Abbreviations: NT, nucleotide; tCP, triplet composon; SARS, severe acute respiratory syndrome, COVID-19, coronavirus disease 2019; orf, open reading frame; pp, polyprotein; NSp, non-structural protein

Keywords: COVID-19, SARS-CoV-2, evolution, triplet composon, coronavirus

† In Memoriam: Prof. Carlos Alonso Bedate.
Abstract

As time passes, identifying new pharmacological targets is becoming more difficult. Shortly, it will be necessary to devise new strategies to tackle the problem. The coronavirus disease outbreak caused by the severe acute respiratory syndrome coronavirus 2, represents a threat to human health serving as example from what we just said. The present study was aimed to collect a set of short RNA motifs with potential biological impact, most of which have not been observed heretofore. Categorizing RNA triplets by their gross-composition, the study collected 88 short RNA motifs, shared by most coronavirus genera independent on the percent identity between genomes. Selected motifs contain all nearest-neighbours of the triplets A, T, G and A, C, G. The high percent identity between severe acute respiratory syndrome coronavirus genomes makes it difficult these peptides to be found by current methods. The results provide 50 motifs in the 1a polyprotein-encoding orf, 27 in the 1b polyprotein-encoding orf, 5 in the spike-encoding orf and 6 in the nucleocapsid-encoding orf. They also provide insights about the validity of the procedure, confirming some motifs interspersed or attached to known relevant functional fragments of the genome, although most of them have not yet been associated to any known function. The high level of preservation of these motifs in most coronavirus genera suggest they might have potential to be used for diagnostic, in vaccines, or as substrate for protease inhibitors.

Introduction

Coronaviridae family contains enveloped, positive-sense, single-stranded RNA viruses. The coronavirus (CoV) taxonomy, by the Coronavirus Study Group of the International Committee for Taxonomy of Viruses (ICTV), classify CoVs into four genera: alpha, beta, gamma and delta. Moreover, bat CoVs are likely the gene source of alpha and beta genera, and avian CoVs are likely the gene source of gamma and delta genera.

In recent years there have been three zoonotic outbreaks of beta-CoVs: the severe acute respiratory syndrome 1 (SARS-CoV-1); the Middle East respiratory syndrome (MERS-CoV) and the severe acute respiratory syndrome 2 (SARS-CoV-2). The spread of the pandemic caused by the beta-SARS-CoV-2 is affecting all around the world. After beta-SARS-CoV-2 genome isolate, a study showed that genomes from both bats and pangolin species would be natural reservoirs of the SARS-CoV-2, highlighting the high identity between genomes of these species and the human SARS-CoV-2. The high genome identity observed when different beta-SARS-CoV genomes are compared, make it difficult to outline a strategy for distinguish highly conserved short RNA motifs with possible biological interest. Consequently, an approximation to the issue independently on the percent identity between genomes would be desirable.

Categorizing NT triplets by their gross-composition, resulted to be a useful strategy when applied to issues regarding coevolution or gene clustering; an analysis between distant strains of picornavirus genomes, as of the foot-and-mouth disease virus and of the human rhinovirus, gave as result, some NT sequence correlations not observed heretofore. Briefly, the strategy proposed works as follows: i) the genome is read fully-overlapping, taking into account the context of each NT of the genome. ii) each NT triplet is categorized by their gross composition; such categorizations, called triplet composons (or tCPs), generate a tCP sequence that can be compared with other tCP sequences. Comparing tCP sequences let the access to short RNA motifs tCP conserved; Short motifs are hard to be found mainly when the percent identity among genomes is very high. Short RNA motifs, here in collected, are clearly distinguished from the known short linear motifs or SLiMs that are peptides mediating protein-protein interactions.
having a functional interaction interface encoded in a short and poorly conserved sequence\textsuperscript{15}. The goal of the present study was to provide proof-of-concept for a rational approach to obtain short highly conserved RNA motifs potentially relevant.

**Methods**

**Data availability.** The CoV genomes were obtained from the GenBank\textsuperscript{\textregistered}, https://www.ncbi.nlm.nih.gov/\textsuperscript{16}. Genomes set to be studied would be: human coronavirus (HCoV)\textsuperscript{17}, porcine epidemic diarrhoea virus (PEDV)\textsuperscript{18}, human severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)\textsuperscript{19}, human severe acute respiratory syndrome coronavirus (SARS-CoV-1)\textsuperscript{20}, civet severe acute respiratory syndrome coronavirus (civet SARS-CoV)\textsuperscript{21}, mouse hepatitis virus (MHV)\textsuperscript{22}, bat severe acute respiratory syndrome coronavirus (bat SARS-CoV)\textsuperscript{23}, avian infectious bronchitis virus (IBV)\textsuperscript{24} and Common moorhen coronavirus\textsuperscript{3}. CoVs data as the strain, the genus, the accession number and references are in Table 1. Abbreviations are taken from the International Committee on Taxonomy of Viruses (ICTV) https://talk.ictvonline.org/. In the paper, T is listed as a nucleotide. Since the viral genomes are RNA, U should be substituted for T. However, very often, RNA is reverse-transcribed into DNA first, and then the DNA is sequenced. This is the reason why GenBank present the single-stranded RNA viral genomes in this way. Moreover, NCBI support indicates that replacing U with T is a GenBank convention, which saves computational resources.

| Virus                        | Strain       | Genus     | Accession   | Length (bp) | references |
|------------------------------|--------------|-----------|-------------|-------------|------------|
| Human HCoV                   | 229E         | Alpha-CoV | AF304460    | 27317       | 17         |
| Porcine PEDV                 | CV777        | Alpha-CoV | AF353511    | 28033       | 18         |
| Human SARS-CoV-2             | WHU01        | Beta-CoV  | MN988668    | 29881       | 19         |
| Human SARS-CoV-1             | Urbani       | Beta-CoV  | AY278741    | 29727       | 20         |
| Civet SARS-CoV               | SZ3          | Beta-CoV  | AY304486    | 29741       | 21         |
| Murine MHV                   | A59          | Beta-CoV  | AY700211    | 31335       | 22         |
| Bat SARS-CoV                 | 279/2005     | Beta-CoV  | DQ648857    | 29741       | 23         |
| Avian IBV                    | Mass 41      | Gamma-CoV | AY851295    | 27475       | 24         |
| Common-moorhen CoV           | HKU21-8295   | Delta-CoV | JQ065049    | 26223       | 3          |

**Table 1.** Genomes of different CoV genera

**Numerical analysis.** Similarities and dissimilarities among genomes of different CoV genera were analysed by the tCP-method\textsuperscript{12}. The justification of the method is based in exclusionary multiplet categorizations characterized by the presence or absence of particular bases. 14 different categorizations, or tCPs, were identified (Table 2), each one containing one or six NT triplets depending on whether they are non-degenerated or degenerated\textsuperscript{11}. The RNA was read fully-overlapping avoiding information loss and guaranteeing all triplets are considered in the study, taking into account, thus, the context of each NT in the RNA sequence.
The tCP distribution in CoV genomes is represented in cumulative tCP usage frequency graphs. The cumulative frequency is the sum of all previous tCP appearances along the sequence up to the current length. To simplify, the cumulative graph and its regression line, hereafter called ‘tCP profile’, were projected on the axis representing the genome length by subtracting the cumulative tCP usage frequency from the regression line. The profile represents the distribution of differences between the tCP events observed, \( tCP_0 \), and the tCP events estimated, \( tCP_e \). It will consider that two RNA sequences share a similar tCP profile when the Pearson correlation coefficient, \( r \), is equal or higher than an arbitrarily cut-off. The election of the cut-off would be done on the basis of the p-value (significance level). It will consider for comparison, cut-offs for \( 0.7 \leq r < 0.8 \) and \( r \geq 0.8 \). The Pearson correlation coefficient obtained are highly significant for both cut-offs due to the dependence of \( r \) on the sample size; in the analysis, the sample size is \( >2\times10^4 \) points. For both cut-offs the \( r \) significance was \( p<0.001 \). Graphs and statistics were carried out with the package OriginPro 8 SRO V8.0724 (B724) ©OriginLab Corporation.

For assessing the applicability of the method, certain criteria must be fulfilled. i) The tCP identity near 100% was discarded when compared CoV genomes, because impedes looking for SLiMs. Thus, the number of shared tCPs must be significantly lower than 14. ii) Viral genomes must be orthologs, because correlation does not imply causation unless we are dealing with highly correlated orthologs. iii) Last but not least, we must search for distant ortholog genomes to look for RNA SLiMs conserved in most CoV genera. This criterion will be met when comparing beta-SARS-CoVs with genomes of the remaining CoV genera.

**Identification of tCPs along CoV genomes.** Identification of tCPs along genomes was done as follows: i) we translated CoV genomes to tCP sequences with the help of Table 2; ii) we compared tCP sequences using the dynamic algorithm created for the global alignment of two sequences; iii) to compare the tCP profiles of different CoVs we computing the correlation coefficient between them; iv) if the correlation coefficient of a pair of given tCP profiles is higher than the cut-off considered, the corresponding shared tCPs were annotated.

### Table 2. List of tCPs and their associated NTs

| COMPOSON | RNA-TRIPLETS ASSOCIATED TO COMPOSONS |
|----------|-------------------------------------|
| <A>      | AAA                                 |
| <T>      | TTT                                 |
| <G>      | GGG                                 |
| <C>      | CCC                                 |
| <AC>     | AAC, CAA, ACA, CCA, ACC, CAC        |
| <AT>     | AAT, TAA, ATA, TTA, ATT, TAT        |
| <AG>     | AAG, AAG, AGA, GGA, AGG, GAG        |
| <CG>     | CCG, GCC, CGG, GGC, CGG, GCG       |
| <GT>     | GGT, TGG, GTG, TTG, GGT, GTG        |
| <CT>     | CCT, TCC, CTC, TTT, CTT, TTT        |
| <AGC>    | AGC, GCA, GAC, AGG, CGA, GAC        |
| <AGT>    | AGT, GAT, TAG, ATG, TGA, GAT        |
| <ACT>    | ACT, CTA, TAC, ATC, TCA, CAT        |
| <TCG>    | TCG, CGT, GTC, TGC, GCT, CTG        |
Results

Alignment of full-length CoV genomes. We started comparing genomes of different CoV genera to verify whether they fulfil the requirements to apply the tCP-method (Table 3). The beta-SARS-CoV-2 genome share about 79% NT identity with others beta-SARS-CoV genomes, and <51% with others CoV genera, suggesting that beta-SARS-CoVs may have diverged from other CoV genera. When alpha, gamma, and delta-CoVs were compared it was observed <60% NT identity. The genomes of Civet SARS-CoV and human SARS-CoV-1, however, share almost 100% NT and tCP identities supporting the zoonotic nature of human SARS-CoV-1; they differ in a 29-NT sequence retained by the civet isolate not found in the human isolate\textsuperscript{21}, sharing, therefore, all the 14 tCPs. In the other cases, the tCP identity is significantly lower than the NT identity.

To sum up, the results presented revealed that selected CoV genomes fulfil the requirements described in Methods and look like adequate for the application of the tCP method to comparisons with the SARS-CoV-2 genome.

Comparison of tCP profiles among human beta-SARS-CoVs-1 and 2. Beta-SARS-CoV-2 and beta-SARS-CoV-1 genomes were compared to illustrate what and how many tCPs are shared by both. The results showed (Table 4) that six tCPs, <AG>, <CG>, <GT>, <AGT>, <C> and <T>, were shared by both, for r≥0.8 (p<0.001); high resemblances were observed for these profiles (Figure 1). However, significant dissimilarities were observed among both genomes when the correlation coefficient decreases (Figure 1, profiles <AT> and <G>). The complete panel representing all the 14 tCP profiles from the comparison (Supporting Figure 1) shows, for r<0.7, notable dissimilarities in the remaining profiles. Often, short highly similar fragments are observed interspersed in low similarity regions. Such short fragments would be of interest in local genome studies but they were not considered in this paper.

| CoVs                  | HSARS CoV-2 | HSARS CoV-1 | Civet SARS CoV | Bat SARS CoV | HCoV | PEDV | MHV | IBV | Common moorhen CoV |
|-----------------------|-------------|-------------|----------------|--------------|------|------|-----|-----|-------------------|
| HSARS CoV-2           | 100         | 79.2        | 79.2           | 78.9         | 49.5 | 48.8 | 50.3| 49.7| 47.8              |
| HSARS CoV-1           | 0           | 4.3         | 4.4            | 4.4          | 36.0 | 36.6 | 34.1| 35.2| 38.5              |
| Civet SARS CoV        | 87.7        | 99.7        | 99.7           | 87.8         | 49.4 | 48.7 | 49.4| 49.7| 47.0              |
| Bat SARS CoV          | 8.7         | 0           | 0.1            | 2.0          | 35.5 | 36.6 | 35.5| 35.6| 39.0              |
| HCoV                  | 59.9        | 73.0        | 73.0           | 100          | 49.3 | 48.8 | 49.3| 49.6| 47.1              |
| PEDV                  | 9.1         | 3.7         | 3.7            | 0            | 35.5 | 36.6 | 35.6| 35.6| 39.1              |
| MHV                   | 30.0        | 29.2        | 29.2           | 29.3         | 100  | 59.3 | 48.0| 50.1| 49.5              |
| IBV                   | 37.3        | 37.7        | 37.9           | 38.6         | 22  | 22  | 37.7| 34.9| 36.3              |
| Common moorhen CoV    | 38.3        | 38.7        | 38.5           | 38.6         | 28.2 | 0   | 37.9| 35.9| 36.9              |

Upper number in the box: bold, NT percent identity; *italics, tCP percent identity. Lower number in the box: percentage of GAPS

Table 3. NT and tCP percent identity between alignments of CoV genomes

5
Figure 1. tCP profiles when genomes of human beta-SARS-CoV-1 (red) and human beta-SARS-CoV-2 (blue) were compared. For r≥0.8, tCP profiles <AG>, <CG>, <GT>, <AGT>, <C> and <T> were represented. tCP profiles <AT> and <G> were represented as an example of non-shared tCPs, for r<0.7. For each tCP, Pearson correlation coefficients, r, were annotated.

The shared tCPs for cut-offs r≥0.8 (black) and 0.7≤r<0.8 (red) were illustrated in Table 4. Most beta-SARS-CoVs would show a high percent identity ranging in between 78.9% to 99.7% for NTs, and 57.6% to 99.3% for tCPs (Table 3); all comparisons show there are many conserved tCPs (6 to 12). All beta-SARS-CoVs, for r≥0.8, share the tCPs <T>, <AG>, <GT> and <AGT> and for 0.7≤r<0.8 share the tCP <AGC>.

Figure 2. Similarities and dissimilarities in the profiles of <AGT> and <G>, when the beta-SARS-CoV-2 genome (red) was compared with genomes from A) alpha-HCoV, B) alpha-PED, C) delta-common moorhen CoV and D) gamma-IBV (blue). tCPs were represented in bold together with the value of the correlation coefficient, r.


| tCP alignments | SARS CoV-2 | SARS CoV-1 | Civet SARS CoV | Bat CoV | HCoV | PEDV | MHV | IBV | Common moorhen CoV |
|----------------|------------|------------|----------------|---------|------|------|-----|-----|--------------------|
| SARS CoV-2     | all the 14 tCPs | <C>, <T>, <AG>, <CG>, <GT>, <AGT>, <AC>, <CT>, <ACT>, <AGC> | <C>, <T>, <AG>, <CG>, <GT>, <AGT>, <AC>, <CT>, <ACT>, <AGC> | <G>, <GT>, <AG>, <CG>, <AT>, <AGC>, <CT>, <ACT> | <C>, <AGT>, <ACG>, <CT>, <ACT>, <AGC> | <C>, <AGT>, <ACG>, <CT>, <ACT>, <AGC> |
| SARS CoV-1     | all the 14 tCPs | all the 14 tCPs | <C>, <T>, <AG>, <CG>, <GT>, <AGT>, <AC>, <CT>, <ACT>, <AGC> | <C>, <T>, <AG>, <CG>, <GT>, <AGT>, <AC>, <CT>, <ACT>, <AGC> | <C>, <AGT>, <ACG>, <CT>, <ACT>, <AGC> | <C>, <AGT>, <ACG>, <CT>, <ACT>, <AGC> |
| Civet SARS CoV | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | <C>, <AG>, <GT>, <CT>, <ACT>, <AGC>, <ACT> | <AC>, <AGC>, <CT>, <ACT> |
| Bat CoV        | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | <AG>, <GT>, <AGC>, <ACT> | <AC>, <CT>, <AG>, <AGC> |
| HCoV           | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | <AC>, <CT> | <AC>, <CT>, <AG>, <AGC> |
| PEDV           | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | <AC>, <AT>, <AG> | <C>, <AGT> |
| MHV            | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | <AGT>, <GT>, <AT>, <CT> |
| IBV            | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | all the 14 tCPs | <AGT>, <GT>, <AT>, <CT> |

Black text, tCPs conserved for r≥0.8; red text, tCPs conserved for 0.7≤r<0.8 (p<0.001; points per profile n>2x10⁴). As expected, all the 14 tCPs were conserved in identical genomes.
Comparison of tCP profiles among beta-SARS-CoV-2 and other CoV genera. NT and tCP percent identities falls up to 33% and 46%, respectively, when the beta-SARS-CoV-2 genome was compared with alpha-HCoV, alpha-PEDV, gamma-IBV and delta-common-moorhen genomes (Table 3); results indicate that <AGT> is the only tCP conserved, except for the alpha HCoV genome that conserves the tCP <G> (Table 4). But, what happen with the tCP <G> in HCoV?

![Figure 3](image)

**Figure 3.** Similarities and dissimilarities observed in the <AGC> profile when the beta-SARS-CoV-2 genome (red) was compared A) with genomes from beta-SARS-CoVs as Human SARS-CoV-1, Civet SARS-CoV and bat SARS-CoV; B) with genomes from other CoV genera as alpha-HCoV, alpha-PEDV, beta-MHV, gamma-IBV and delta-common-moorhen-CoV. The Pearson correlation coefficient, r, was also represented.

To answer the question, we further investigated the resemblance of both <AGT> and <G> profiles when the genome of SARS-CoV-2 was compared with other alpha-CoVs, as well as, with genomes from other CoV genera. Figure 2 shows profiles of both tCPs obtained from alignments of the SARS-CoV-2 genome with genomes of HCoV, PEDV, IBV and common moorhen-CoV genera. The figure illustrates the changes observed in both tCPs along genomes. <AGT> shows an increase in the correlation coefficient from r=0.63 (for HCoV) to r=0.92 (for IBV) in different CoV genera. All <AGT> profiles maintain a notable resemblance with the exception of the HCoV 5’region that shows important dissimilarities, explaining the decrease in the correlation coefficient observed in HCoV relative to other CoVs. However, a high identity was observed in the HCoV 3’region of the genome, suggesting that dissimilarity observed in the 5’ region of the genome was species specific. <AGT> is shared also by SARS-CoVs suggesting that <AGT> would be evolutionary important in all CoV genera. Data further supporting the notion that even for the case of HCoV the resemblance of tCP profiles is notable despite the local dissimilarity in the 5’ region.

Changes most pronounced were observed, however, for <G>. The correlation coefficient relative to SARS-CoV-2 for <G> profile changes notably (Figure 2), increasing progressively from avian (r=-0.24) to mammals (r=0.8). The profile resemblance observed for <AGT> is, however, lost for <G>, indicating a notable change in the <G> usage between avian and mammals. As it is shown in Table 3, the tCP <AGC> would also be conserved in the most CoV genomes
analysed for a cut-off $r \geq 0.7$ (p<0.001). The high significance of the correlation coefficient ensures that at least 49% of the variance will be shared when CoV genomes are compared (determination coefficient, $r^2 = 0.49$). Therefore, the resemblances of <AGC> profile would be also guaranteed providing useful evolutionary information.

Figure 3 shows a detailed analysis of <AGC> distribution when the beta-SARS-CoV-2 genome was compared with others beta-SARS-CoVs genomes (Figure 3A) and also with genomes from other CoV genera (Figure 3B). The <AGC> profile shows, in all CoV genera, a correlation coefficient $r > 0.7$ (p<0.001) with the exceptions of both MHV ($r = 0.33$) and common moorhen CoV ($r = 0.47$), although, in both genomes, resemblances in the 5’ region were notable (Figure 3B); All profiles maintain, indeed, a substantial resemblance, as expected, including MHV and common moorhen CoV, but in these cases with notable dissimilarities 3’ regions of tCP profiles. There are not significant differences in correlation coefficients ($r < 0.7$) for <AGC> profile except for bat SARS-CoV ($r = 0.88$), indicating evolutionary closeness between human SARS-CoV-2 and bat SARS-CoV. The tCP <AGC> would be, then, considered as significantly conserved in most CoV genera. Not only it is shared in all SARS-CoVs but also in the most CoV genera.

To sum up, the results presented revealed that two tCPs <AGT> and <ACG> were conserved in nearly all CoV genera. Despite dissimilarities in the HCoV 3’ region of the genome and the high similarity in the 5’ region for <AGT>, this is in keeping with the idea of a common origin for all CoV genera. We suggest, with a high confidence level, that RNA short motifs generated by tCPs <AGT> and <AGC> would be shared by most CoVs, independently on the NT similarity between their genomes.

List of short RNA motifs tCP conserved in CoV genomes. Having established, with high confidence level, that tCPs <AGT> and <AGC> were conserved by most CoV genera, and that <T>, <AG>, <GT>, <AGT> and <AGC> were conserved in beta-SARS-CoVs we are now interested in identifying short RNA motifs tCP conserved by most CoV genera.

First of all, we will create a target selection strategy to isolate short genomic motifs following the next criteria: i) RNA motifs must be tCP conserved in all beta-SARS-CoVs; the criterion imply that selected motifs will share NT-triplets from <T>, <AG>, <GT>, <AGT> and <AGC>. ii) Selected motifs will contain more than 50% of NT-triplets from the conserved tCPs <AGT> and <AGC>. iii) The RNA fragments chosen would be as long as possible; limited exceptions to this rule would be fragments of shorter length having very high percentages of NTs contained in <AGT> and <AGC>. Following these criteria, Table 5 shows 88 selected RNA motifs, as well as, their associated amino acids. The size, position and percentage of conserved NTs in tCPs <AGT> and <AGC> were also annotated. The SARS-CoV-2 harbours a wide-range of selected RNA motifs ranging in size from 12 to 49 bps (from 4 to 16 amino acids).

From now on, in the main text, we will referred to selected RNA motifs using the notation from the end of Table 5: upper case letters would represent NTs-tCP conserved in all beta-SARS-CoVs; underlined-bold upper case letters represent NTs-tCP conserved in most CoV genera; codons would be represented by separated NT triplets in the correct reading frame and amino acids tCP conserved in most CoV genera would be represented by bold upper case letters below their corresponding codons.
| Conserved RNA motifs** | N° | Protein encoding ORF | N° amino acids | Genome location | % |
|------------------------|----|----------------------|----------------|----------------|---|
| GGTGTGACCGGAAGAGTAAGATG6 GAG AGC CTT GT | 1 | pp1a | 4 | 246-278 | 51 |
| 5’ NC region M E S L | | | | | |
| ATG GTT AGG GTT GAG CTT G A G A A | 2 | pp1a | 8 | 512-537 | 69 |
| V M V E L V A E | | | | | |
| TAGT TAC GCC GCC GAT | 3 | pp1a | 5 | 666-681 | 69 |
| S Y G A D | | | | | |
| AAG CTT GAT GGC TTT AGT GGT AGA A | 4 | pp1a | 8 | 1137-1162 | 60 |
| K L D G F M G R | | | | | |
| TAGT TCG GAT GAA GAG AAG T | 5 | pp1a | 7 | 2349-2167 | 53 |
| D W L E E K | | | | | |
| AA GCC TTT AGT AAT TGA GGT GAA | 6 | pp1a | 6 | 2366-2385 | 60 |
| A L N L G E | | | | | |
| AAG GGA TTT AGC TAC AGA AAG TGT G T T | 7 | pp1a | 8 | 2404-2427 | 50 |
| K G L Y R K C V | | | | | |
| GCA TTG TTA GAT GAG TGG AGT ATG GC | 8 | pp1a | 8 | 2945-2972 | 79 |
| I D L D E W S M | | | | | |
| CAG GAT GAG TTA TTA GAT GAT GAT AGT | 9 | pp1a | 15 | 3044-3092 | 55 |
| D E D E E G D C E E | | | | | |
| GAG TTT GAG CC | | | | | |
| E E E | | | | | |
| AT GAG TAT GGA A | 10 | pp1a | 3 | 3104-3115 | 100 |
| E V G | | | | | |
| T GAA GAG GCA GAA GAA GAT TGG TTA GAT GAT GAT AGT | 11 | pp1a | 13 | 3177-3216 | 60 |
| E E E Q E E D W L D D S | | | | | |
| TT GAA GGT AAT AGT TTT AGT GAT T | 12 | pp1a | 7 | 3332-3355 | 67 |
| E V N S F S G | | | | | |
| AGC TTT TGG AAT AGG AGT G | 13 | pp1a | 7 | 3828-3850 | 51 |
| S F L E M K S | | | | | |
| TA GCG AAA GTG TT G A A A A A A G T G C C | 14 | pp1a | 7 | 4202-4226 | 60 |
| A K A L R K V | | | | | |
| TAG TGA GAG GCA AAG AGA AG T | 15 | pp1a | 7 | 4278-4299 | 59 |
| E V E A K T V | | | | | |
| TAG GAT GAA GTT | 16 | pp1a | 4 | 4901-4914 | 86 |
| D G E V | | | | | |
| T T G A G A A A G T G A A G G A C | 17 | pp1a | 5 | 4947-4964 | 61 |
| L R E V R | | | | | |
| A C G T G T T G A G G T T T G A G G A | 18 | pp1a | 6 | 5145-5166 | 57 |
| T R E V E A F E | | | | | |
| AAG ACA GT A G A G G T A G T A G T A G A G A A | 19 | pp1a | 11 | 5422-5454 | 73 |
| K T V G E L G D V R E | | | | | |
| AAG GGT G1 A G A A G T G T T A G T C A C T A C T A G T G C A | 20 | pp1a | 10 | 5551-5581 | 58 |
| K G V E A V M Y M G | | | | | |
| TG AAG TCA GAG GGC CGA GGG AGT GAT | 21 | pp1a | 9 | 6347-6375 | 79 |
| K S E D A Q G M D | | | | | |
| T T G A G T G A T T G G A A | 22 | pp1a | 4 | 6449-6464 | 88 |
| E C N V | | | | | |
| AGT GGT TTA GAT TCT TTA GAC | 23 | pp1a | 7 | 7141-7159 | 57 |
| S G L D S L D | | | | | |
| TA G T C A G A G T G T G G T T G C A | 24 | pp1a | 7 | 7232-7254 | 61 |
| V A E W F L A | | | | | |
| ATG GTT AGA ATG TA | 25 | pp1a | 4 | 7402-7416 | 79 |
| M V R M | | | | | |
| AT GTA TGG AAA AGT TAT GTG CAT GTT GTA GAC GGT TGT A | 26 | pp1a | 12 | 7439-7477 | 72 |
| V W K S Y V V V D G C | | | | | |
| TAGT GAT GAA GTT GG C G A G A G A C | 27 | pp1a | 8 | 7659-7684 | 69 |
| T S D E V A R D L | | | | | |
| TAGT GAT GGT GAT AGT GCG GAA GTT GCA GGT | 28 | pp1a | 10 | 7998-8028 | 65 |
| D V G D S A E V A V | | | | | |
| TAGT ATG GAC TGG ATG GCG C | 29 | pp1a | 6 | 8334-8353 | 60 |
| C I D C S A | | | | | |
| ATG TTT AGC CAG GGT GGT GGT AGT T | 30 | pp1a | 8 | 8772-8797 | 58 |
| W F S Q R G G S | | | | | |
| ATG GAT GAC TTT TTA CAT TT | 31 | pp1a | 5 | 8897-8915 | 53 |
| G D F L H | | | | | |
| TAGT GCA GGT GGT A | 32 | pp1a | 4 | 8931-8944 | 86 |
Table 5. List of short RNA motifs and their associated peptides

Distribution of NTs-tCP conserved in short selected RNA motifs along the SARS-CoV-2 genome.

Differences in the frequency of appearance of NTs-tCP conserved in short selected RNA motifs were observed in Table 5. In order to gain a better sense of their distribution along the genome, Figure 4A shows a simple schematic of the genome organization of the SARS-CoV-2 \(^{28}\), together with the distribution of NTs-tCP conserved at the same scale (Figure 4B) in selected RNA motifs. When compared Figure 4A and 4B, it was observed the bulk of NTs-tCP conserved, located in the replicate-encoding orf 1ab, being the region with most NTs-tCP conserved and consequently the most populated with selected RNA motifs. The orf 1ab encode two overlapping polyproteins (abbreviated pp1a and pp1b) that regulate replication and transcription processes\(^{17}\) being then an optimum target for a possible therapy. The figure shows a comparative between NTs-tCP conserved when compared the SARS-CoV-2 and other SARS-CoVs (black), and the SARS-CoV-2 and other CoV genera (grey); differences observed in the frequency of appearance between both
cases are a consequence of differences in the number of tCPs conserved between the first and second cases (Table 4). The larger size motifs appear in the pp1a region following by the pp1b region. Four genome stretches of at least 500 bps do not contain any selected tCP motif in the pp1ab-encoding orf. One of them is located at 5’ region and other in the middle of pp1a-encoding orf; there are two other in the middle of the pp1b-encoding orf. This means that despite the high similarity shown by this genome region, lacks of motifs tCP conserved in SARS-CoVs and also in other CoV genera. The spike-pp encoding-orf also contains broad regions non tCP conserved specially in the pp1b/spike interface. Many other genome regions lacking conserved tCP motifs as the protein-encoding orfs of the surface, the envelope and also the matrix (Figure 4B).

![Genomic organization of SARS-CoV-2](image1)

**Figure 4.** A) Genomic organization of SARS-CoV-2. B) Distribution of NTs-tCP conserved in human SARS-CoV-2 relative to other SARS-CoVs (black) and relative to most CoV genera (grey). The nomenclature as in Table 5.

To sum up, the results presented revealed that encoding proteins of SARS-CoV-2 as pp1a, pp1b, the spike polyprotein (Spp) and the nucleocapsid protein, contain tCP conserved fragments separated by genome fragments lacking tCP conservation. A high number of tCP conserved motifs were observed in the pp1ab-encoding orf. In contrast, fragments and genome regions lacking tCP conserved motifs were observed specially near the 5’ regions of the genome.

**Discussion**

We will discuss whether or not some selected RNA motifs from Table 5 would have biological relevance. As it is not possible the analyses of all motifs from the table, we centre the attention on motifs 1, 9, 24, 39, 68, 79, 81 and 82, appearing in the literature inserted or attached to known relevant fragments of CoV genomes. It must be taken into account that conserved amino acids encoded by motifs of different CoV genera are there, attending possibly, to reasons of fitness or survival and could, therefore, be endowed with biological relevance.

The first point to be underlined, has to do with the fact that amino acids encoded by tCP-conserved codons would be tCP-conserved too; as a result (see Table 2), asparagine (D)-encoding
codons GAT and GAC, serine (S)-encoding codons AGT and AGC, valine (V)-encoding codon GTA, threonine (T)-encoding codon ACG, arginine (R)-encoding codon CGA, glutamine (Q)-encoding codon CAG, and alanine (A)-encoding codon GCA would be also conserved, as well as the initiation (ATG) and termination codons (TGA and TAG). Motif 1 would be an example of a highly conserved region surrounding the initiation of pp1ab.

There are two identical RNA motifs, (TGTAGCTAGT), encoding the motif V-A-S (valine-alanine-serine). The motif appears in two regions of the SARS-CoV-2 genome, inserted in tCP conserved motifs 68 and 81 (double-underlined), namely, (AT GTA GCT AGT GCA) and (CGG CGG GCA CGT AGT GTA GCT AGT), respectively; one is in the pp1ab-encoding orf and the other in the Spp-encoding orf. The V-A-S motif present in SARS-CoV-2 Spp, is absent, however, in other beta-SARS-CoV Spp’s, namely, human SARS-CoV-1 and civet and bat SARS-CoVs; this is a notable difference between SARS-CoV-1 and SARS-CoV-2 genomes, suggesting that the motif found in the Spp would be species specific. The motif appears profusely in beta-SARS-CoVs pp1ab and also in alpha-HCoV pp1ab. Its presence is, however, very limited in other regions of CoV genomes (Supporting Table 1).

The motif V-A-S is attached to the R-R-A-R fragment of motif 81 (R-R-A-R-S-V-A-S) through a highly conserved serine, a cleavage site for the host protease furin necessary for activation of Spp of SARS-CoV-1 and MERS-CoV as proposed by a two-step sequential protease cleavage model29,30; this supports the importance of the motif in SARS-CoVs Spp because pp1a and pp1b help to release functional polypeptides by the papain-like protease (PLpro) and the 3C-like protease (3CLpro) situated in the non-structural protein (NSp) region, that could serve as substrate for some type of inhibitors. The short motif V-A-S appears often associated to SARS-CoV-1 3CLpro cleavage sites, into canonical recognition positions; more specifically, at the same positions of NSp4/5 and NSp8/931. Moreover, sequences containing V-A-S could be considered as potential substrates for some protease inhibitors in SARS-CoVs. The sequence of SARS-CoV-1 3CLpro L-V-A-S-L, containing the V-A-S-motif, serve as substrate for classic protease inhibitors as 5-(2′-aminoethyl) aminonaphthalene sulfonic acid (EDANS)32,33, amino benzoyl (abz)34,35, or the colorimetric label pNA34. It is the first time, as far as we know, that this sort motif has been suggested as a key fragment in certain experimentally deduced motifs having biological significance. This is supported by data describing that amino acid sequence alignment of SARS-CoV-2/SARS-CoV-1 replicase polyproteins show a high overall identity, resulting in a noticeable conservation at the polyprotein cleavage sites36. The motif appears up to 10 times in the SARS-CoV-2 genome, 9 times in the pp1ab-encoding orf and one more time in the Spp-encoding orf at positions 18841-18860 and 23605-23628, respectively (Table 5). A detailed analysis referred to the V-A-S-motif shows why information supplied by the tCP code-like is complementary to that obtained from the genetic code. As can be observed, the fact that the V-A-S-motifs are identical in SARS-CoVs does not imply necessarily they must be highly tCP-conserved in all CoV genera. In fact, there are notable differences in their tCP conservation, indicating, then, the evolutionary relevance of each one. The V-A-S-motif can be considered an example of how viruses incorporate simple motifs into their protein sequences to mimic human proteins and enhance their functional capabilities in host cells during infections. In addition, we suspect the V-A-S-motif could have a biological role in CoV genomes due to the high number of copies of the motif present in the pp1ab relative to the rest of viral proteins where few if any exist. Despite the size of the motif its presence in the pp1ab is much higher than expected relative to any other genome region (Supporting Table 1).

Other motif with possible biological relevance in beta-SARS-CoVs Spp could be the motif-82 encoding G-K-Y-E-Q, (TT GGA AGT TAT GAG CAG TA). It is located at the end of the Spp
and it is attached to the Spp fragment Y-I-K-W-P-W-Y-I-W-L, a highly conserved tryptophan W-rich membrane proximal external region (MPER) present on members of Coronaviridae family and more specifically in beta-SARS-CoV-2\textsuperscript{37-42}, forming the polypeptide G-K-Y-E-Q-Y-I-K-W-P-W-Y-I-W-L. Mutational studies have shown that W residues in the MPER are essential for effective viral infection and that MPER could serve as inhibitor for the entry process\textsuperscript{43}; the functional reason for the presence of the tCP conserved motif G-K-Y-E-Q attached to MPER would be related to the highly hydrophobic nature of the MPER peptide requiring the inclusion of the upstream charged residues K-Y-E-Q to increase its aqueous solubility\textsuperscript{44}.

Table 5 contains other repeated and conserved sequences in the SARS-CoV-2 genome, namely, (TTGCAGAGTGGTTTT) and (GAGTATGG), both located in the pp1a-encoding orf: The first one (double underlined), is inserted on selected RNA-motifs 24 and 39, namely, (\textbf{TA GTT GCA GAG TGG TTT TTG GCA}) and (\textbf{CA GCT GTT TTG CAG AGT GGT TTT AGA A}), respectively (Table 5). The peptides encoded by both RNA motifs are, V-A-E-W-F-L-A and A-V-L-Q-S-G-F-R, respectively. The peptide V-A-E-W-F-L-A appear related with the identification of potential cytotoxic T lymphocyte epitopes in the SARS-CoV-2 Indian isolate which can act as an effective vaccine epitope candidate\textsuperscript{45}; the peptide A-V-L-Q-S-G-F-R, could serve as a starting point for determining effective drug candidates for the treatment of SARS-CoV; the octapeptide naturally becomes an ideal good inhibitor for the SARS proteinase\textsuperscript{46,47}. The second repeated motif (double underlined), is inserted on selected RNA-motifs 8 and 10, namely, (\textbf{GC ATT GAT TTA GAT GAG TGG AGT ATG GC}) and (\textbf{AT GAG TAT GGT A}), respectively (Table 5). The peptides encoded by both RNA motifs are, I-D-L-D-E-W-S-M and E-Y-G, respectively; both are close together in the genome of SARS-CoV-2 (between positions 2945 and 3115), at the initio of the pp1a; they show 79\% and 100\% NTs-tCP conserved in all CoV genera, respectively. We have not found verified information in the literature about both motifs. However, the high level of tCP conservation between CoV genera make them good candidates for study.

The peptide (\textbf{D-E-D-E-E-G-D-C-E-E-E-F-E}) encoded by RNA motif 9, (\textbf{CA GAT GAG GAT GAA GAA GAA GGT GAT TGT GAA GAA GAA GAG TTT GAG CC}) (Table 5) has also interest. Segments of the motif have identity with functional fragments of the high mobility group box 1 (HMGB1) protein, actively secreted by inflammatory cells in response to pathogen-associated molecular patterns\textsuperscript{48,49} and probably related with SARS-CoVs pathology\textsuperscript{39}. An anti-HMGB1 monoclonal antibody has been developed\textsuperscript{50} recognizing specifically the C-terminal tail of the motif 9, D-E-D-E-E-E. As can be observed, the motif contains a long acidic tail comprising, mostly glutamic (E) and aspartic (D) acids, similar to that described in HMGB1\textsuperscript{51}. A current hypothesis is that immunization with anti-HMGB1 antibodies would confer protection against SARS-CoV-2 injuries\textsuperscript{52} by inhibiting the protease. Proteolytic processing of the CoV replicase is essential for ongoing viral RNA synthesis\textsuperscript{53}. The enzymatic activity of both PL\textsuperscript{50} and 3CL\textsuperscript{pro} are essential for the viral life cycle. That is why, the beta-SARS-CoVs proteases are attractive targets for the development of antiviral drugs to reduce viral replication and pathogenicity\textsuperscript{54}.

The Spp is utilized by CoVs to bind to their cellular receptors. The crystal structure of the beta-SARS-CoV-2 S receptor-binding domain (RBD) interacting with the ACE2 receptor has been determined\textsuperscript{55}. The binding activates the fusion between the cell and SARS-CoV membranes for the virus entry into the cell\textsuperscript{55,56}. Into the SARS-CoV-2 RBD there is a receptor binding motif (RBM) containing most of the contacting residues for ACE2 binding\textsuperscript{55}. Inserted in the RBM, is the RNA motif 80, (\textbf{TG TAT AGA TTT AGG AAG T})-encoding the peptide Y-R-L-F-R-K (Table 5), highly conserved in beta-CoVs although less conserved in other CoV genera. We think the motif conserve the residues Y and R because they could be determinant for the effectiveness of the Spp/ACE2 interaction. We based the suggestion on the high percent identity of tCP-conserved NTs from other CoV genera (67\%). Only five motifs, following tCP selection criteria,
have been selected in the Spp-encoding orf, and only one is located in the RBM. In the Spp RBM, the percent tCP-conserved residues between SARS-CoVs and other CoV genera is discrete, indicating that the protein motif would be specific of beta-SARS-CoVs; therefore, the relevance of the motif 80 is dictated by its location in the RBM and its implication in the Spp/ACE2 interaction. The Spp also contains the peptides R-R-A-R-S-V-A-S- and R-G-D-E-V-R encoded in RNA-motifs 81 and 79, namely, (CGG CGG GCA CGT AGT GTA GCT AGT) and (TAGA GGT GAT GAA GTC AGA C). Motif 79 is located in the RBD out of the RBM, and motif 81 is located out of the RBD, having 85% and 75% percent identity of tCP-conserved NTs, and also 67% and 75% tCP-conserved residues in other CoV genera different from beta-SARS-CoVs. Therefore, the V-A-S-fragment inserted in motif 81 appears again, now located in the SARS-CoV-2 Spp being relevant for two main reasons: first, it is known that effective conformational changes of S protein leading to membrane fusion, require both, the receptor binding and a proper protease activation. A R-R-A-R-furin site was found in the S protein motif R-R-A-R-S-V-A-S between S1 and S2 subunits in SARS-CoV-2 S protein inserted in RNA motif 81 containing also the highly conserved V-A-S-fragment; second, this RNA-motif is found in the beta-SARS-CoV-2 S protein is absent, however, in human, civet and bat SARS-CoV S proteins, highlighting, thus, a fundamental difference between those CoVs in a CoV region specially sensible for the survival. It is known that the SARS-CoV Spp can be activated by host cell proteases, with proteolytic cleavage at the S1/S2 boundary and adjacent to a fusion peptide in the S2 domain. Thus, although elastase-mediated activation of SARS-CoV Spp has been considered an important factor for the severe pneumonia seen in SARS-CoV-infected patients, studies with neutrophil elastase (NE) failed to give any consistent activation of SARS-CoV S-mediated fusion. The V-A-S-motif appears also in the Spp and it is known that neutrophil elastase, prefers substrates containing the V-A-S-motifs at the P1 position. This fact endowing the V-A-S-motif even more evolutionary-relevance. Other factors such as the unique R-R-A-R-furin cleavage site at the S1/S2 boundary of the SARS-CoV-2 Spp could play a role in facilitating the rapid human-to-human transmission.

The MPER has an important role in the function of the S protein. The MPER, rich in aromatic residues with three or four W and two or three Y residues are present in all coronavirus S proteins. Such sequence conservation is also found in other viruses with class I fusion protein, such as human immunodeficiency virus type 1 (HIV1), feline immunodeficiency virus, influenza virus, and Ebola virus. The selected RNA-motif 82 (Table 5) attached to MPER acquire biological relevance principally for the following reasons: i) When working with synthetic peptides derived from CoV S proteins, the highly hydrophobic nature of the MPER peptide required the inclusion of upstream charged residues K-Y-E-Q in the peptide to increase its aqueous solubility and facilitating purification; ii) In SARS-CoV-2 Spp the upstream charged residues K-Y-E-Q-motif 82 are naturally attached to MPER (this paper); iii) K-Y-E-Q-Y-I-K shares sequential homology with the cholesterol recognition amino acid consensus motif, that was believed to result in a peptide or protein that preferentially associates with cholesterol, a principal component of lipid rafts.

The matrix protein M is an integral membrane protein involved in budding which interacts with the nucleocapsid and S proteins. As a component of the viral envelope plays a role in morphogenesis and assembly by interaction with other viral proteins. M protein is connected with infectivity through binding to S protein and the surface receptor, to promoting membrane fusion. In this context, the absence of the tCP-conserved V-A-S-motif in the beta-SARS-CoV-2 M protein and its presence in the beta-SARS-CoV M protein is, between others, one of the most notable differences between both human beta-SARS-CoVs. Supporting Table 1 illustrates that, since the point of view of the M protein, the V-A-S-motif, as it occurs with human SARS-CoV, is also present in civet and bat beta SARS-CoVs and also in human and porcine alpha-CoVs and
will be absent in the rest of CoVs. The presence of the V-A-S-motif in CoV Spp imply its absence in the matrix protein and vice versa, with one exception, the porcine alpha PEDV that having the motif in both. Moreover, the motif is present in SARS-CoV-2, murine MHV, porcine PEDV and avian CoV M proteins but not in SARS-CoV, civet and bat M proteins. The above data suggest more evolutionary proximity of civets and bats with human SARS-CoV with human SARS-CoV-2 that could be more evolutionary close to murine MHV. The presence of the V-A-S-motif in beta-SARS-CoV M protein could play a role in M protein because neutrophil elastase, having activity against a broad range of extracellular matrix proteins, prefers substrates with Val>Ala>Ser, Cys at the P1 position.

The E (envelope) protein is the smallest of the structural proteins and it is abundantly expressed in the infected cell although, only a small portion is assimilated into the virion envelope. Most M proteins are restricted at the site of subcellular trafficking where it participates in CoV assembly and budding. In SARS-CoV-2 the tCP-method does not selected any RNA motif for this small protein as occurred for the M protein. The only specie having a V-A-S-motif in the E protein was the porcine alpha PEDV (Supporting Table 1).

The SARS-CoV-2 N (nucleocapsid) protein is a RNA-binding protein playing vital roles in i) forming helical ribonucleoproteins, ii) regulating viral RNA synthesis in replication/transcription, and iii) modulating infected cell metabolism, being a highly immunogenic and expressed protein during infection, inducing immune responses against human SARS-CoV and SARS-CoV-2. Most CoVs studied, with the exception of common moorhen delta-CoV (Supporting Table 1), lack the V-A-S-motif in the N protein, however, the tCP-method selected six RNA motifs conserved in beta-SARS-CoVs with percent of RNA conservation in other CoV genera ranging in between 52% and 72% (Table 4).

To sum up, the results presented revealed a new set of short RNA motifs in the SARS-CoV-2 by reading the genome fully-overlapping and categorizing RNA triplets by gross composition (Table 2). Some motifs show functional characteristics experimentally deduced and the remaining motifs must be taken into account in studies of functionality although today be unknown. For example, to our knowledge, the V-A-S-peptide, codified in motifs 68 and 81, has not been described before and however it is present in many protein fragments functionally relevant in CoVs. The same occur with other peptides as those codified in motifs 9, 68, 79, 80, 81, 82 (Table 5). In some cases, as the selected motif 80, it is inserted in a region functionally relevant as the RBM of the beta-SARS-CoV-2 S RBD to bind to the ACE2 receptor; in other cases, as the motif 82, it is attached to fragments functionally relevant, as the MPER, to increase its aqueous solubility and to facilitate its purification. It is unlikely to emerge those new motifs at random, inserted or flanking functional CoV genome fragments in general and in SARS-CoV-2 in particular. We suspect that selected RNA motifs could have significant roles in the fitness and survival of CoVs due their high degree of NTs-tCP conserved inter genera. They could be suitable for studies of inhibition and vaccines, more valuable because these motifs have not been observed heretofore.

References

1. de Groot, R.J. et al., in Coronaviridae, Elsevier Academic Press, Philadelphia, PA (2011).
2. Woo, P.C. et al. 2009. Comparative analysis of complete genome sequences of three avian coronaviruses reveals a novel group 3c coronavirus. J Virol 83, 908-917, https://pubmed.ncbi.nlm.nih.gov/18971277/ (2009).
3. Woo, P.C. Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. J Virol 86, 3995-4008, https://pubmed.ncbi.nlm.nih.gov/22278237/ (2012).
4. Dong, B.Q. et al. Detection of a novel and highly divergent coronavirus from asian leopard cats and Chinese ferret badgers in Southern China. J Virol 81, 6920-6926, https://pubmed.ncbi.nlm.nih.gov/17459938/ (2007).
5. Mihindukulasuriya, K.A., Wu, G., Leger, J., Nordhausen, R.W., Wang, D. Identification of a novel coronavirus from a beluga whale by using a panviral microarray. J Virol 82, 5084-5088, https://pubmed.ncbi.nlm.nih.gov/18353961/ (2008).
60. Fleming, J. O., Shubin, R. A., Sussman, M. A., Casteel, N., Stohlman, S. A. Monoclonal antibodies to the matrix (E1) glycoprotein of mouse hepatitis virus protect mice from encephalitis. *Virology* **168**, 162-167, https://pubmed.ncbi.nlm.nih.gov/2535900/ (1989).

61. Bode, W., Meyer, E., Jr., Powers, J. C. Human leukocyte and porcine pancreatic elastase: X-ray crystal structures, mechanism, substrate specificity, and mechanism-based inhibitors. *Biochemistry* **28**, 1951-1963, https://pubmed.ncbi.nlm.nih.gov/2655701/ (1989).

62. Venkatagopalan, P., Daskalova, S. M., Lopez, L. A., Dolezal, K. A., Hogue, B. G. Coronavirus envelope (E) protein remains at the site of assembly. *Virology* **478**, 75-85, https://pubmed.ncbi.nlm.nih.gov/25726972/ (2015).

63. Nieto-Torres, J. L. et al. Subcellular location and topology of severe acute respiratory syndrome coronavirus envelope protein. *Virology* **415**, 69-82, https://pubmed.ncbi.nlm.nih.gov/21524776/ (2011).

64. Cong, Y. et al. Nucleocapsid Protein Recruitment to Replication-Transcription Complexes Plays a Crucial Role in Coronaviral Life Cycle. *J Virol* **94**, e01925-19 https://pubmed.ncbi.nlm.nih.gov/31776274/ (2020).

65. Nelson, G. W., Stohlman, S. A., Tahara, S. M. High affinity interaction between nucleocapsid protein and leader/intergenic sequence of mouse hepatitis virus RNA. *J Gen Virol* **81**, 181-188, https://pubmed.ncbi.nlm.nih.gov/10640556/ (2000).

66. Stohlman, S. A. et al. Specific interaction between coronavirus leader RNA and nucleocapsid protein. *J Virol* **62**, 4288-4295, https://pubmed.ncbi.nlm.nih.gov/2845141/ (1988).

67. Ahmed, S. F., Quadeer, A. A., McKay, M. R. Preliminary Identification of Potential Vaccine Targets for the COVID-19 Coronavirus (SARS-CoV-2) Based on SARS-CoV Immunological Studies. *Viruses* **12**, 254, https://pubmed.ncbi.nlm.nih.gov/32106567/ (2020).

68. Liu, S. J. et al. Immunological characterizations of the nucleocapsid protein based SARS vaccine candidates. *Vaccine* **24**, 3100-3108, https://pubmed.ncbi.nlm.nih.gov/16494977/ (2006).

Acknowledgements

I would like to express my deep gratitude to Prof Carlos Alonso Bedate who died on April 13, 2020 in the middle of the SARS CoV-2 pandemic for his advice, guidance, friendship and encouragement throughout the last twenty-five years of my life.

This work was funded by a program of the Instituto de Salud Carlos III-Redes Temáticas de Investigación Cooperativa en Salud (ISCIII-RETIC RD06/0021/0008 program) and Laboratorios LETI. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. An institutional grant from Fundación Ramón Areces is also acknowledged.

Author contributions

M.A.F. and C.A. designed this study and wrote the main manuscript text. M.A.F. analyzed the data and participated in data collection. Authors contributed to the discussion, edited the manuscript, and approved this manuscript.

Competing interests

The authors declare no competing interest
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupportingMaterials.pdf