“Coronaviruses: unknown negative genes in the positive RNA genomes of coronaviruses”

O.P.Zhirnov [1, 2], S.V.Poyarkov [1)]

The Russian-German Academy of Medico-Social and Biotechnological Sciences, Moscow 129009, Russia [(1)]

The D.I.Ivanovsky Institute of Virology, The N.F. Gamaleya National Research Center of Epidemiology and Microbiology, Moscow 123098, Russia [(2)]

Abstract

Coronavirus family has a single-stranded RNA genome encoding 25-30 proteins in different viruses by the mechanism of positive-sense strategy. Extended open reading translation frames (genes) were found to locate under a negative-sense polarity in all coronaviruses genomes. These negative-sense genes varies in the range of 150-450 nt to encode negative genes polypeptides (NGP) with mol. wt. 5-30 kDa. It implies that coronaviruses besides positive genome strategy may have "a dark side of the Moon" expressing genes and virions through the negative strategy. It is noteworthy, that positive- and negative-sense genes colocolized in the same RNA regions of coronavirus genome, so called stacking genes. Ambisense stacking of genes in coronavirus genomes significantly increases virus diversity, genetic potential and extend virus-host adaptation pathway possibilities.

Key words: coronaviruses, ambisense strategy, virus diversity, host adaptation

There are four ambisense virus genera (phlebo-, tospo-, arena-, and bunyaviruses), which are well known to realize both positive- and negative-sense genome RNA strategies to encode viral proteins **[1].** Ambisense genes of these virus genera locate in separate areas of the genome RNA without their overlapping and stacking **[1].** Earlier, we have observed that influenza A virus (the orthomyxovirus) having the negative sense RNA genome is able to encode viral proteins via positive sense genome strategy **[2].** It is noteworthy, that in the case of flu viruses positive- and negative-sense genes colocolized in the same RNA regions, so called stacking genes **[2].**

Coronavirus family is composed of positive sense RNA genome viruses containing lipid envelope, a positive stranded RNA enveloped viruses **[3].** Coronavirus genome encodes 2 sets of positive sense genes: nonstructural nsp1-nsp16 group and structural group integrating structural proteins S, M, E, HE, N and several accessory proteins (asp). All these proteins are well known to be expressed through positive RNA strategy**.** Here we have scanned genomes of human and animal coronaviruses by in silico approach to find novel genes in negative sense genome polarity.

This analysis have identified prolonged open reading frames (ORFs) (150-450 nucleotide long) in all coronaviruses known so far (fig.1; table 1). Notably, similar to flu A virus genomes, positive- and negative-sense genes of coronaviruses are colocolized in the same regions of the RNA genome, so called stacking genes. These frames started with either classical AUG or alternative, such as CUG **[4],** codons and terminated by stop-codons (UGA, GGA, or UAG) and, thus, were potentially able to express negative genes polypeptides (NGP) with the mol. wt. range of 5-30 kD. The genome location pattern of these ORFs may be considered as a specific marker (virus signature) in evolutionary and diagnostic studies of coronavirus family (fig.1). For example, SARS-Cov2 and BAT/RTG13 viruses are more closely related having one large AUG/NGP4 gene, whereas pangolin-Cov virus has 2 negative-sense genes in this area of the genome RNA and seems to be evolved more distantly from the first ones.

Next, we have suggested that to be expressed in infected cells, the revealed ORFs could have the IRES (the Internal Ribosomal Entry Site). This idea has been examined by in silico approach with IRESpred program (HTTP://bioinfo.net.in/IRESPred), which uses the algorithm of structural IRES similarities with previously identified ones in many viruses **[5].** Our analysis performed with SARS-Cov2 virus has found a translation unit of at least 4 NGP genes (NPG1-NPG4) at the 3'-end of virus the (-)RNA genome replica (Fig.2A). This program has allowed to identify a clear IRES-like structures enriched with 10 and 16 canonical "hair-pins" in the regions 8100-8599 nt (IRES 1) and 6792-6488 nt (IRES 2) (counting from the 5' end of the genome (+)RNA) and

preceded the start AUG of the NGP4 (6489-6187 nt) (Fig. 2B). These IRES motifs clearly suggest that 3'-unit of translation cassette in the (-)cRNA can function as template RNA for synthesis of four unique coronavirus proteins NGP1-NGP4 with mol. wt. 15-30 kD. Moreover, this genome cassette has additional properties of gene expression unit, such as 3'- poly A (a start signal) and 5' poly U (a termination signal) tracts. These signals can be recognized by the viral RNA polymerase to initiate its binding and terminate the synthesis of the NGP mRNA containing 5' IRES and 3' poly A tail (it is shown on Fig. 2A). Similar translation cassette at the 3'-end of complimentary virus (-)cRNA molecule was also found in MERS-Cov, Bat, Pangolin, and other animal coronaviruses.

Expression and function of the revealed negative sense genes of coronaviruses remain to be determined. (i) It should be mentioned that these NGP genes are evolutionary stable and exist for a long time in all human and animal coronaviruses that assume their biological determination. (ii) Ambisense stacking of genes in coronavirus genomes significantly increases virus diversity, genetic potential and extend virus-host adaptation pathway possibilities. (iii) Existence of numerous ambisense genes opens a new avenue for coronavirus reproduction where one virus genome can produce a multiple progeny population of virions possessing identical genome RNA and different protein compositions. In this case, a part of virions decorated with NGP proteins could be hidden from us, as "the dark side of the Moon". (iv) The expression of coronavirus "negative" genes may have a host (tissue)-dependent regulation facilitating immune escape of overcovered virions and specific pathogenetic pathways in the host(s) where the up-expression of the virus NGP genes occurs. Further studies will shed light on this ambisense concept of human and animal coronaviruses.

Authors sincerely gratitude G.P. Georgiev and D.K. Lvov, the Academicians of the Russian Academy of Sciences, for helpful discussions and advises. The work has no financial support. The authors don't have a conflict of interests. This study does not develop animal experiments.

References

1. Nguyen M, Haenni AL. Expression strategies of ambisense viruses. Virus Res. 2003 Jun;93(2):141-50. Review. PubMed PMID: 12782362.
2. Zhirnov OP. Unique Bipolar Gene Architecture in the RNA Genome of Influenza A Virus. Biochemistry (Mosc). 2020 Mar;85(3):387-392. doi: 10.1134/S0006297920030141. PMID: 32564743; PMCID: PMC7222887.
3. Fehr R. A., Perlman S. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol. Biol. 2015;1282:1–23. doi: 10.1007/978-1-4939-2438-7_1.
4. Kearse M.G., and Wilusz J.E. Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. 2017 Sep 1; 31(17): 1717–1731. doi: 10.1101/gad.305250.117 PMCID: PMC5666671
5. Kolekar P, Pataskar A, Kulkarni-Kale U, Pal J, Kulkarni A. IRESPred: Web server for Prediction of Cellular and Viral Internal Ribosome Entry Site (IRES). Sci Rep. 2016 Jun 6;6:27436. doi: 10.1038/srep27436. PMID: 27264539; PMCID: PMC4893748.

Legends to the figures

Fig. 1. Location of negative-sense genes in the RNA genomes of human and animal coronaviruses.

Genome location of the large negative-sense ORFs (≥300 nt) starting with classical AUG initiation codon in genomes of different coronaviruses are shown by arrows supplied with start nucleotide position. Analysis of the ORFs location was done based on Genbank sequencing data. Viruses: (a) – human coronavirus HCov-229E (ac.n. NC_002645.1); (b) SARS-Cov (NC_004718.3); (c) MERS-Cov (NC_019843.3); (d) SARS-Cov2 (MT635445.1); (e) Pangolin-Cov (MT040335.1); (f) avian bronchitis coronavirus (NC_001451.1); (g) porcine coronavirus HKU-15 (NC_039208.1); (h) bat coronavirus Bat/RATG13 (MN996532.1); (k) human coronavirus HKU1 (NC_006577.2). Nucleotide scale number counting from the 5' end in the virus genome (+)RNA is outlined at the bottom. Areas of positive sense nsp and structural genes (S, E, HE, M, N and accessory proteins) are indicated on the top.

Fig.2. IRES structures and translation cassette unit in the (-)cRNA of SARS-Cov2.

[A]. 3’ end area of the subgenomic (-)cRNA complimentary to the virus genome 5’ end (+)vRNA of SARS-Cov2 (ac.n. MT635445.1) is displayed. Five ORFs for NGP1-NGP5 beginning either classical AUG (NGP4) or alternative CUG (NGP1-3, NGP5) codons are shown by arrows. Nucleotide counting number from the 5’ end of (+)vRNA are shown for each NGP ORFs. Phases of the translation frame (fr) for each NGP are estimated regarding the frame of NGP4 (fr.0) as follows: NGP1and 2 (fr. +1), NGP3 (fr.0). **[B].** IRES-like structures enriched with 16 and 10 canonical “hair-pins” RNA elements in the regions 8100-8599 nt (IRES 1) and 6488-6792 nt (IRES 2), respectively, were predicted by the IRESpred program **[5].** These IRES-like structures 1 and 2 have significant free energy value as low as -99,4 and -73,8 kkal/mol, respectively.

Table 1.

Characteristics of negative sense genes revealed in genomes of different coronaviruses.

Viruses and viral genomes ac.n.	Number of NGP ORFs in virus genome [1)]	M.W. range of the large NGPs [2)]
Alpha-coronaviruses		
HCov-229E NC_002645.1	29/1/29/5	12.37-14.41
Beta-coronaviruses		
SARS-Cov1 NC_004718.3	34/0/35/2	11.47 - 14.98
MERS NC_019843.3	32/8/23/3	11.07 - 18.55
SARS-Cov2 MT635445.1	21/1/26/4	10.85 - 17.18
Pangolin-CoV MT040335.1	29/3/17/4	10.8 -19.9
HCov-HKU1 NC_006577.2	15/1/13/2	11.47 - 14.98
Bat coronavirus RATG13 MN996532.1	17/2/29/1	10,95- 19,7
Murine hepatitis virus A59 FJ884687.1	29/5/42/7	11,2-36,8
Bovine coronavirus BCoV-ENT NC_003045.1	25/1/26/0	20,8
Gamma-coronaviruses		
Avian infectious bronchitis virus NC_001451.1	20/6/8/3	12.73 - 26.46
Delta-coronaviruses		
Porcine coronavirus HKU15 NC_039208.1	26/5/29/3	11.19 - 17.35

1) Negative sense genes were identified by in silico approach using the Open Reading Frame Finder program (https://www.ncbi.nlm.nih.gov/orffinder/). First and second digits show overall and large gene numbers ORFs starting with classical AUG, respectively. Third and fourth numbers show overall and large gene numbers ORFs starting with alternative CUG, respectively. Large genes have more than 300 nt long. GenBank ac.n. of the virus genomes are indicated.

2) A range of mol. wt. NGP polypeptides encoded by the large negative polarity genes (≥300 nt) starting either with AUG or CUG codons are outlined.