The SARS-CoV-2 arginine dimers

Antonio R. Romeu (antonioramon.romeu@iubilo.urv.cat)
University Rovira i Virgili

Enric Ollé
University Rovira i Virgili

Short Report

Keywords: SARS-CoV-2, Sarbecovirus, Arginine dimer, Arginine codons usage, Bioinformatics

DOI: https://doi.org/10.21203/rs.3.rs-770380/v1

License: ☕️ This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
The SARS-CoV-2 arginine dimers

Antonio R. Romeu\textsuperscript{1} and Enric Ollé\textsuperscript{2}

\textsuperscript{1}: Chemist. Professor of Biochemistry and Molecular Biology. University Rovira i Virgili. Tarragona. Spain. Corresponding author. Email: antonioramon.romeu@iuibo.urv.cat

\textsuperscript{2}: Veterinarian, Biochemist. Associate Professor of the Department of Biochemistry and Biotechnology. University Rovira i Virgili. Tarragona. Spain. Email: enric.olle@urv.cat

Abstract

Arginine is present, even as a dimer, in the viral polybasic furin cleavage sites, including that of SARS-CoV-2 in its protein S, whose acquisition is one of its characteristics that distinguishes it from the rest of the sarbecoviruses. The CGGCGG sequence encodes the SARS-CoV-2 furin site RR dimer. The aim of this work is to report the other SARS-CoV-2 arginine pairs, with particular emphasis in their codon usage. Here we show the presence of RR dimers in the orf1ab related non structural proteins nsp3, nsp4, nsp6, nsp13 and nsp14A2. Also, with a higher proportion in the structural protein nucleocapsid. All these RR dimers were strictly conserved in the sarbecovirus strains closest to SARS-CoV-2, and none of them was encoded by the CGGCGG sequence.

Key words

SARS-CoV-2, Sarbecovirus, Arginine dimer, Arginine codons usage, Bioinformatics

Introduction

Arginine (R) is a polar and non-hydrophobic amino acid, with a positive charged guanidine group, a physiological pH, linked a 3-hydrocarbon aliphatic chain. Arginine participates in the binding of negatively charged substrates and/or protein actives sites (1). Consistently, arginine is involved in viral polybasic proteolitic cleavage sites, even as a dimers, as substrate of the ubiquitously expressed serine protease furin (2,3).

In this context, a notable characteristic of the SARS-CoV-2 is the acquisition of a polybasic cleavage site (PRRAR) at the S1-S2 boundary of the S protein, which is recognized by the furin protease, that greatly mediates the fusion of human cell and viral membranes, and the rapid human-to-human virus transmission (4-7). That acquisition was achieved through an insertion of four amino acids (PRRA) in the S protein. Neither sarbecoviruses nor bat sarbecovirus strains closest to SARS-CoV-2, have a polybasic cleavage site (4).

However, this site is common in viral proteins, such as the hemagglutinin (H5) protein of the avian influenza viruses (2) or the S protein of some of the seventh coronavirus known to infect humans: HCoV-HKU1 (RRKRR-756, coordinate based on S protein), HCoV-OC43 (RRSRR-764) and MERS-CoV (MLKRR-700) (4).

Another notable characteristic of the SARS-CoV-2 is the CGG arginine codon used encoding the RR dimer of the polybasic furin site. The CGGCGG nucleotide sequence was not previously seen encoding RR pairs of viral furin sites (8). Arginine is encoded by the codons CGU, CGC, CGA, CGG, AGA, and AGG. In SARS-CoV-2 the arginine CGG codon is the less often used (frequency 0.03) (8).

The aim of this work is to report the other RR dimers in the SARS-CoV-2. That is, how many are there? In what proteins? How are they coded? And, going a little further, are these RR dimers conserved in the sarbecovirus strains closest to SARS-CoV-2? How are they encoded?
Methods

The source of information were NCBI GenBank and GISAID databases. The SARS-CoV-2 reference sequence (NCBI, GenBank NC_045512.2 and GISAID, EPI_ISL_402124), WIV04 isolate, was also used as reference sequence in this study. The sarbecoviruses used were the following: four human SARS-CoV-2 (WIV04, Wuhan-Hu-1, CDC-CruiseA-12, INMI1), nine bat closest to SARS-CoV-2 (RmYN02 (9), RatG13 (19), RShsTT200, RShsTT182 (11), RpYN06 (12), RacCS203 (13), PrC31, ZC45, ZXC21 (14)), six from pangolin (MP789, GX-P5L, GX-P4L, GX-P1E, GX-P5E, GX-P2V), two bat sarbecovirus more phylogenetically distant to SARS-CoV-2 (BM48-31, BtKY72) and two SARS-CoV sarbecovirus (human, HKU-39849; and ferret Tor 2/FP1-10912) (Table 1). First, it was identified the proteins of the SARS-CoV-2 having the RR pair. Second, for the sarbecovirus used in this study, it was created an in-house database of their genes and their orthologous. Since not all virus genomes were fully annotated, for each of these sarbecovirus, the gene sequences were obtained through successive pairwise BLASTn analyses, using gene sequences of the SARS-CoV-2 reference sequence as a query. The protein sequences were obtained by translating the gene sequences. Multiple sequence alignments were created by the EMBL Clustal Omega (v.1.2.4) using default parameters. Gene and protein multi-alignments were done separately. In the gene multi-alignments, the RR dimer codons were identified, from the position of the RR dimers in the corresponding protein multi-alignments multiplied by three (the RR pairs codons were not identified using codon multiple alignments). All Tables are sorted according the pairwise BLASTn identity percentage between the sarbecovirus genome sequence and the SARS-CoV-2 genome reference sequence.

Results and Discussion

According to the SARS-CoV-2 genome structure (5), from the 5' end, there are two large open reading (ORF1a and ORF1b) covering two-thirds of the RNA genome, and encoding 15 non-structural proteins (nsp) that compose the viral replication and transcription complex. It is followed by the other third of the genome, already in the region of the 3' end that the encodes structural proteins, namely spike (S), envelope (E), membrane (M), nucleocapsid (N). All these SARS-CoV-2 proteins were analysed for the presence of RR dimers. Results are shown in Table 2.

In the S protein, the only RR dimer is that of the polybasic furin cleavage site. However, RR pairs were found in some orf1ab related non-structural proteins that make up by the RNA polymerase complex. Also, RR pairs were found and in the structural protein nucleocapsid, encoded by the N gene. Among the non-structural proteins, RR pairs were in the papain-like protease (nsp3); the proteins involved in the formation of replication compartments (nsp3, nsp4 and nsp6); the helicase (nsp13); and, and in the 3'–5' exonuclease that assists RNA synthesis with a unique RNA proofreading function (nsp14) (5).

Regarding the nucleocapsid protein, it is involved in the package of the positive strand viral genome RNA into a helical ribonucleocapsid (15). Keeping in mind the positive charged of the arginine guanidine group, the presence of the four RR dimers in the SARS-CoV-2 nucleocapsid protein, agrees with what is known that arginine residues are essential in viral capsid assembly (1).

Interestingly, the RR dimers found in that SARS-CoV-2 proteins (apart from that of the S protein) were strictly conserved in the sarbecoviruses strains closely phylogenetically related to it. This suggests a fundamental biological role of these RR doublets. Table 3 shows again the sarbecoviruses used in this study, with the genomic coordinates of the genes, or part of them, that harbour the encoding RR dimers.

Table 4 shows these RR dimers in detail. However, in some orf1ab related non structural proteins from the more distant sarbecoviruses strains, the lysine (K) took the place of the arginine. Lysine, that is also amphipathic, is also highly present in viral polybasic fury cleavage sites (2).
However, the main meaning of table 4 was to show the triplets that encode the RR dimers. It was not a surprise that the AGA arginine majority codon in SARS-CoV-2 (frequency 0.445) (8) was the most frequent encoding the RR dimers. In few occasions the repetition of the same codon appeared, but when it took place, it was in the AGAAGA sequence. On the other hand, the CGGCGG sequence, encoding the RR dimer of the SARS-CoV-2 furin site, did not appear in any case.

Acknowledgements

This work has not been awarded grants by any research-supporting institution.

Competing interest declaration

All authors declare that they have no conflicts of interest.

References

1. Michael J. Harms, Jamie L. Schlessman, Gloria R. Sue, and Bertrand Garcia-Moreno E. Arginine residues at internal positions in a protein are always charged. Proceedings of the National Academy of Sciences of the United States of America (PNAS). 108(47):18954-18959, 2011. PMID: 22080604. doi:10.1073/pnas.1104808108.

2. Elisabeth Braun, Daniel Sauter. Furin-mediated protein processing in infectious diseases and cancer. Clin. Transl. Immunol. E1073, 2019. PMID: 31406574. doi.org/10.1002/cti2.1073.

3. Imène Kara, Marjorie Poggi, Bernadette Bonardo, Roland Govers, Jean-François Landrier, Sun Tian, Ingo Leibiger, Robert Day, John W M Creemers, Franck Peiretti. The Paired Basic Amino Acid-cleaving Enzyme 4 (PACE4) Is Involved in the Maturation of Insulin Receptor Isoform B. J. Biol. Chem. 290:2812-2821, 2015. PMID: 25527501. doi: 10.1074/jbc.M114.592543.

4. Kristian G. Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, Robert F Garry. The proximal origin of SARS-CoV-2.

5. Philip V'kovski, Annika Kratzel, Silvio Steiner, Hanspeter Stalder, Volker Thiel. Coronavirus biology and replication: implications for SARS-CoV-2. Nat. Rev. Microbiol. Oct 28;1-16, 2020. PMID: 33116300. doi: 10.1038/s41579-020-00468-6.

6. Shuai Xia, Qiaoshuai Lan, Shan Su, Xinling Wang, Wei Xu, Zezhong Liu, Yun Zhu, Qian Wang, Lu Lu, Shibo Jiang. The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. Signal Transduct. Target Ther. 5:92, 2020. PMID: 32532959. doi.org/10.1038/s41392-020-0184-0.

7. Markus Hoffmann, Hannah Kleine-Weber, Stefan Pöhlmann. Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Mol. Cell 78:779-784, 2020. PMID: 32362314. doi: 10.1016/j.molcel.2020.04.022.

8. Antonio R. Romeu, Enric Ollé. SARS-CoV-2 and the Secret of the Furin Site. Preprints 2021, 2021020264 (doi: 10.20944/preprints202102.0264.v1).

9. Hong Zhou, Xing Chen, Tao Hu, Juan Li, Hao Song, Yanran Liu, Peihan Wang, Di Liu, Jing Yang, Edward C. Holmes, Alice C. Hughes, Yuhai Bi, Weifeng Shi. A Novel Bat Coronavirus Closely Related to SARS-CoV-2
Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. Current Biology. 30(11):P2196-2203.e3, 2020. PMID:34147139. doi: 10.1016/j.cell.2021.06.008.

10. Peng Zhou, Xing-Lou Yang, Xian-Guang Wang, Ben Hu, Lei Zhang, Wei Zhang, Hao-Rui Si, Yan Zhu, Bei Li, Chao-Lin Huang, Hui-Dong Chen, Jing Chen, Yun Luo, Hua Guo, Ren-Di Jiang, Mei-Qin Liu, Ying Chen, Xu-Rui Shen, Xi Wang, Xiao-Shuang Zheng, Kai Zhao, Quan-Jiao Chen, Fei Deng, Lin-Lin Liu, Bing Yan, Fa-Xian Zhan, Yan-Yi Wang, Geng-Fu Xiao, Zheng-Li Shi. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273, 2020. PMID: 32015507. doi:10.1038/s41586-020-2012-7. (Addendum, 588(7836):E6, 2020. PMID: 3319918. doi:10.1038/s41586-020-2951-z).

11. Vibol Hul1, Deborah Delaune, Erik A Karlsson1, Alexandre Hassanin, Putita Ou Tey, Artem Baidaliuk, Fabiana Gámbaro, Vuong Tan Tu, Lucy Keatts, Jonna Mazet, Christine Johnson, Philippe Buchy, Philippe Dussart, Tracey Goldstein, Etienne Simon-Lorière, Veasna Duong. A novel SARS-CoV-2 related coronavirus in bats from Cambodia. bioRxiv. January 26, 2021. doi.org/10.1101/2021.01.26.428212.

12. Hong Zhou, Jingkai Ji, Xing Chen, Yuhai Bi, Juan Li1, TaoHu, Hao Song, 3Yanhua Chen, Mingxue Cui, Yanyan Zhang, Alice C. Hughes, Edward C. 4Holmes, Weifeng Shi1. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. bioRxiv. March 8, 2021. doi/10.1101/2021.03.08.434390.

13. Supaporn Wacharapluesadee, Chee Wah Tan, Patarapol Maneenorn, Prateep Duengkae, Feng Zhu, Yutthana Joyjinda, Thongchai Kaewpom, Wan Ni Chia, Weenassarin Ampoot, Beng Lee Lim, Kanthita Worachotsueptrakun, Vivian Chih-Wei Chen, Nutthinee Sirichan, Chanida Ruchrisiarod, Apaporn Rodpan, Kirana Noradechanon, Thanawadee Phaihana, Niran Jantarat, Boonchu Thongnumchaima, Changchun Tu, Gary Crameri, Martha M Stokes, Thiravat Hemachudha, Lin-Fa Wang. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. Nature communications. 12(1):972, 2021. PMID: 33563978. doi: 10.1038/s41467-021-21240-1.

14. Dan Hu, Changqiang Zhu, Lele Ai, Ting He, Yi Wang, Fuqiang Ye, Lu Yang, Chenxi Ding, Xuhui Zhu, Ruicheng Lv, Jin Zhu, Bachar Hassan, Youjun Feng, Weilong Tan & Changjun Wang. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. Emerging Microbes & Infections. 7:154, 2018. PMID: 30209269. doi: 10.1038/s41426-018-0155-5.

15. 12. UniProtKB - P0DTC9 (NCAP_SARS2). SARS-CoV-2 Nucleoprotein. Accessed July 25, 2021. https://www.uniprot.org/uniprot/P0DTC9.
Table 1. Sarbecovirus used in the analysis for the conserved SARS-CoV-2 arginine doublets

| Sarbecovirus name | Host            | Country  | Year | Accession num. | % Identity* |
|-------------------|-----------------|----------|------|----------------|-------------|
| WIV04             | Human (SARS-CoV-2) | China    | 2019 | MN906528.1     | Reference   |
| Wuhan-Hu-1        | Human (SARS-CoV-2) | China    | 2019 | NC_045512.2    | 100.00      |
| CDC-CruiseA-12    | Human (SARS-CoV-2) | USA      | 2020 | MT159709.2     | 99.99       |
| INMI1             | Human (SARS-CoV-2) | Italy    | 2020 | MT066156.1     | 99.99       |
| RmYN02            | Bat             | China    | 2019 | EPI_ISL_412977 | 97.19       |
| RatG13            | Bat             | China    | 2013 | MN996532.2     | 96.14       |
| RShsTT200         | Bat             | Cambodia | 2020 | EPI_ISL_852604 | 94.57       |
| RShsTT182         | Bat             | Cambodia | 2020 | EPI_ISL_852605 | 94.57       |
| RpYN06            | Bat             | China    | 2020 | EPI_ISL_699446 | 94.00       |
| RacCS203          | Bat             | Tailandia| 2020 | Mw251308.1     | 94.00       |
| PrC31             | Bat             | China    | 2018 | Mw703458.1     | 94.00       |
| MP789             | Bat             | China    | 2019 | MT121216.1     | 90.11       |
| ZC45              | Bat             | China    | 2017 | MG772933.1     | 89.12       |
| ZXC21             | Bat             | China    | 2015 | MG772934.1     | 88.65       |
| GX-P5L            | Pangolin        | China    | 2017 | MT040335.1     | 85.98       |
| GX-P4E            | Pangolin        | China    | 2017 | MT040333.1     | 85.97       |
| GX-P1E            | Pangolin        | China    | 2017 | MT040334.1     | 85.95       |
| GX-P5E            | Pangolin        | China    | 2017 | MT040336.1     | 85.95       |
| GX-P2V            | Pangolin        | China    | 2018 | MT072864.1     | 85.94       |
| BM48-31           | Bat             | Bulgaria | 2008 | NC_014470.1    | 89.79       |
| BtKY72            | Bat             | Kenya    | 2007 | KY352407.1     | 88.84       |
| HKU-39849         | Human (SARS-CoV) | China    | 2003 | GU553363.1     | 82.30       |
| Tor 2/FP1-10912   | Ferret (SARS-CoV) | USA      | 2003 | JX163926.1     | 82.30       |

* Pairwise NCBI-BLASTn percentage of identity, using the NC_045512.2 SARS-CoV-2 reference sequence (WIV04) as a query.
Table 2. The SARS-CoV-2 proteins having RR dimer(s) and their genes. The coordinates are based on the reference sequence. Data of the RR pair involved in the polybasic furin cleavage site of the S protein is highlighted in red.

| Gene     | Genome coord. | RR     | Protein | Codons               | Gene coord. |
|----------|---------------|--------|---------|----------------------|-------------|
| nsp3     | 2720-8553     | GVRR   | 1614    | GGTGTTAGAAGG         | 4842        |
| nsp4     | 8555-10054    | RFRR   | 306     | AGGTGTGGAAGA         | 918         |
| nsp6     | 10973-11842   | LKRR   | 401     | CTAAAGAGACGT         | 1203        |
| nsp13    | 16237-18039   | GARR   | 138     | GGTGCTAGGAAGA        | 414         |
| nsp14A2  | 18040-19620   | CIRR   | 22      | TGCTACAGTAGA         | 66          |
|          |               | TCRR   | 443     | ACTTTGC6GCGT         | 1329        |
|          |               | IPRR   | 595     | ATTCACG6TAGG         | 1785        |
| S        | 21563-25384   | TYRR   | 53      | ACCTATAGAAAGA        | 159         |
| ORF9 (N) | 28274-29533   | KQRR   | 41      | AAACAACG6TCGG        | 123         |
|          |               | YTRR   | 89      | TACTACCAGAAGA        | 267         |
|          |               | ATRR   | 93      | GCTACACAGCGA         | 279         |
|          |               | FGRR   | 277     | TCTCGACAGACGT        | 831         |

Nucleocapsid
Table 3. Selected sarbecovirus genes encoding proteins with the presence of the arginine doublet. For genes derived from the orf1ab polyprotein, it is shown the coordinates of the genomic region, that in the pairwise BLAST search matched with the corresponding gene of the reference genome (WIV04). For the N gene, structural nucleocapsid protein N, in the genomes which are annotated, the GenBank access number is shown.

| Sarbecovirus name | nsp3         | nsp4         | nsp6         | nsp13        | nsp14A2      | Nucleocapsid (N) |
|-------------------|--------------|--------------|--------------|--------------|--------------|------------------|
| WIV04             | 2720-8553    | 8555-10054   | 10973-11842  | 16237-18039  | 18040-19620   | QHR63268.1       |
| Wuhan-Hu-1        | 2720-8553    | 8555-10054   | 10973-11842  | 16237-18039  | 18040-19620   | YP_009724397.2   |
| CDC-CruiseA-12    | 2720-8553    | 8555-10054   | 10973-11842  | 16237-18039  | 18040-19620   | QII57215.1       |
| INMI1             | 2720-8553    | 8555-10054   | 10973-11842  | 16237-18039  | 18040-19620   | QIA98561.1       |
| RmYN02            | 2708-8535    | 8537-10036   | 10955-11824  | 16219-18021  | 18022-19602   | 28104-29363      |
| RatG13            | 2720-8550    | 8552-10051   | 10970-11839  | 16234-18036  | 18037-19617   | QHR63398.1       |
| RShsTT200         | 2720-8532    | 8534-10033   | 10952-11821  | 16216-18018  | 18019-19599   | 28177-29436      |
| RShsTT182         | 2714-8526    | 8528-10027   | 10946-11815  | 16210-18012  | 18013-19593   | 28171-29430      |
| RpYN06            | 2720-8550    | 8552-10051   | 10970-11839  | 16234-18036  | 18037-19617   | 28186-29445      |
| Rac9S203          | 2762-8553    | 8555-10054   | 10973-11842  | 16237-18039  | 18040-19620   | QQM18872.1       |
| PrC31             | 2682-8512    | 8514-10013   | 10932-11801  | 16196-17997  | 17999-19579   | QSQ26168.1       |
| MP789             | 3127-8412    | 8414-9913    | 10832-11761  | 16096-17988  | 17899-19479   | QIG59561.1       |
| ZC45              | 2719-8540    | 8542-10041   | 10966-11829  | 16224-18026  | 18027-19601   | AVP78038.1       |
| ZXC21             | 2719-8474    | 8476-9975    | 10894-11763  | 16158-17960  | 17961-19535   | AVP78049.1       |
| GX-P5L            | 3312-8530    | 8601-10024   | ns           | 16214-18016  | 18017-19597   | QIA48639.1       |
| GX-P4L            | 3312-8530    | 8601-10024   | ns           | 16214-18016  | 18017-19597   | QIA48621.1       |
| GX-P1E            | 3312-8530    | 8601-10024   | ns           | 16214-18016  | 18017-19597   | QIA48630.1       |
| GX-P5E            | 3312-8530    | 8601-10024   | ns           | 16214-18016  | 18017-19597   | QIA48648.1       |
| GX-P2V            | 3294-8512    | 8583-10006   | ns           | 16196-17998  | 17999-19579   | QIQ54651.1       |
| BM48-31           | 4063-8379    | 8569-9883    | ns           | 16066-17858  | 17872-19439   | YP_003858591.1   |
| BtkY72            | ns           | ns           | 10850-11698  | 16903-17885  | 17896-19470   | APO40586.1       |
| HKU-39849         | 3843-8443    | ns           | ns           | 16127-17929  | 17930-19500   | ADC35492.1       |
| Tor 2/FP1-10912   | 3843-8443    | ns           | ns           | 16127-17929  | 17930-19500   | AFR58723.1       |

ns: no significant similarity found in the pairwise BLASTn search using the corresponding gene of the NC_045512.2 SARS-CoV-2 reference s genome (WIV04) query
Table 4. Arginine doublet y their codons of SARS-CoV-2 and conserved in the closest sarbecovirus species to it. Data was extracted from del protein and encoding gene multiple alignment, respectively. Numbers on the right stand for the amino acid and nucleotide position in the reference genome (WIV04) sequences. * represents strictly conserved residues in the multiple alignment.

| Sarbecovirus | Pair 1 | nbecovirus | Pair 1 | Pair 2 | Sarbecovirus | Pair 1 |
|--------------|--------|------------|--------|--------|--------------|--------|
| WIV04        | RR 1614| AGAAGG 4842| WIV04  | RR 306 | AGAAGA 918   | RR 401 | AGACGT 1203| WIV04  | RR 138 | AGGAGA 414 |
| Wuhan-Hu-1   | RR AGAAGG| Wuhan-Hu-1  | RR AGAAGA| RR AGACGT| Wuhan-Hu-1  | RR AGGAGA|
| CDC-CruiseA-12| RR AGAAGG| CDC-CruiseA-12| RR AGAAGA| RR AGACGT| CDC-CruiseA-12 | RR AGGAGA|
| INMI1        | RR AGAAGG| INMI1      | RR AGAAGA| RR AGACGT| INMI1       | RR AGGAGA|
| RmYN02       | RR AGAAGG| RmYN02    | RR AGAAGA| RR AGACGT| RmYN02      | RR AGGAGA|
| RaTG13       | RR AGAAGG| RaTG13     | RR AGAAGA| RR AGACGT| RaTG13      | RR AGGAGA|
| RShSTT200    | KR AAAAGG| RShSTT200 | RR AGAAGA| RR AGACGT| RShSTT200   | RR AGGAGA|
| RShSTT182    | KR AAAAGG| RShSTT182 | RR AGAAGA| RR AGACGT| RShSTT182   | RR AGGAGA|
| RpYN06       | RR AGAAGG| RpYN06    | RR AGAAGA| RR AGACGT| RpYN06      | RR AGGAGA|
| RacCS203     | RR AGAAGG| RacCS203  | RR AGAAGA| RR AGACGT| RacCS203    | RR AGGAGA|
| PrC31        | RR AGAAGG| PrC31     | RR AGAAGA| RR AGACGT| PrC31       | RR AGGAGA|
| MP789        | RR AGAAGA| MP789     | RR AGAAGA| RR AGACGT| MP789       | RR AGGAGA|
| ZC45         | RR AGAAGG| ZC45      | RR AGAAGA| RR AGACGT| ZC45        | RR AGGAGA|
| ZXC21        | RR AGAAGG| ZXC21     | RR AGAAGA| RR AGACGT| ZXC21       | RR AGGAGA|
| GX-P5L       | RR AGAAGA| GX-P5L    | RR AGAAGA| RR CGTAGA| BtKY72      | ** ** *|
| GX-P4L       | RR AGAAGA| GX-P4L    | RR AGAAGA| RR CGTAGA| ** ** *      | ** ** *|
| GX-P1E       | RR AGAAGA| GX-P1E    | RR AGAAGA| RR CGTAGA| ** ** *      | ** ** *|
| GX-P5E       | RR AGAAGA| GX-P5E    | RR AGAAGA| RR CGTAGA| ** ** *      | ** ** *|
| GX-P2V       | RR AGAAGA| GX-P2V    | RR AGAAGA| RR CGTAGA| ** ** *      | ** ** *|
| BM48-31      | KK AGAAGA| BM48-31   | RR AGGCGT| KR AGAAGA| ** ** *      | ** ** *|
| HKU-39849    | KR AGAAGA|           | ** ** * | ** ** *   |              | ** ** * |
| Tor2-FP1-10912| KR AGAAGA|           | ** ** * | ** ** *   |              | ** ** * |
Table 4. continued

**Table:**

| nsp13. Zinc-binding domain. Helicase, RNA 5’ triphosphatase | nsp14A2. Guanosine N7- methyltransferase. 3’ to 5’ exoribonuclease. Proofreading |
|---|---|
| **becovirus** | **Pair 1** | **Pair 2** | **Pair 3** | **becovirus** | **Pair 1** | **Pair 2** |
| WIV04 | RR 22 | CGTATA 66 | RR 443 | CGGCGT 1329 | RR 595 | CGTATG 1785 | WIV04 | RR 53 | AGAAGA 159 | RR 213 | AGACGTG 639 |
| Wuhan-Hu-1 | RR | CGTATA | RR | CGGCGT | RR | CGTATG | Wuhan-Hu-1 | RR | AGAAGA | RR | AGAAGA |
| CDC-CruiseA-12 | RR | CGTATA | RR | CGGCGT | RR | CGTATG | CDC-CruiseA-12 | RR | AGAAGA | RR | AGAAGA |
| INM11 | RR | CGTATA | RR | CGGCGT | RR | CGTATG | INM11 | RR | AGAAGA | RR | AGAAGA |
| RaYN02 | RR | CGTATA | RR | CGGCGT | RR | CGTATG | RaYN02 | RR | AGAAGA | RR | AGAAGA |
| RShSTT200 | RR | CGTATA | RR | CGGCGT | RR | CGTATG | RShSTT200 | RR | AGAAGA | RR | AGAAGA |
| RShSTT182 | RR | CGTATA | RR | CGGCGT | RR | CGTATG | RShSTT182 | RR | AGAAGA | RR | AGAAGA |
| RpYN06 | RR | CGTATA | RR | CGGCGT | RR | CGTATG | RpYN06 | RR | AGAAGA | RR | AGAAGA |
| RacCS203 | RR | CGTATA | RR | CGGCGT | RR | CGTATG | RacCS203 | RR | AGAAGA | RR | AGAAGA |
| PrC31 | RR | CGCAGA | RR | CGTCGC | RR | CGTATG | PrC31 | RR | AGAAGA | RR | AGAAGA |
| MP789 | RR | AGGAGA | RR | CGGCGT | RR | CGTATG | MP789 | RR | AGAAGA | RR | AGAAGA |
| ZC45 | RR | AGGAGA | RR | CGGCGT | RR | CGTATG | ZC45 | RR | AGAAGA | RR | AGAAGA |
| ZXC21 | RR | AGGAGA | RR | CGGCGT | RR | CGTATG | ZXC21 | RR | AGAAGA | RR | AGAAGA |
| GX-P5L | RR | AGGAGA | RR | CGGCGT | RR | CGTATG | GX-P5L | RR | AGAAGA | RR | AGAAGA |
| GX-P4L | RR | AGGAGA | RR | CGGCGT | RR | CGTATG | GX-P4L | RR | AGAAGA | RR | AGAAGA |
| GX-P1E | RR | AGAAGA | RR | CGGCGT | RR | CGTATG | GX-P1E | RR | AGAAGA | RR | AGAAGA |
| GX-P5E | RR | AGAAGA | RR | CGGCGT | RR | CGTATG | GX-P5E | RR | AGAAGA | RR | AGAAGA |
| GX-P2V | RR | AGAAGA | RR | CGGCGT | RR | CGTATG | GX-P2V | RR | AGAAGA | RR | AGAAGA |
| BM4S-31 | RR | AGGCGA | RR | CGTAGG | RK | CGTAAA | BM4S-31 | RR | AGAAGA | RR | AGAAGA |
| BtKY72 | RR | AGGCGA | RR | CGTAGG | RK | CGTAAA | BtKY72 | RR | AGAAGA | RR | AGAAGA |
| HKU-39849 | RR | AGGAGA | RR | CGCCGT | RR | CGTATG | HKU-39849 | RR | AGAAGA | RR | AGAAGA |
| Tor2-FP1-10912 | * | * | * | * | * | CGTCGC | Tor2-FP1-10912 | * | * | * | * | * |
orf9 N. Nucleocapsid
Packages the positive strand viral genome RNA. Virion assembly

| Sarbecovirus         | Pair 1 | Pair 2 | Pair 3 | Pair 4 |
|----------------------|--------|--------|--------|--------|
| WIV04                | RR 41  | CGTCGG | RR 89  | CGAAGA | RR 93  | AGACGA | RR 277 | AGACGT 831 |
| Wuhan-Hu-1           | RR     | CGTCGG | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| CDC-CruiseA-12       | RR     | CGTCGG | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| INMI1                | RR     | CGTCGG | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| RpYN02               | RR     | CGTCGT | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| RatG13               | RR     | CGTCGG | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| RshSTT200            | RR     | CGTCGT | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| RshSTT182            | RR     | CGTCGT | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| RpYN06               | RR     | CGTCGT | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| RacC3203             | RR     | CGTCGT | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| PrC31                | RR     | CGTCGT | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| MP789                | RR     | CGTCGT | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| ZC45                 | RR     | CGTCGA | RR     | CGTACA | RR     | AGACGA | RR     | AGACGT |
| ZXC21                | RR     | CGTCGA | RR     | CGTACA | RR     | AGACGA | RR     | AGACGT |
| GXR-P5L              | RR     | CGAAAG | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| GXR-P4L              | RR     | CGAAGG | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| GXR-P1E              | RR     | CGAAAG | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| GXR-P5E              | RR     | CGAAAG | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| GXR-P2V              | RR     | CGAAAG | RR     | CGAAGA | RR     | AGACGA | RR     | AGACGT |
| BM48-31              | RR     | AGAAGA | RR     | CGAAGA | RR     | CGACGA | RR     | AGACGA |
| BtKY72               | RR     | AGAAGA | RR     | AGAAGA | RR     | CGACGA | RR     | AGACGA |
| HKU-39849            | RR     | CGCCGA | RR     | CGAAGA | RR     | CGACGA | RR     | AGACGT |
| Tor2-FP1-10912       | RR     | CGCCGA | RR     | CGAAGA | RR     | CGACGA | RR     | AGACGT |

** Note: CGTCGG and CGAAGA are common motifs.