3

Detection and Sequence Characterization of the 3′-End of Coronavirus Genomes Harboring the Highly Conserved RNA Motif s2m

Christine Monceyron Jonassen

Abstract

A remarkably conserved 43-nucleotide-long motif present at the 3′-end of the genomes of several members of the polyadenylated RNA virus families *Astroviridae*, *Coronaviridae*, and *Picornaviridae* can be used for the detection and sequence characterization of the viruses harboring it. The procedure makes use of a primer located in the most conserved core of s2m toward a generic anchored oligo(dT) primer in a semispecific PCR. This strategy allows the sequencing of some 50–100 nucleotides from the 3′-end of the virus genome, representing sufficient sequence information for initiation of further genomic characterization in a rapid amplification of cDNA ends (5′-RACE) and primer walking strategy.

Key words: coronavirus; s2m; PCR; sequencing; 5′-RACE; detection; diagnosis

1. Introduction

S2m is a conserved motif present in the genomes of several members of the RNA virus families *Astroviridae*, *Coronaviridae*, and *Picornaviridae **(1)***. These viruses have single-stranded positive-sense RNA genomes, 6.5 to 32 kb long, with a poly(A) tail at their 3′-ends. S2m is 43 nucleotides long and forms a highly conserved RNA structure ***(2)***. In these very different and rapidly evolving viruses, the remarkably conserved nature of s2m, in both RNA sequence and folding, suggests an essential, but as yet unknown function. This motif is located

From: *Methods in Molecular Biology, vol. 454: SARS- and Other Coronaviruses*,
Edited by: D. Cavanagh, DOI: 10.1007/978-1-59745-181-9_3,

50–150 nucleotides upstream of the 3' poly(A) tail of these virus genomes. Phylogenetic considerations and the genomic position of s2m suggest that ancestors of some of these viruses acquired s2m by horizontal gene transfer.

All of the known group 3 coronaviruses, including infectious bronchitis virus (IBV), turkey coronavirus, pheasant coronavirus, and the recently characterized coronaviruses infecting different wild bird species ***(3)***, as well as the group 2 coronavirus associated with the severe acute respiratory syndrome (SARS) ***(4,5)*** have been found to have s2m.

As the core of s2m is highly conserved in nucleotide sequence, it is an ideal target for identification of the viruses that contain it, and, together with a coronavirus replicase sequence, it was one of the probes that first identified the etiology of SARS as a coronavirus ***(4,6)***. In this chapter the use of s2m as a handle in amplification strategies for obtaining sequence information of novel coronavirus genomes is described.

2. Materials

2.1. RT-PCR

1. The reagents for reverse transcription (RT) include: Superscript III reverse transcriptase (200 U/μl) and 5X first-strand buffer (250 mM Tris-HCl, 375 mM KCl, 15 m*M* $MgCl_2$), DTT (100 mM), dNTP mix (10 mM each), and RNaseOUT recombinant RNase inhibitor (40 U/μL) (all reagents from Invitrogen).
2. The primers used in the initial RT and polymerase chain reaction (PCR) are: anchored oligo$(dT)_{20,}$ 5′-$(T)_{20}$VN, (2.5 μl/μg) (Invitrogen) or anchored oligo$(dT)_{18,}$ 5′-$(T)_{18}$VN, as reverse primers, and s2m-core: 5′ CCG AGT A(C/G)G ATC GAG GG as the sense primer.
3. Reagents for PCR: dNTP mix (10 mM), RNase and DNase-free water (both from Invitrogen), and HotStar Taq DNA polymerase (5 U/μl) and 10X buffer (containing 15 mM $MgCl_2$) (Qiagen).
4. Thermocyclers (MJ Research) are used both for RT and PCR.

2.2. Sequencing and Analysis

1. Agarose (Applied Biosystems) for visualization of PCR products.
2. Sequencing primers are s2m-coreseq: 5′-GAG TA(C/G) GAT CGA GGG TAC, AV12: 5′-$(T)_{18}$GC for sequencing the initial RT-PCR products (3.1) or coronavirus gene-specific primers for sequencing the 5′-RACE PCR products (3.3).
3. Oligo software version 6.68 is used to check for primer hairpin structures and primer dimers and to calculate annealing temperatures for PCR and cycle sequencing.

4. BigDye Terminator Cycle Sequencing kit version 1.1 or version 3.1 (Applied Biosystems) is used for sequencing of short (from initial PCR) or long (5′-RACE PCR) PCR products, respectively.
5. Cycle sequencing reaction is performed on an MJ Research thermocycler.
6. 3130xl Genetic Analyser for sequencing (Applied Biosystems).
7. Sequence analysis programs: Sequencher version 4.1.4 and BioEdit version 7.0.1.

2.3. 5′-Race

1. All reagents for RT are described in Section 2.1.1.
2. For RNase treatment of newly synthesized cDNA, a master mix of two-thirds part RNase H (2 U/μl) (Invitrogen) and one-third part RNase T1 (100–150 U/μl) (Roche) is made. The mix is made fresh, and stored at 4°C until use.
3. cDNA purification: QIAquick PCR purification kit (Qiagen).
4. cDNA yields are measured using a NanoDrop Spectrophotometer (Saveen & Werner).
5. Tailing reagents: Terminal deoxynucleotidyl transferase (TdT) (15 U/μl), 5X TdT buffer (0.5 M potassium cacodylate, 10 mM $CoCl_2$, 1 mM DTT), and dCTP (2 mM) (all from Invitrogen)
6. Sense primer in the 5′-RACE PCR: 5′-RACE abridged anchor primer (5′-GGC CAC GCG TCG ACT AGT ACG GGI IGG GII GGG IIG) (Invitrogen).
7. PCR enzyme used in the long-range 5′-RACE PCR: BD Advantage 2 Taq polymerase with 10X BD Advantage 2 PCR buffer (400 mM Tricine-KOH, 150 mM KOAc, 35 mM $Mg(OAc)_2$, 37.5 μg/μl BSA, 0.05 % Tween 20, and 0.05 % Nonidet-P40) (Clontech).

3. Methods

Determining the 3′-end sequences of unknown RNA can be cumbersome, and the present method makes use of s2m in some of the coronavirus genomes as a handle for determination of the sequences between s2m and the 3′ poly(A) tail of these genomes. The method can be used on RNA extracted from clinical samples (e.g., tracheal or cloacal swabs) that have been confirmed to be positive for coronavirus, e.g., by RT-PCR and sequencing, using pancoronavirus primers ***(7)*** (*see* **Note 1**).

As s2m is located only about 100 nucleotides upstream of the 3′ poly(A) tail when it is present in the genomes of coronaviruses, only limited sequence information can be acertained from the initial amplification product obtained using a primer located within s2m toward a generic primer for polyadenylated RNA. This limited sequence information is, however, sufficient to design virus-specific primers that can be used in a 5′-RACE strategy that, together with primer walking, allow sequencing toward the 5′-end of the virus genomes ***(8)***.

3.1. Initial Amplification Using an s2m-Based Primer toward a Generic Primer of Polyadenylated Genomes in Coronavirus Positive Samples

1. For RT, prepare the following mix (per sample): 4 μl 5X first-strand buffer, 1 μl 100 mM DTT, 1 μl dNTP (10 mM), 1 μl anchored oligo(dT) primer (*see* **Note 2**), 1 μl RNaseOUT (40 U/μl), and 1 μl Superscript III reverse transcriptase. The mix should be made fresh, and stored at 4°C until use.
2. Add 11 μl RNA extract to the 9-μl RT mix, with care taken to avoid air bubbles, as these can oxidize the reverse transcriptase. The RT is performed at 50°C for 30 min, followed by an enzyme inactivation step at 70°C for 15 min, and cooling to 4°C thereafter.
3. PCR is performed using the s2m-core primer toward a generic anchored oligo(dT) primer. Prepare the following PCR mix (per sample): 35.7 μl RNase and DNase-free water, 5 μl 10X PCR buffer, 1 μl dNTP (10 mM), 1 μl sense primer s2m-core (25 μM), 2 μl reverse primer anchored oligo$(dT)_{18}$ (25 μM) and 0.3 μl HotStar Taq DNA polymerase. The mix is made fresh, and stored at 4°C until use.
4. A 5-μl cDNA sample is added to 45 μl PCR mix. The following PCR program is performed: 95°C 15 min for activation of the polymerase, followed by 40 cycles (94°C, 40 sec; 55°C, 20 sec; 72°C, 40 sec); 72°C, 5 min; and cooling at 8°C

3.2. Sequencing and Analysis of the Initial PCR Products

1. 15 μl of the PCR products are visualized on a 2% agarose gel. If s2m is part of the coronavirus genome, one should obtain specific amplification of the 3′-end (*see* **Note 3**). The PCR products are purified directly prior to sequencing reaction, using the QIAquick PCR purification kit according to the manufacturer's instructions, with a final elution volume of 30 μl.
2. The cycle sequencing reaction is performed using the BigDye Terminator version 1.1, especially suited for short PCR products. Add 4 μl BigDye terminator mix, 1.3 μl primer s2m-coreseq (2.5 μM) (*see* **Note 4**) to 4.7 μl purified PCR product, and perform the cycle sequencing reaction as follows: 96°C for 1 min, followed by 25 cycles (96°C, 10 sec; 55°C, 5 sec; 60°C, 4 min), and cooling to 4°C thereafter.
3. The cycle sequencing products are purified from unincorporated labeled dideoxynucleotides and separated by capillary electrophoresis on a 3130xl Genetic Analyser for sequence determination, according to Applied Biosystems' instructions.
4. The sequence obtained gives information on the 3′-end of the coronavirus genome. As s2m is located about 100–150 nucleotides upstream of the poly(A) tail of the coronaviruses harboring it, 50–100 nucleotides can be determined that way, and two new primers are designed, which can be partly overlapping, for use in PCR and sequencing in the 5′-RACE strategy (*see* **Note 5**).

3.3. 5′-RACE for Sequencing and Primer Walking Strategy

1. A new RT is performed, with care taken not to vortex samples, mixing by gentle pipetting to avoid breakage of the genomic RNA molecule and achieve long-range

cDNA synthesis, to get long sequence stretches in the 5′-RACE procedure (*see* **Note 6**). A 10-μl RNA extract is added to 1 μl anchored oligo(dT)$_{20}$ primer (2.5 μg/μl), and incubated for 5 min at 65°C, to get rid of RNA secondary structures. The sample is then rapidly chilled on ice, for 1 min. The sample is then spun down, and 4 μl 5X first-strand buffer, 1 μl 10 mM dNTP, and 1 μl 100 mM DTT are added and mixed gently by pipetting up and down. The samples are then incubated for 1 min at 55°C, and 1 μl RNaseOUT (40 U/μl) and 2 μl Superscript III (200 U/μl) are added to each sample, and the samples are rapidly put back into the thermocycler used for RT. The RT is performed at 55°C, for 80 min, followed by enzyme inactivation at 70°C and cooling at 4°C thereafter (*see* **Note 7**).

2. The samples from reverse transcription are centrifuged for 20 sec at 1538 × *g* to collect all droplets, and incubated at 37°C in a thermocycler. A 1-μl RNase mix is then added to each sample (*see* **Note 8**) and mixed gently by pipetting up and down, and the samples are put back into the thermocycler for further incubation for 30 min at 37°C. When the incubation is finished, the samples are spun down and cooled.
3. The newly synthesized cDNA is purified using the QIAquick PCR purification kit according to the manufacturer's instructions, with the following modifications: all mixing steps are performed gently, and the sample is eluted in 30 μl RNase/DNase-free water prewarmed at 65°C. The cDNA amount is measured spectrophotometrically on a NanoDrop.
4. Cytosine residues are added to the 3'-ends of purified cDNA, in order to make a homopolymeric (C) tail to serve as a handle for semispecific amplification with a specially designed anchor primer. The tailing is performed as follows: 10 μl of the purified cDNA is added to 6.5 μl RNase/DNase-free water, 5 μl 5X tailing buffer, and 2.5 μl 2 mM dCTP. In order to avoid secondary structure in the cDNA that might impair the tailing, a short denaturation step is performed at 94°C for 3 min. The sample is then immediately chilled on ice and spun down. One μl TdT (15 U/μl) is then added, and the tailing reaction is performed at 37°C for 10 min, followed by an enzyme inactivation step at 65°C for 10 min. The sample is then spun down and cooled on ice.
5. The 5′-RACE PCR is performed as follows (*see* **Note 9**): 5 μl tailed sample is added to 45 μl PCR mix consisting of 34.5 μl RNase/DNase-free water, 5 μl 10X BD Advantage 2 PCR buffer, 1 μl dNTP (10 mM), 1 μl coronavirus-specific reverse primer (25 μM), 2.5 μl abridged anchor primer (10 μM), and 1 μl BD Advantage polymerase mix. The cycling conditions include an initial enzyme-activation step at 95°C for 1 min, followed by 35 cycles of 94°C for 30 sec, 55°C for 20 sec, and 68°C for 7 min (*see* **Note 10**). The cycling is followed by a step at 68°C for 7 min and cooling to 4°C thereafter.
6. Prior to sequencing reaction, the PCR product, appearing as a weak smear on a 1% agarose gel electrophoresis, is purified directly using the QIAquick PCR purification kit according to the manufacturer's instructions, with a final elution volume of 30 μl (*see* **Note 11**). Cycle sequencing is performed using BigDye Terminator version 3.1, for long PCR product sequencing, with the following reagents: 4 μl BigDye Terminator mix, 1.3 μl coronavirus-specific reverse primer 2 (2.5 μM), and 4.7 μl PCR product. The cycle sequencing program consists of an initial step

at 96°C for 1 min, followed by 25 cycles of 96°C for 10 sec, 50°C for 5 sec, and 60°C for 4 min, followed by cooling at 4°C.

7. The cycle sequencing products are purified from unincorporated labeled dideoxynucleotides and separated by capillary electrophoresis on a 3130xl Genetic Analyser for sequence determination, according to Applied Biosystems' instructions (*see* **Note 12**).

4. Notes

1. In addition to coronaviruses, this procedure has been successfully applied to amplify and sequence novel astrovirus and picornavirus genomes, and potentially all polyadenylated viral genomes harboring s2m can be identified and characterized with this procedure. However, when used on biological samples that do not contain s2m-harboring viruses, nonspecific amplification of host cell ribosomal RNA is obtained on some occasions.
2. We have generally used 1 μl anchored oligo(dT)$_{18}$ (25 μ*M*) in the RT, but 1 μl Invitrogens anchored oligo(dT)$_{20}$ (2.5 μg/μl, corresponding to ca. 350 μM), or Invitrogens standard oligo(dT)$_{20}$ (50 μM), has also been used successfully as a reverse primer.
3. The obtained PCR products often give a smear on visualization on agarose gel, owing to serial mispriming of the anchored oligo(dT) primer on the 3′ poly(A) tail, leading to long-range accumulation of As in the PCR products that can be observed upon sequencing. However, if amplification is successful, all PCR products, regardless of the number of terminal As, have the same start at s2m, as the s2m-core primer is highly specific, and the PCR product smear can be sequenced using s2m-core or s2m-coreseq primer.
4. As several of the viruses harboring s2m do terminate the genome with GC(A)n, AV12 (2.5 μM) can often be used for sequencing the opposite strand. If low signal intensity is obtained for the sequences, the following modifications can be made in the cycle sequencing program: lowering the annealing temperature to 50°C and/or adding 5–10 cycles in the cycle sequencing reaction.
5. Once a specific reverse primer can be designed as described in Section 3.2, an alternative strategy to 5′-RACE/primer walking is to amplify directly a long PCR product using upstream primers in conserved family-specific motifs, e.g., in the polymerase gene. This will be achieved more easily with viruses that have small genomes, such as astrovirus and picornavirus, especially when the polymerase is encoded toward the 3′- end of the genome, rather than with coronaviruses that have a large genome. In our laboratory, we were able to get sequence information by using the sense 2Bm primer designed by Stephensen et al. in 1999 *(7)* toward the coronavirus specific primer at the 3′-end of the coronavirus genome. However, the PCR product yield was low, and only short sequence stretches could be obtained in this way.
6. The RT in the 5′-RACE procedure is laborious, as no mix containing all the reagents is prepared in advance, but, rather, primer and enzymes are added

individually to each sample. In addition, the 5′-RACE procedure described in the Invitrogen handbook suggests that the RNase treatment of the sample be performed on freshly synthesized cDNA, that has not been frozen. This is why the initial RT step described in Section 3.1, performed in order to check novel coronavirus genomes for the presence of s2m, makes use of a simpler RT protocol.

7. An increased amount of the Superscript III reverse transcriptase (400 U per reaction instead of the normally 200 U) is used for long-range cDNA synthesis in the 5′-RACE strategy. In addition, this enzyme allows the reaction to be performed at a temperature as high as 55°C, which minimizes secondary structures in RNA and is RNase H negative, which increases the yield of full-length cDNAs.
8. It is important to degrade the template RNA from the newly synthesized cDNA, to avoid hybrid renaturation prior to the tailing step. The RNase mix consists of both RNase H to hydrolyze RNA bound to cDNA, and RNase T1 to hydrolyze free RNA molecules.
9. The 5′-RACE PCR is semispecific, using an anchor primer as sense primer, which amplifies the poly(C)-tailed template, and a newly designed coronavirus-specific reverse primer. Even if care has been taken throughout the whole procedure to avoid synthesis of short cDNAs from template RNA, there might be some shorter cDNA molecules that will be more easily amplified in PCRs. It is therefore important to use a PCR enzyme with a proofreading activity, in order to get PCR products from the longer cDNA molecules as well.
10. The long elongation step is chosen to allow amplification of longer cDNAs.
11. As cDNA molecules of different lengths are generated during RT, the 5′-RACE PCR products will be of different lengths, so the 5′-RACE abridged anchor primer cannot be used in cycle sequencing. As all the 5′-RACE PCR products have aligned 3′-ends, the coronavirus-specific 5′-RACE reverse PCR primer or, preferably for increased specificity, a primer slightly upstream of this PCR primer is used in cycle sequencing.
12. Using this procedure, one should obtain at least 500 nucleotides, which would allow the design of new primers both in sense direction for sequence verification and reverse direction for acquisition of new sequence data toward the 5′-end of the genome in a primer walking strategy. The new primers can be used directly in step 5 of the 5′-RACE procedure, if all the steps have been optimized for long-range cDNA synthesis. However, usually no more than 2000–3000 nucleotides are efficiently sequenced from the initial 5′-RACE-RT, and the whole 5′-RACE procedure has to be repeated with a primer for cDNA synthesis as upstream as possible.

References

1. Jonassen, C. M., Jonassen, T. O., and Grinde, B. (1998) A common RNA motif in the 3’ end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus *J. Gen. Virol.* **79**, 715–718.

2. Robertson, M. P., Igel, H., Baertsch, R., Haussler, D., Ares, M., Jr., and Scott, W. G. (2005) The structure of a rigorously conserved RNA element within the SARS virus genome *PLoS Biol.* **3**, E5.
3. Jonassen, C. M., Kofstad, T., Larsen, I. L., Løvland, A., Handeland, K., Follestad, A. and Lillehaug, A. (2005) Molecular identification and characterization of novel coronaviruses infecting graylag geese (*Anser anser*), feral pigeons (*Columbia livia*) and mallards (*Anas platyrhynchos*) *J. Gen. Virol.* **86**, 1597–1607.
4. Ksiazek, T. G., Erdman, D., Goldsmith, C. S., Zaki, S. R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J. A., Lim, W., Rollin, P. E., Dowell, S. F., Ling, A. E., Humphrey, C. D., Shieh, W. J., Guarner, J., Paddock, C. D., Rota, P., Fields, B., DeRisi, J., Yang, J. Y., Cox, N., Hughes, J. M., Le Duc, J. W., Bellini, W. J., and Anderson, L. J. (2003) A novel coronavirus associated with severe acute respiratory syndrome *N. Engl. J. Med.* **348**, 1953–1966.
5. Marra, M. A., Jones, S. J., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S., Khattra, J., Asano, J. K., Barber, S. A., Chan, S. Y., Cloutier, A., Coughlin, S. M., Freeman, D., Girn, N., Griffith, O. L., Leach, S. R., Mayo, M., McDonald, H., Montgomery, S. B., Pandoh, P. K., Petrescu, A. S., Robertson, A. G., Schein, J. E., Siddiqui, A., Smailus, D. E., Stott, J. M., Yang, G. S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T. F., Bowness, D., Czub, M., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla,A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li,Y., Normand, S., Stroher, U., Tipples, G. A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R. C., Krajden, M., Petric, M., Skowronski, D. M., Upton, C., and Roper, R. L. (2003) The genome sequence of the SARS-associated coronavirus *Science* **300**, 1399–1404.
6. Wang, D., Urisman, A., Liu, Y. T., Springer, M., Ksiazek, T. G., Erdman, D. D., Mardis, E. R., Hickenbotham, M., Magrini, V., Eldred, J., Latreille, J. P., Wilson, R. K., Ganem, D., and Derisi, J. L. (2003) Viral discovery and sequence recovery using DNA microarrays *PLoS Biol.* **1**, E2.
7. Stephensen, C. B., Casebolt, D. B., and Gangopadhyay, N. N. (1999) Phylogenetic analysis of a highly conserved region of the polymerase gene from 11 coronaviruses and development of a consensus polymerase chain reaction assay *Virus Res.* **60**, 181–189.
8. Jonassen, C. M., Jonassen, T. O., Sveen, T. M., and Grinde, B. (2003) Complete genomic sequences of astroviruses from sheep and turkey: comparison with related viruses *Virus Res.* **91**, 195–201.