SARS-CoV-2 genomic and subgenomic RNAs in diagnostic samples are not an indicator of active replication

Soren Alexandersen[1,2,3,*], Anthony Chamings [1,2], and Tarka Raj Bhatta [1,2,]

[1]Geelong Centre for Emerging Infectious Diseases, Geelong, VIC 3220, Australia; [2]Deakin University, Geelong, VIC 3220, Australia; [3]Barwon Health, University Hospital Geelong, Geelong, VIC 3220 Australia

* Correspondence: soren.alexandersen@deakin.edu.au; Tel.: +61-0-342159635

Abstract

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) emerged in China in late December 2019 and has spread worldwide. Coronaviruses are enveloped, positive sense, single-stranded RNA viruses and employ a complicated pattern of virus genome length RNA replication as well as transcription of genome length and leader containing subgenomic RNAs. Although not fully understood, both replication and transcription are thought to take place in so-called double-membrane vesicles in the cytoplasm of infected cells. We here describe detection of SARS-CoV-2 subgenomic RNAs in diagnostic samples up to 17 days after initial detection of infection, and provide a likely explanation not only for extended PCR positivity of such samples, but also for discrepancies in results of different PCR methods described by others. Overall, we present evidence that subgenomic RNAs may not be an indicator of active coronavirus replication/infection, but that these RNAs, similar to the virus genome RNA, may be rather stable, and thus detectable for an extended period, most likely due to their close association with cellular membranes.

Introduction

Human coronavirus disease 2019 (COVID-19) emerged in late December 2019 in Wuhan, Hubei Province, China [1,2] and a novel betacoronavirus, subsequently named severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), shown to be the cause. This virus could rather easily transmit from person to person and rapidly spread worldwide [3,4]. SARS-CoV-2 belongs to the Order *Nidovirales*, Family *Coronaviridae*, Subfamily *Orthocoronavirinae*, Genus *Betacoronavirus*, Subgenus *Sarbecovirus*, Species *Severe acute respiratory syndrome-related coronavirus* and Individuum SARS-CoV-2 with the addition of the strain/sequence, e.g. SARS-CoV-2 Wuhan-Hu-1 as the reference strain [5].

Similar to other coronaviruses, SARS-CoV-2 is an enveloped, positive sense, single stranded RNA virus with a genome of nearly 30,000 nucleotides [6]. After having entered the host cell, replication of coronaviruses initially involves generation of a complementary negative sense genome length RNA for amplification of plus strand virus genome RNA as well as transcription of a series of plus strand subgenomic RNAs all with a common leader joined to gene sequences in the 3'-end of the virus genome. Both virus replication and transcription are thought to be facilitated by a virus replication complex derived from virus proteins encoded in the 5' two thirds of the virus genome (termed Open Reading Frame (Orf) 1a and 1b) by a minus 1 ribosomal frameshift between Orf1a and 1b, and proteolytic processing of the generated polyproteins, and translated from the full length plus sense virus genome RNA. As mentioned, a set of subgenomic RNAs are also generated, most likely from a complex mechanism involving paused negative sense RNA synthesis leading to a nested set of negative sense RNAs from the 3'end of the virus genome joined to a common 5'-leader sequence of approximately 70 nucleotides [7]. These nested negative sense RNAs in turn serve as templates for generation of plus strands thus able to serve as a nested set of virus mRNAs for translation of specific proteins from the 3'-third of the virus genome [7]. These subgenomic mRNAs of SARS-CoV-2 are thought to encode the following virus proteins: structural proteins spike (S), envelope (E), membrane (M) and nucleocapsid protein (N) and several accessory proteins for SARS-CoV-2 thought to include 3a, 6, 7a, 7b, 8 and 10 [8]. The mechanism for generation of these subgenomic mRNAs, are not fully understood [7,9], but thought to be tightly regulated to ensure the optimal ratio of virus proteins and to involve pausing of the virus replication/transcription complex at so-called transcription-regulatory sequences (TRS) located immediately adjacent to open

reading frames for these virus genes [8]. Furthermore, it appears that the expression of the N protein is required for efficient coronavirus subgenomic mRNA transcription [7]. The subcellular site/s of coronavirus RNA replication and transcription in the cytoplasm of infected cells is not fully defined, but thought to involve so-called "double-membrane vesicles" (DMV) in or on which the virus replication complex synthesise the needed double and single stranded full length genomic and subgenomic RNAs [7,9]. While it is still unclear whether this RNA synthesis takes place inside or on the outside of these vesicles, it is thought that the membranes somehow "protect" the synthesised RNA, including double stranded RNA, from host cell recognition and response and also from experimental exposure to RNase [9,10]. In addition, it has been shown that coronavirus cytosolic RNA is protected from so-called "nonsense-mediated decay" (NMD) by the virus N protein and thus are more stable in that environment compared to what would have been expected for non-spliced RNA [11]. Finally, while it was originally thought that coronavirus virions contained subgenomic RNAs in addition to the virus plus strand genomic length RNA, it has now been shown that these subgenomic RNAs do not contain a packaging signal and are not found in highly purified, cellular membrane free, coronavirus virions [12]. However, this further indicates, that unless specific steps to remove cellular membranes are used for sample preparation, such coronavirus RNAs are tightly associated with membrane structures, and interestingly, appear to be found in a consistent ratio of around 1-8 to 1 as compared to virus genomic RNA [13].

As mentioned above, the exact mechanism for generation of coronavirus genomic and subgenomic RNAs are not fully understood, but thought to involve the virus replication complex, the TRS, the N protein and double-membrane vesicles in the cytoplasm of infected cells. Although one study has been published looking at the abundance of subgenomic RNAs for SARS-Cov-2, that study employed virus culture in Vero cells [8]. That study indicated that while the predicted spike (S; Orf2), Orf3a, envelope (E; Orf4), membrane (M; Orf5), Orf6, Orf7a and nucleocapsid protein (N; Orf9) subgenomic RNAs were found at high levels in cell culture, only low levels of the Orf7b subgenomic RNA was detected and the Orf10 subgenomic RNA was detected at extremely low level (1 read detected, corresponding to only 0.000009% of reads analysed) [8]. This far, very little has been published in regards to the presence of SARS-CoV-2 subgenomic RNAs in samples from infected people. As single study by Wolfe et al [14], looked specifically for the presence of the E gene subgenomic RNA by a specific PCR and took the presence of subgenomic RNA as an indication of active virus infection/transcription. That study could detect E gene subgenomic RNA at a level of only 0.4% of virus genome RNA,

in sputum samples from day 4-9 of infection, but only up to day 5 in throat swab samples [14]. However, while that study assumed a correlation between the presence of the subgenomic E mRNA and active virus replication/transcription and thus active infection, this assumption may not be accurate considering what has been mentioned above about the membrane associated nature of coronavirus RNA and their stability/protection from the host cell response and from RNases.

We here describe the detection of SARS-CoV-2 subgenomic RNAs in routine diagnostic oropharyngeal/nasopharyngeal swabs subjected to next generation sequencing (NGS) and extend the study to also look at subgenomic RNAs being present in selected read archives selected from the NCBI Sequence Read Archive. We show that such subgenomic RNAs are present in most samples, but that the overall, and individual, subgenomic RNA abundance varies among samples and may be related to stage of infection and, importantly, more related to how samples were taken and treated before testing/sequencing. In turn, this finding may help explain the extended period of PCR positivity observed in infected individuals and also, at least in part, explain conflicting findings around "reinfection" as well as discrepancies among diagnostic PCRs detecting targets in different parts of the SARS-CoV-2 genome [15,16].

Results

Detection, classification and abundance of subgenomic RNAs in SARS-CoV-2 positive samples

As indicated in the Materials and Methods, in our analysis of subgenomic RNAs we have included 12 SARS-CoV-2 positive samples of which two were amplified using two different polymerases for a total of 14 positive samples included in our study together with a virus-negative sample and a few samples amplified with an inflammation panel as an additional negative control for potential unspecific mapping of reads. As also mentioned, manual inspection of reads indicated the presence of subgenomic RNAs and mapping against a reference (Supplementary Information S1 [file: Wuhan-Hu-1-NC_045512-21500-and-subgenomics-SA4.fasta]) designed to specifically map the 10 potential subgenomic RNAs, indicated the presence of subgenomic RNAs in all SARS-CoV-2 positive samples while no reads were found in negative or control samples. Interestingly, the abundance of different subgenomic RNAs and the overall number of these subgenomic RNAs varied widely among the 14 samples as indicated in more details below and summarised in Table 1. Overall, of the

56 million NGS reads generated from the 14 virus-positive samples, nearly 800,000 reads mapped to one of the 10 SARS-CoV-2 subgenomic RNAs (Table 1A). However, no reads mapped to the tentative Orf10/15 RNA and only 5 reads were mapped to the tentative Orf7b RNA. In contrast, reads were mapped to the other 8 subgenomic RNAs, and although it differed among samples, S (Spike) were consistently mapped at a low level followed in increasing order by subgenomic RNAs for Orf3A, E, M and Orf6. Orf7a, Orf8 and N were mapped in the highest abundance although, as mentioned, this was not consistent for all samples (Table 1B and Figures 1-2). As indicated above, the abundance, although overall more or less as expected based on assumed subgenomic RNA abundance [7-9,13,17], differed widely among samples, most likely depending on sample quality and overall virus/subgenomic RNA abundance. However, comparing 2 samples amplified with two different polymerases generated a somewhat similar picture although overall abundance differed (Supplementary Figures S1-S2; samples 33, 34 compared with samples 37, 38, noting that sample 34 is relatively poor quality, see below) and comparing different samples with high quality reads and high virus coverage did also, although with some variability from sample to sample, generate a somewhat comparable pattern (Supplementary Fig. S3; samples 37, 38, 60, 61 and 66). However, looking at sample quality, as determined by overall average read length of a given sample mapped to the full virus genome (Table 1), strongly indicated that sample quality/read length influenced levels of subgenomic RNAs detected, likely due to these subgenomic RNA amplicons incidentally being shorter than most of the virus genome amplicons (Supplementary Table S1; ThermoFisher SARS-CoV-2 Ampliseq panel include 237 virus amplicons ranging from 125-275 nucleotides in size if amplifying virus genomic RNA). Consequently, comparing samples with average read lengths of 115 nucleotides or less with samples having an average read length of 150 nucleotides or more (Table 1), clearly shows the preference for amplification of subgenomic RNAs from short read length samples (Fig 3 and Table 1, comparing good quality samples 35 and 36 with relatively poor quality samples 62, 63 and 64). To look further into this, we analysed the mapping results of two samples already known to be of poor quality, having been suspended in water rather than PBS/transport medium before coming to our laboratory. Although these two samples had a low Ct (high virus load) in the PCR, the NGS generated mostly very short reads (Sample 65 and 68, Table 1) and had a different pattern with a very high abundance of subgenomic RNAs dominated by the Orf7a subgenomic amplicon (Supplementary Fig. S4, samples 65 and 68). As mentioned, this is most likely due to this amplicon being short (sequence length between leader sequence forward primer and nearest pool 2 reverse primer of only 85 nucleotides, although most other subgenomic amplicons would also be expected to be

short (supplementary Table S1). Our sample set analysed here also included samples from two individuals sampled 11-17 days apart and representing early and late infection; sample 37 early and sample 60 late and sample 62 even later (poor quality) as well as sample 38 early and sample 61 late as well as sample 63 (poor quality) taken in between samples 38 and sample 61. As can be seen in the comparison of these samples (Supplementary Fig. S5), the levels of subgenomic RNAs detected are higher in the late infection samples than in the early ones. Although this may possibly indicate a rather long period of virus replication/transcription, we believe it to be more likely due to coronavirus membrane-associated RNAs being partly, albeit not fully, protected from host and environmental degradation and that samples with partly degraded RNA, represented as shorter average read lengths, have some short subgenomic RNA amplicons preferentially amplified (Table 1 and Fig. 4), and thus, as mentioned, more consistent with such samples mainly containing partly degraded virus genomic as well as subgenomic RNAs.

Detection of subgenomic RNAs mapped to the virus genome by filtering reads containing the partial leader sequence

To validate our results detailed above, we looked at the NGS reads to find likely subgenomic RNAs already mapped to the virus genome but filtering so only reads containing part of the leader sequence were included and then look at where these reads had been mapped. Between 8 and 256,123 reads containing the leader sequence were found in our positive samples while none was detected in the negative sample 67 (Supplementary Table S2). Reads were mapped to the location of the TRS of the 10 known subgenomic RNAs, however, only barcodes 37, 38 and 66 possessed reads, in a low number, mapping to the start of ORF7b, and only barcode 66 contained any reads mapping to the start of ORF10 (16 reads or 0.001% of total reads in this sample). The percentage of the mapped reads with a leader sequence to the corresponding ORF in the SARS-CoV-2 genome are shown in Supplementary Fig. S6. In all samples except sample 64 and 66, over 50% of the reads with a leader mapped to the start of a known subgenomic RNA. In samples 37, 60, 61, 62, 63, 65 and 68, over 75% of reads containing the reader sequence were mapped to subgenomic RNAs. While, the percentages varied among the samples, the three subgenomic RNAs with the highest median number of reads with the leader sequence were the E gene/ORF4 (4.1%), ORF7a (17.4%), ORFF8 (4.3%) and N gene/ORF 9 (10.7%).

The samples with the highest number of reads mapping to cryptic or unknown TRS were the poorer quality samples 34, 64 and 65 and no consistent pattern was observed in the mapping of reads with the leader sequence to any individual unrecognised TRS site.

Searching the NCBI SRA to identify reads mapping to SARS-CoV-2 subgenomic RNAs

The final step in our analysis included searching the NCBI SRA and selection of a few deposited NGS reads from studies using either the same SARS-CoV-2 Ampliseq panel used by us as well as a few generated by different methods, such as direct RNAseq or even virus RNA targeted capture (virus RNA targeted capture SRX8336076; Ion Torrent RNAseq e.g. SRX8346911 USA Vero cells; Ion Torrent Ampliseq e.g. SRX8340472 India; SRX8155154 Tiger USA; MinIon eg ERX4009132 Spain; GridIon SRX788484 SARS-CoV-2/Australia/VIC01/2020, RNAseq oligo-dT selection; Illumina e.g. SRX7777164 Vero cells and SRX7777160 swab USA random PCR). For these selected SRA studies, although not abundant for all of them, reads representing subgenomic RNAs rather than virus genomic RNA could be found by simple analysis using e.g. BlastN. Again, as in our own data, we detected no or very little subgenomic RNA of Orf7b and no evidence for Orf10/15 subgenomic RNA.

To look at this in more detail, we downloaded a selection of SRA's generated from different sample types, on different sequencing platforms (Illumina, Ion Torrent and Nanopore) and employing different library amplification strategies (Amplicon based, RNA-Seq and random sequencing). Reads covering most of the SARS-CoV-2 genome were present in all archives, except for SRR11454612 which had low coverage over the SARS-CoV-2 genome when mapped to reference MN908947.3. Reads belonging to subgenomic RNA could be identified in all samples except that sample (SRR11454612), representing RNAseq on a sputum sample from an infected human (Supplementary Table S3 and S4). The two selected Ion Torrent Ampliseq SRA's (SRR11810731 and SRR11810737) produced the highest number of subgenomic reads, followed by an RNA-Seq experiment performed in cell culture using a Nanopore platform (SRR11267570). The selected RNA-Seq experiments performed on clinical samples, typically generated very low levels of reads mapping to the leader sequence and consequently to the virus genome. The Artic network primers also detected subgenomic reads in virus culture experiments (ERR4157962 and ERR4157960).

The proportion of reads mapped to the 5' UTR, the subgenomic RNAs and to unknown/unrecognised locations are shown in Supplementary Fig S7. The subgenomic RNA's with the highest median number of reads mapped in the SRA's were the N gene (10.6%) and

ORF7a (7.1%), followed by ORF3a (1.7%) and M gene (1.2%). The subgenomic S gene and ORF6 were typically low (0.1% and 0.9% respectively). No reads were mapped to the subgenomic ORF10 in any sample, and only sample SRR11267570 and SRR11810737 had any reads mapped to the subgenomic ORF7b (0.2-0.3% of reads having the leader sequence).

Discussion

We here describe the specific detection and mapping of SARS-CoV-2 leader-containing subgenomic RNAs in routine diagnostic oropharyngeal/nasopharyngeal swabs subjected to next generation sequencing (NGS). We present results from two different approaches to map subgenomic RNAs and extend the analysis to also look at subgenomic RNAs being present in selected read archives from the NCBI Sequence Read Archive. Aided by the current understanding of the cell biology of coronavirus infections (please see Introduction for details and specific references), in particular the known association of virus RNAs with, and at least in partial protection by, cellular double-membrane vesicles (DMVs), we present an integrated interpretation of our results based on detailed analysis of relative abundance of the different subgenomic RNAs in samples collected early and late in infection, samples of different quality and in samples subjected to cell/membrane lysis in water. Our integrated interpretation of the results overall, is that both virion and subgenomic RNAs are most likely rather stable *in vivo* and that detection of subgenomic RNAs in clinical samples, importantly, do not necessarily signify active virus replication/transcription, but instead is due to such RNAs being part of double-membrane vesicles and thus relatively stable compared to e.g. cellular mRNA. Although subgenomic RNAs are thought to be particular abundant in early infection, reportedly up to 70 times more abundant than virus genomic RNA at the peak of RNA transcription, that may only be observed in cell culture, as it occurs as early as 6-8 hours after infection [18] and after that is found at a more steady state level of a ratio of 1:1-1:8 [13].

Our specific detection of subgenomic RNAs in clinical samples and the indication that these RNAs are rather stable and most likely found in, and protected by, membrane structures and that their relative abundance may be more related to sample quality and storage before testing/sequencing than to the actual stage of infection, may help explain not only the extended period of PCR positivity observed in infected individuals, but also, at least in part, help explain conflicting findings around “reinfection” as well as discrepancies among diagnostic PCRs detecting targets in different parts of the SARS-CoV-2 genome [15,16]. For example, our analysis

indicated that some subgenomic RNAs may be more abundantly detected in poor samples, possibly because of partly degraded RNA in such samples and the increased ability of PCR, including most diagnostic PCRs as well as NGS employing various amplification steps, to amplify short targets. We believe that this is particular evident in the Ampliseq and other multi primer-pair sequencing strategies, because partly degraded targets will increase amplification of short amplicons as the "competition" with longer target amplicons is decreased. This notion is supported by our findings in inadvertently poor samples, in our sample material represented by two diagnostic samples that for unrelated reasons had been suspended in water rather than in PBS or transport medium as for our other samples. These particular two samples (sample 65 and 68, Table 1), in which cells and membrane vesicles were almost certainly disrupted [19], and thus any coronavirus RNAs exposed to the environment and likely to RNases, were dominated by short reads, i.e. short amplicons, and strikingly, with a very high abundance of reads mapped to subgenomic RNAs, in particular the Orf7a RNA. Clearly, original sample abundance cannot change just because cellular membranes are lysed by the hypotonic treatment, so the observed increase in these reads may almost certainly be caused by preferred amplification of shorter or more efficient amplicons in such samples. Nevertheless, taken together, it shows that these subgenomic RNAs are present in our diagnostic samples, but that relative ratios are highly dependent on sample handling.

Our mapping of specific subgenomic RNAs indicated that samples had no or very low levels of subgenomic RNA specific for Orf7b and that a subgenomic RNA specific for Orf10/15 was absent. This is consistent with what has been described for SARS-CoV-2 in cell culture [8]. A recent publication [20] did report conflicting results for diagnostic samples; however, we believe this to be due to misinterpretation of the results of that study, as NGS reads of only 75 nucleotides in length were mapped to the SARS-CoV-2 genome and assigned to subgenomic RNA regions, completely ignoring the fact that only leader containing reads would be specific for any given subgenomic RNA. Consequently, their interpretation of high levels of e.g. Orf10 is most likely incorrect, and their data simply a reflection of a higher coverage of the 3'-end of the genome which is expected as all subgenomic RNAs extend to the 3'-end of the virus genome. However, the apparent coverage observed in that study is approximately 8 times higher towards the 3'-end, consistent with what would be expected in diagnostic material with virus RNAs protected by cellular membranes [10,13] and consistent with the overall abundance of subgenomic RNAs observed in our study. Nevertheless, we emphasise that these studies are

not comparable, as we believe the study described by Zhang et al. [20] is not mapping subgenomic RNAs but simply reports coverage for the different parts of the virus genome.

In conclusion, by combining knowledge of general coronavirus cell biology and replication/transcription with careful mapping of SARS-CoV-2 subgenomic RNAs in clinical samples taken at different times of infection and of different quality, we present information that helps understand prolonged and sometimes inconsistent PCR-positivity and may pave the way for development of better diagnostic PCRs and NGS strategies to define active SARS-CoV-2 infection as opposed to extended presence of what most likely represent highly stable virus genomic and subgenomic RNAs present in, and at least in part protected by, cellular membranes, so-called double-membrane vesicles (DMVs). Our findings are likely to be relevant also for other coronaviruses and possibly also other viruses in the Order *Nidovirales*. That coronaviruses, and their RNA, may be extremely resistant when part of a membrane matrix is well known, and was demonstrated for example when another coronavirus, the porcine epidemic diarrhea virus, entered and infected pigs in Canada in early 2014 by feed containing spray-dried porcine plasma [21,22]. Detergent treatment and ultracentrifugation indicated that this coronavirus RNA was initially bound to membranes, but could be pelleted by ultracentrifugation after detergent release and other studies, e.g. our own using yet another coronavirus, the avian infectious bronchitis virus, further support that detergent treatment will release the coronavirus RNA and make it susceptible to nuclease degradation while passing through 0.8, but not 0.45 micron filtres, support the fact that the majority of such coronavirus RNA is membrane bound [23]. Consequently, we believe that the methods described here to detect and look at relative abundance of SARS-CoV-2 RNAs in clinical samples together with insights in what is known about coronavirus cell biology overall, will have general interest and applicability not only for SARS-CoV-2, but also for other coronaviruses and related viruses.

Methods

Samples

We here describe extended analysis of samples already subjected to next generation sequencing (NGS) at the Geelong Centre for Emerging Infectious Diseases (GCEID). The samples and the results of SARS-CoV-2 genomic consensus sequencing have been described previously (submitted to the Journal Viruses). That study included combined nasopharyngeal and oropharyngeal swab samples collected from individuals in the region of Greater Geelong,

Victoria, Australia between the 28th of January to the 14th of April 2020. The study included NGS of 11 SARS-CoV-2 PCR-positive samples and 1 negative sample as control. The 11 PCR-positive samples were obtained from 7 individuals as NGS was done on samples taken at three time points from two individuals to monitor their infection (submitted to the Journal Viruses). For the studies described here we added one additional positive sample collected on the 24th of April 2020. This sample was subjected to NGS in exactly the same way as described for the samples mentioned above (details of methods given in (submitted to the Journal Viruses)). For the study described here we also included additional negative controls, including some samples being amplified with a panel designed to amplify human transcripts involved in inflammation (see Results). The study complied with all relevant ethical regulations and has been approved by the Barwon Health Human Research Ethics Committee (Ref HREC 20/56) and participants provided opt-out consent.

In addition to analysis of the NGS reads obtained from the samples mentioned above, we also searched the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) and used selected SRA studies to support the findings from our own samples.

Nucleic acid extraction, cDNA synthesis and SARS-CoV-2 Ampliseq NGS at GCEID

Nucleic acid extraction and cDNA synthesis was performed as described including heating extracted nucleic acids at 70°C for 5 minutes and rapid cooling on ice before cDNA synthesis using SuperScript™ VILO™ Master Mix (Thermofisher Scientific, Victoria, Australia) as per manufacturers' instructions and described previously (submitted to the Journal Viruses) [24]. Prepared cDNA samples were then amplified using the Ion Ampliseq™ Library Kit 2.0 (Thermofisher Scientific, Victoria, Australia) as described earlier [25,26] and a commercially available SARS-CoV-2 Ampliseq panel kindly provided by Thermofisher Scientific, Victoria, Australia. In addition, two of the samples that yielded a low virus coverage by this method, were amplified separately using essentially the same method, but with the Ampliseq Hi Fi mix replaced with Amplitaq Gold 360 Master mix for the amplification step. Amplification was done for either 21, 27 or 35 cycles depending on the estimated virus load in the samples as described previously (submitted to the Journal Viruses) and libraries prepared and run on Ion Torrent 530 chips on an Ion S5 XL genetic sequencer (Thermofisher Scientific) at a concentration of 50pM as per the manufacturer's protocols and as described previously (submitted to the Journal Viruses) [24,27]. As described previously (submitted to the Journal Viruses), generated sequence reads were then mapped to a SARS-CoV-2 reference genome

using the TMAP software included in the Torrent Suite 5.10.1 [28], and virus genomic consensus sequences generated using additional Torrent Suite plugins supplied by Thermofisher Scientific, and visualized in Integrative Genomic Viewer [29] (IGV 2.6.3) (Broad Institute, Cambridge, MA, USA). Near complete and partial SARS-CoV-2 genomes were aligned using Clustal-W [30] in MEGA 7 software [31] and near full length sequences submitted to the Global Initiative on Sharing All influenza Database (GISAID) [32,33] (https://www.gisaid.org/) as described in our previous study (submitted to the Journal Viruses).

Analyses for the detection of SARS-CoV-2 subgenomic mRNAs in the NGS reads

Although the SARS-CoV-2 Ampliseq panel used for the NGS has been designed to generate near full length SARS-CoV-2 genomic sequences, it uses simultaneous amplification of sample cDNA with a total of 242 primer pairs of which 237 primer pairs cover the near full genome of SARS-CoV-2 and an additional 5 amplicons targeting cellular genes in two primer pools (submitted to the Journal Viruses). Close inspection of all primers included in the panel, indicated that two of the forward primers (specifically the first forward primer in each of primer pool 1 and 2, see Thermofisher Scientific for details) have their 3'-end at SARS-CoV-2 (NCBI Accession MN908947.3) [2] nucleotide 42 and 52, respectively, and consequently have a perfect match to a sequence included in the SARS-CoV-2 leader sequence with an estimated 27 or 17 nucleotides downstream of these primers also being part of the leader sequence [7,8,14,17].

Consequently, we concluded that the Ampliseq panel used here would potentially also amplify SARS-CoV-2 subgenomic RNAs by amplification from these two forward primers together with the closest downstream primer included in the same Ampliseq pool. Although this was not evident when assembling full virus genome sequences (submitted to the Journal Viruses), a close inspection of reads around expected subgenomic RNA Transcription Regulatory Sites (TRS) [17] indicated that a significant number of NGS reads may have been amplified from subgenomic RNAs rather than from virus genomic RNA. To analyse this in more detail, we first assembled an exploratory composite reference for remapping using the Torrent Suite T-map reanalyse function. This initial assembled reference consisted of a composite reference with one sequence containing the first 21500 nucleotides of the SARS-CoV-2 genome used for the initial assembly (NCBI Accession MN908947.3) [2] to map reads most likely corresponding to the virus genome while we in addition, assembled 10 tentative subgenomic RNA sequences containing 28 nucleotides from the 3'end of the leader sequence (of which the first 11 nucleotides would be from the forward primer from primer pool 2, if not enzymatically

removed by the NGS process) and this leader then followed by the assumed TRS and gene specific sequence for the next 72 nucleotides. Consequently, this reference contained the first 21500 nucleotides of the virus genome as well as 10 composite references corresponding to the assumed 5'-end of the 10 potential subgenomic RNAs; S, Orf3a, E, M, Orf6, Orf7a, Orf7b, Orf8, N and Orf10/15 [7,8,17]. This initial analysis indicated that this was an efficient way of mapping reads corresponding to subgenomic RNAs, and for our final analysis we updated the subgenomic RNA sequences in this composite reference to include the full leader sequence from nucleotide 1-69 and extended the gene specific sequences to ensure that they would include a reverse primer from each primer pool without extending into the next specific gene sequence. This final composite reference used for mapping then included the first 21500 nucleotides of the SARS-CoV-2 genome and the 10 subgenomic RNA specific sequences, each including the leader and gene specific sequences and having a length of 233-364 nucleotides (Supplementary Information S1 [file: Wuhan-Hu-1-NC_045512-21500-and-subgenomics-SA4.fasta]). Mapped reads were visualised in IGV at a minimal alignment score of 60 and a mapping quality (MAPQ) of 84.

The next step in the analysis was to look at whether such subgenomic RNAs could be detected by a somewhat different type of analysis based on using reads already mapped to the full virus sequence and then filter these to only look at reads containing part of the leader sequence. This type of analysis should give an unbiased view as to where reads containing the leader may be located on the genome, whether those sites correspond to the assumed position of the genomic leader and proposed TRS of the leader-containing subgenomic RNAs, whether the abundance somewhat correspond to that found by the method mentioned above and finally, whether any additional subgenomic RNAs or cryptic TRS sites may be detected. Reads within the mapped reads BAM files were filtered on whether they had a MAPQ of 32 or higher and contained the partial leader sequence GTAGATCTGTTCTCT, using a custom script written in BASH 4.4.20 and AWK 4.1.4 (GNU project, www.gnu.org) using samtools 1.7 [34]. This sequence corresponded to nucleotides 52-67 within the SARS-CoV-2 leader sequence in GenBank sequence MN908947.3 [2]. Reads with this sequence immediately upstream from the mapped region of the read, or within the 5' end of the mapped region, were retained. The script then generated a spreadsheet giving the nucleotide position of where the leader sequence of each read finished in relation to the reference genome. These reads were then visualized in IGV. To assign the reads to the corresponding subgenomic RNA, the reads were grouped by the nucleotide position at the end of the leader sequence and tallied in Excel. Based on the end

position of the leader mapped to the reference genome, each read was assigned to the corresponding subgenomic TRS. Typically, the leader sequence sat within a soft-clipped portion of each read, although depending on the reference sequence, the Ion Torrent TMAP algorithm did occasionally include the start of the leader sequence within the mapped portion of some reads, and at times included spurious insertions or deletions within this section of the mapping in its attempt to map the leader to the reference. Therefore, any read with the leader ending within 10nt of the start of the known subgenomic TRS sequences were assigned to the respective TRS. Some reads did not map to any known TRS, and these were assigned to an "Unknown TRS".

The final step in our analysis included searching the NCBI SRA from where we selected a few deposited NGS reads from studies using the same SARS-CoV-2 Ampliseq panel used by us and in addition selected a few generated by different methods. SAM files from 15 SRA accessions were downloaded with the NCBI SRAtoolkit sam-dump 2.8.2 and mapped to the Ampliseq SARS-CoV-2 reference MN908947.3 [2] using NCBI Magic-BLAST 1.3.0 with a minimum alignment score of 50 and percentage identity of 90% or higher. The script and analysis method to identify reads containing the leader sequence described above was used on the Magic-BLAST mapped SAM files for each of the SRA archives, and the number of reads corresponding to the start of each subgenomic RNA tallied.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The sequence reads and the subgenomic RNA mapping file for our SARS-CoV-2 positive samples reported here have been deposited in the NCBI Sequence Read Archive (SRA) under SRA accession: PRJNA636225. All other data supporting the findings of this manuscript are available in the Supplementary Information files or from the corresponding author upon reasonable request. A reporting summary for this Article is available as a Supplementary Information file.

References

1 Hui, D. S. *et al.* The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health- the latest 2019 novel coronavirus outbreak in Wuhan, China. *Int J Infect Dis* **91**, 264-266 (2020).

2 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269, doi:10.1038/s41586-020-2008-3 (2020).

3 Chan, J. F.-W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet* **395**, 514-523 (2020).

4 Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* (2020).

5 Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* **5**, 536-544, doi:10.1038/s41564-020-0695-z (2020).

6 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).

7 Sola, I., Almazan, F., Zuniga, S. & Enjuanes, L. Continuous and discontinuous RNA synthesis in coronaviruses. *Annual review of virology* **2**, 265-288 (2015).

8 Kim, D. *et al.* The architecture of SARS-CoV-2 transcriptome. *Cell* (2020).

9 Snijder, E. J. *et al.* A unifying structural and functional model of the coronavirus replication organelle: tracking down RNA synthesis. *BioRxiv* (2020).

10 Hofmann, M. A., Sethna, P. B. & Brian, D. A. Bovine coronavirus mRNA replication continues throughout persistent infection in cell culture. *Journal of Virology* **64**, 4108-4114 (1990).

11 Wada, M., Lokugamage, K. G., Nakagawa, K., Narayanan, K. & Makino, S. Interplay between coronavirus, a cytoplasmic RNA virus, and nonsense-mediated mRNA decay pathway. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E10157-E10166, doi:10.1073/pnas.1811675115 (2018).

12 Escors, D., Izeta, A., Capiscol, C. & Enjuanes, L. Transmissible gastroenteritis coronavirus packaging signal is located at the 5' end of the virus genome. *J Virol* **77**, 7890-7902, doi:10.1128/jvi.77.14.7890-7902.2003 (2003).

13 Zhao, X., Shaw, K. & Cavanagh, D. Presence of subgenomic mRNAs in virions of coronavirus IBV. *Virology* **196**, 172-178 (1993).

14 Wölfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature*, 1-5 (2020).

15 Tom, M. R. & Mina, M. J. To Interpret the SARS-CoV-2 Test, Consider the Cycle Threshold Value. *Clinical Infectious Diseases*, doi:10.1093/cid/ciaa619 (2020).

16 Xiao, A. T., Tong, Y. X. & Zhang, S. Profile of RT-PCR for SARS-CoV-2: a preliminary study from 56 COVID-19 patients. *Clin Infect Dis*, doi:10.1093/cid/ciaa460 (2020).

17 Pascual, M. R. Coronavirus SARS-CoV-2: Analysis of subgenomic mRNA transcription, 3CLpro and PL2pro protease cleavage sites and protein synthesis. *arXiv preprint arXiv:2004.00746* (2020).

18 Wu, H.-Y. & Brian, D. A. Subgenomic messenger RNA amplification in coronaviruses. *Proceedings of the National Academy of Sciences* **107**, 12257-12262, doi:10.1073/pnas.1000378107 (2010).

19 Lew, V. L., Hockaday, A., Freeman, C. J. & Bookchin, R. M. Mechanism of spontaneous inside-out vesiculation of red cell membranes. *J Cell Biol* **106**, 1893-1901, doi:10.1083/jcb.106.6.1893 (1988).
20 Zhang, H. *et al.* Metatranscriptomic Characterization of COVID-19 Identified A Host Transcriptional Classifier Associated With Immune Signaling. *Clinical Infectious Diseases*, doi:10.1093/cid/ciaa663 (2020).
21 Ojkic, D. *et al.* The first case of porcine epidemic diarrhea in Canada. *Can Vet J* **56**, 149-152 (2015).
22 Pasick, J. *et al.* Investigation into the role of potentially contaminated feed as a source of the first-detected outbreaks of porcine epidemic diarrhea in Canada. *Transbound Emerg Dis* **61**, 397-410, doi:10.1111/tbed.12269 (2014).
23 Vibin, J. *et al.* Metagenomics detection and characterisation of viruses in faecal samples from Australian wild birds. *Sci Rep* **8**, 8686, doi:10.1038/s41598-018-26851-1 (2018).
24 Bhatta, T. R., Chamings, A., Vibin, J. & Alexandersen, S. Detection and characterisation of canine astrovirus, canine parvovirus and canine papillomavirus in puppies using next generation sequencing. *Sci. Rep.* **9**, 1-10 (2019).
25 Chamings, A. *et al.* Evolutionary analysis of human parechovirus type 3 and clinical outcomes of infection during the 2017–18 Australian epidemic. *Scientific reports* **9**, 1-9 (2019).
26 Alexandersen, S., Nelson, T. M., Hodge, J. & Druce, J. Evolutionary and network analysis of virus sequences from infants infected with an Australian recombinant strain of human parechovirus type 3. *Sci. Rep.* **7**, 1-12 (2017).
27 Bhatta, T. R., Chamings, A., Vibin, J., Klaassen, M. & Alexandersen, S. Detection of a Reassortant H9N2 Avian Influenza Virus with Intercontinental Gene Segments in a Resident Australian Chestnut Teal. *Viruses* **12**, 88 (2020).
28 Caboche, S., Audebert, C., Lemoine, Y. & Hot, D. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC genomics* **15**, 264 (2014).
29 Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178-192 (2013).
30 Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *bioinformatics* **23**, 2947-2948 (2007).
31 Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular biology and evolution* **33**, 1870-1874 (2016).
32 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33-46 (2017).
33 Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **22** (2017).
34 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

Acknowledgments: We acknowledge clinical and laboratory diagnostic staff for providing samples and for doing the initial diagnostic SARS-CoV-2 testing. We gratefully acknowledge Thermofisher Scientific, Victoria, Australia, for supplying the Ampliseq panel used. We also acknowledge Jason Hodge, laboratory manager of the GCEID laboratory for his technical input. Finally, we gratefully acknowledge the authors and originating and submitting laboratories for the sequence read archives we have used in this study (Supplementary Table S3).

Author Contributions: S.A. initiated the study, coordinated all work carried out at GCEID and did the initial mapping of SARS-CoV-2 subgenomic RNAs. A.C. did the script-based mapping and details in regards to selected SRA mapping. T.R.B. provided sample information and participated in the NGS and initial analysis to look for subgenomic RNAs. S.A. drafted the initial manuscript with inputs from A.C. and T.R.B. All authors contributed to the final submitted version. All authors have read and agreed to the final version of the manuscript.

Funding: This research was funded by Deakin University, Barwon Health and CSIRO and from the National Health and Medical Research Council (NHMRC) equipment grant number GNT9000413 to S.A.

Conflicts of Interest: The authors declare no competing or conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Tables and Figure Legends

Table 1: Details of the sample number, average read length, number of reads, average virus coverage and number of reads mapped to each subgenomic SARS-CoV-2 RNA are shown (A). The number of reads given in (B) is adjusted so they represent reads normalised to a total of 5 million reads for each sample for easier comparison. These normalised numbers are used for the Figure1-4 and Supplementary Figures S1-S5.

Table 1A

Sample	Average read length	Number of reads in milions	Average virus coverage	S Orf2	Orf3a	E Orf4	M Orf5	Orf6	Orf7a	Orf7b	Orf8	N Orf9	Orf10
33	153	6	432	15	57	77	99	60	261	0	42	174	0
34	73	1.2	312	8	29	98	63	58	592	0	250	35	0
35	149	1.6	96	1	17	17	9	14	72	0	17	81	0
36	163	1.3	6	0	4	0	0	1	1	0	1	1	0
37	185	5.2	26840	243	6798	5691	5548	6551	12898	0	1539	33261	0
38	184	3.3	16640	179	2464	1715	1702	1587	4107	0	406	5071	0
60	150	5.2	16640	204	74	11117	304	12972	6902	0	1463	51614	0
61	115	2.3	3088	0	0	16067	283	0	5132	0	0	14442	0
62	100	2.6	288	0	0	0	89	0	19847	0	0	15130	0
63	83	1.3	44	0	0	133	1	972	0	0	0	0	0
64	92	3	556	0	0	0	88	0	9408	0	1052	3966	0
65	84	5.1	2552	2	10	16581	1249	16655	158755	0	11990	1841	0
66	207	14.3	80968	1611	17702	18165	5793	7066	19443	0	807	20156	0
67	76	1.8	0	0	0	0	0	0	0	0	0	0	0
68	84	3.9	3268	13	21	35701	4453	4265	166578	5	3328	942	0
87	126	3	0	0	0	0	0	0	0	0	0	0	0
89	135	4.1	0	0	0	0	0	0	0	0	0	0	0

Table 1B

Sample	Average read length	Number of reads in milions	Average virus coverage per 5 mill reads	S Orf2	Orf3a	E Orf4	M Orf5	Orf6	Orf7a	Orf7b	Orf8	N Orf9	Orf10
33	153	6	360	13	48	64	83	50	218	0	35	145	0
34	73	1.2	1300	33	121	408	263	242	2467	0	1042	146	0
35	149	1.6	300	3	53	53	28	44	225	0	53	253	0
36	163	1.3	23	0	15	0	0	4	4	0	4	4	0
37	185	5.2	25808	234	6537	5472	5335	6299	12402	0	1480	31982	0
38	184	3.3	25212	271	3733	2598	2579	2405	6223	0	615	7683	0
60	150	5.2	16000	196	71	10689	292	12473	6637	0	1407	49629	0
61	115	2.3	6713	0	0	34928	615	0	11157	0	0	31396	0
62	100	2.6	554	0	0	0	171	0	38167	0	0	29096	0
63	83	1.3	169	0	0	512	4	3738	0	0	0	0	0
64	92	3	927	0	0	0	147	0	15680	0	1753	6610	0
65	84	5.1	2502	2	10	16256	1225	16328	155642	0	11755	1805	0
66	207	14.3	28310	563	6190	6351	2026	2471	6798	0	282	7048	0
67	76	1.8	0	0	0	0	0	0	0	0	0	0	0
68	84	3.9	4190	17	27	45771	5709	5468	213562	6	4267	1208	0
87	126	3	0	0	0	0	0	0	0	0	0	0	0
89	135	4.1	0	0	0	0	0	0	0	0	0	0	0

Figure 1. Average coverage per 5 million reads for the samples included in this study and mapped to either the virus genome or to each subgenomic RNA. The sample is shown on the X-axis and the coverage/number of reads on the Y-axis. Reads mapped to the individual subgenomic RNAs in the fasta file used for mapping are indicated with different colours.

Figure 1

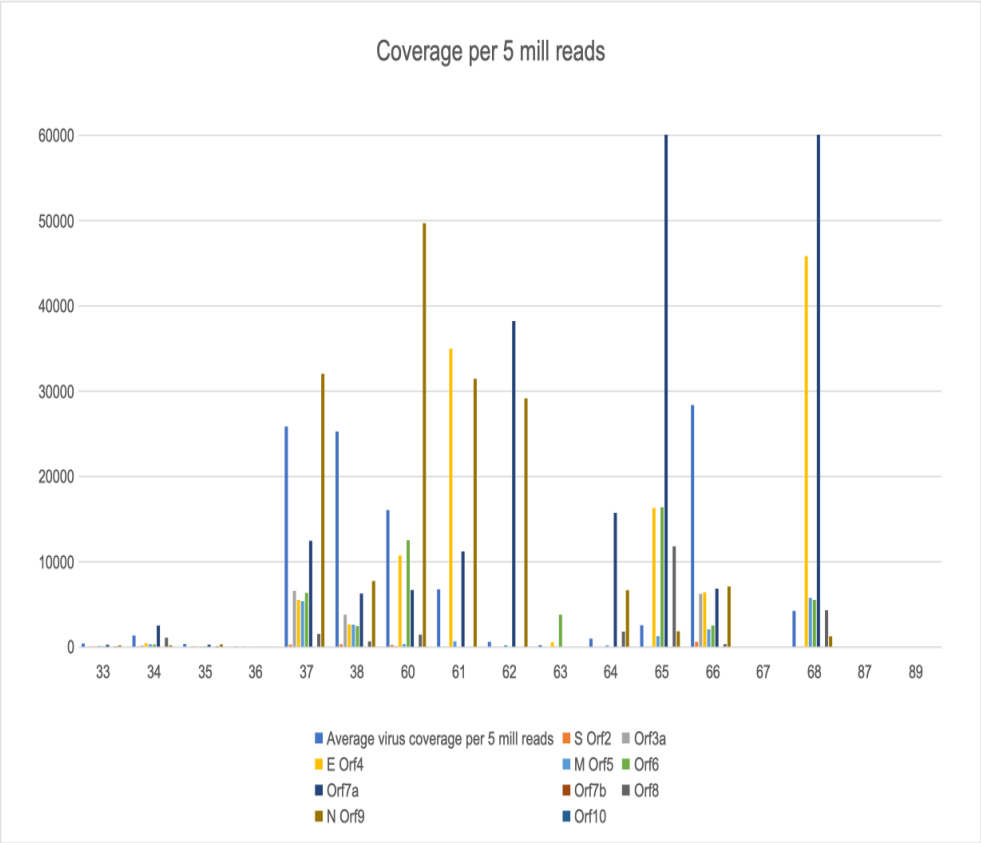


Figure 2. Shows essentially the same data as Fig. 1, but with the different genomic or subgenomic RNAs on the X-axis and the coverage/number of mapped reads per 5 million reads on the Y-axis. Different samples are indicated with different colours.

Figure 2

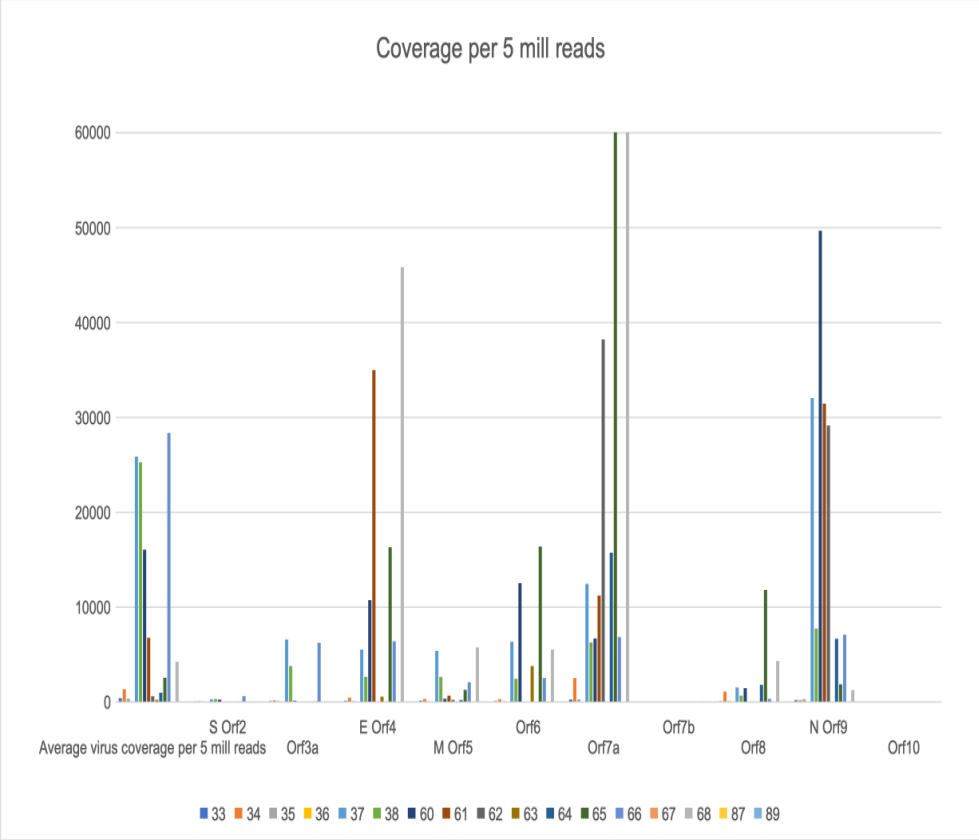


Fig 3. As indicated for Fig. 2, but focusing on samples 35, 36, 62, 63 and 64.

Figure 3

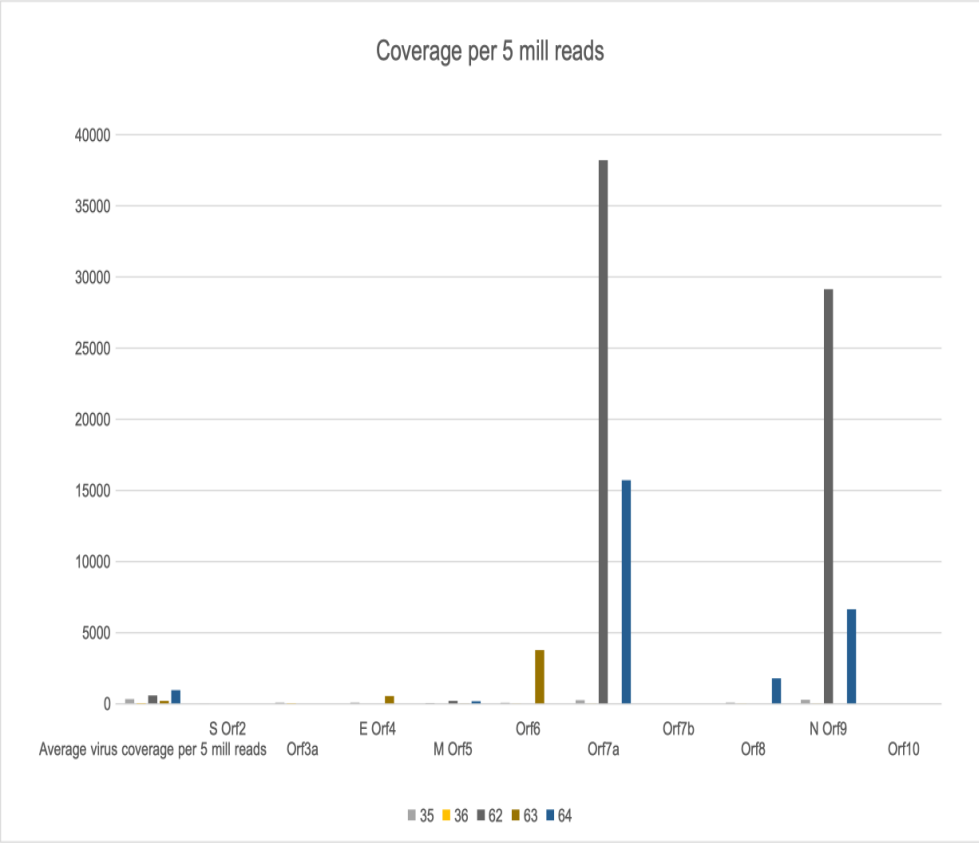
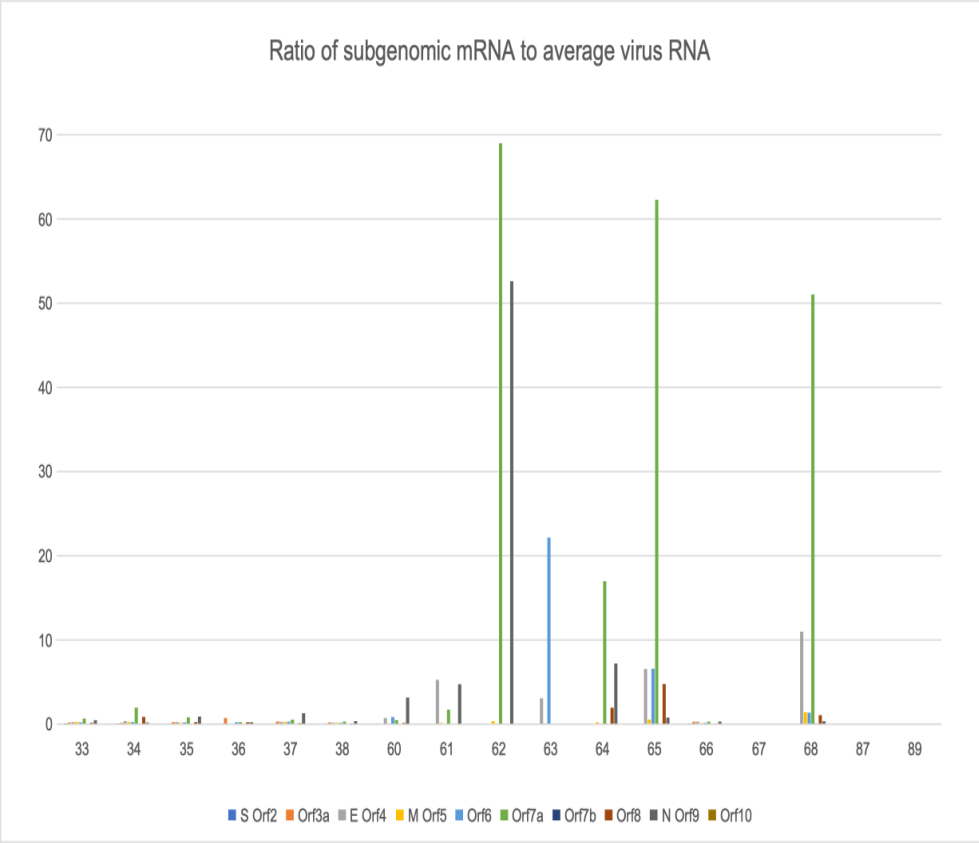


Fig 4. Figure showing the ratio of subgenomic RNA to average virus genomic RNA detected focusing on our finding that samples with shorter average read lengths preferably amplify some subgenomic RNAs.

Figure 4



Supplementary Materials: The following is available online; Supplementary Information file S1, Supplementary Tables S1-S3 and Supplementary Figures S1-S7.