Noncanonical junctions in subgenomic RNAs of SARS-CoV-2 lead to variant open reading frames.

Jason Nomburg[1,2,3], Matthew Meyerson[1,2,4,5*], James A. DeCaprio[1,3,5*]

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston MA
[2]Broad Institute of MIT and Harvard, Cambridge, MA
[3]Harvard Program in Virology, Harvard University Graduate School of Arts and Sciences, Boston, MA
[4]Department of Genetics, Harvard Medical School, Boston, MA
[5]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

*Correspondence to:
James A. DeCaprio
james_decaprio@dfci.harvard.edu

Matthew Meyerson
matthew_meyerson@dfci.harvard.edu

Abstract

SARS-CoV-2, a positive-sense RNA virus in the family *Coronaviridae*, has caused the current worldwide pandemic, known as coronavirus disease 2019 or COVID-19. The definition of SARS-CoV-2 open reading frames is a key step in delineating targets for vaccination and treatment for COVID-19. Here, we report an integrative analysis of three independent direct RNA sequencing datasets of the SARS-CoV-2 transcriptome. We find strong evidence for variant open reading frames (ORFs) encoded by SARS-CoV-2 RNA. A variant transcript for the matrix protein (M) lacking its N-terminal transmembrane domain, initiated by a TTG start codon, is produced by a strong transcriptional regulatory sequence (TRS)-mediated junction within the M ORF and represents up to 19% of all M ORFs. Sporadic non-canonical junctions in the spike (S) ORF lead to N-terminal truncations that remove the N-terminal and receptor-binding domains from up to 25% of S ORFs. Surprisingly, nearly all ORFs from ORF1a identified in these transcriptome sequences were variant. These ORFs contain the first 200-800 amino acids of ORF1a and may represent a mechanism to regulate the relative abundance of ORF1a nonstructural proteins. We show there is strong transcriptome and junctional support for variant ORF1a ORFs in independent direct RNA sequencing and short-read RNA sequencing datasets, and further show that up to 1/3 of these ORFs are expected to have C-terminal fusions with downstream genes. Finally, we show that currently unannotated ORFs are abundant in the SARS-CoV-2 transcriptome. Together, these analyses help to elucidate the diverse coding potential of SARS-CoV-2.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) emerged from Wuhan, China in December 2019 and rapidly led to a world-wide pandemic, known as COVID-19 [1]. The initial sequencing report found that this virus maintains 89.1% nucleotide identity to SARS-CoV-1 [1].

Coronaviruses (CoV) including SARS-CoV-2 contain large, ~ 30kb, positive-sense, single-stranded RNA genomes with a unique genome organization structure. The 5' end of the genome contains two large open reading frames (ORFs), ORF1a and ORF1b. ORF1a produces a large polyprotein and ribosomal slippage at an RNA pseudoknot structure and slippery sequence at the end of ORF1a occasionally causes a frameshift and subsequent translation of a joint ORF1a and ORF1b polyprotein [2]. ORF1a and ORF1ab polyproteins contain protease activity that can cleave the polyproteins into a variety of nonstructural proteins (nsp) capable of facilitating viral transcription and replication and modulating host transcription and translation. Following ORF1b, there is an arrangement of ORFs encoding structural and accessory proteins that tend to vary in content and order across the CoV family. While structural proteins are incorporated into emergent virions, accessory proteins are thought to be dispensable for replication *in vitro* but to increase fitness *in vivo* [3-5].

Initial viral translation of ORF1ab can occur directly from incoming viral genomic RNA, generating the ORF1a and ORF1ab polyproteins and initiating infection. However, because the viral genome is so large and downstream ORFs are far from the 5' end of the genome, effective translation of these ORFs requires the generation of subgenomic RNAs [6]. To generate these subgenomic RNAs, it is thought that CoV undergoes a process of discontinuous extension during synthesis of negative-stranded RNA templates. As the virus-encoded RNA-dependent RNA polymerase progresses from the 3' end towards the 5' end of the genome, it encounters transcriptional regulatory sequences (TRS) upstream of each major ORF (TRS body sites). Here, the polymerase can skip to a similar TRS just upstream of a shared leader sequence at the 5' end of the CoV genome (TRS leader). These antisense subgenomic RNAs are then transcribed, resulting in a series of tiered subgenomic RNAs all containing the same 5' leader sequence [6].

Several studies have characterized SARS-CoV-2 subgenomic RNAs using Nanopore direct RNA sequencing (dRNAseq). In contrast to short-read sequencing, which produces read pairs as long as 150bp each, dRNAseq is capable of directly sequencing entire transcripts including the full length of the viral genome (29903 bases). This allows dRNAseq to identify RNA isoforms and variants that may be challenging to deconvolute using Illumina sequencing. Taiaroa et al. identified eight major subgenomic RNAs, as well as a small number of non-canonical junctions between the 5' leader sequence and downstream regions of the SARS-CoV-2 genome [7]. Kim et al. identify a total of 9 distinct subgenomic RNAs, as well as a series of RNAs that contain unexpected recombination events between 5' and 3' sequences [8]. Using short-read DNA nanoball sequencing data, they identified the existence of noncanonical junctions at various sites across the genome. Davidson et al. identified a series of transcripts encoding the nucleocapsid (N) protein, but with distinct small internal deletions, and validated this finding with proteomic evidence [9].

While these existing studies have identified noncanonical RNA junctions and unexpected RNA species, it is unclear if these junctions influence the landscape of open reading frames or if they result in novel ORFs. We utilized the three independent dRNAseq datasets from Taiaroa et al. [7], Kim et al. [8], and Davidson et al. [9], and three short read Illumina datasets from Blanco-Melo et al. [10], to characterize the effect of unexpected RNA junctions on viral ORFs.

Results

Analysis of three independent dRNAseq datasets reveals canonical and noncanonical junctions.

To determine the major RNA species present in SARS-CoV-2-infected cells, we applied a uniform computational approach to three independent dRNAseq datasets from Taiaroa et al., Kim et al., and Davidson et al. [7-9]. Because dRNAseq can have an error rate of greater than 10% [11], we generated "coordinate-derived transcripts" by determining the start, end, and junction coordinates of each raw transcript by mapping to SARS-CoV-2 (NC_045512.2) and retrieved the equivalent genome sequence using these coordinates (**Figure 1A**). This approach is capable of overcoming sequencing errors and can resolve major RNA species but is not capable of resolving small or structurally complex RNA rearrangements.

Analysis of the 5' and 3' locations of junctions separated by at least 1000 bases revealed a majority of junctions consisting of a 5' end originating in the first 100 bases and a 3' end at distinct sites towards the 3' region of the genome (**Figure 1B,C**). These are canonical junctions between the leader sequence and 3' sequences that generate the subgenomic RNAs [6]. These junctions have 5' origins near the leader TRS core sequence located from position 69 to 75, but there is some heterogeneity in the exact 5' position of each junction (**Figure 1C).** In addition, there is an unexpectedly diffuse pattern of noncanonical junctions across the genome (**Figure 1B**), suggesting that there are unexpected RNA species. Had there been only canonical junctions, the only color in this plot would be at defined points in contact with the x-axis (**Figure 1B**).

A major junction point falls within the M open reading frame.

The 3' end of canonical junctions are strongly enriched at distinct sites that are generally immediately 5' to a target gene (**Figure 1D**) and result in the major subgenomic RNAs. S, ORF3a, E, M, ORF7a, ORF8, and N each have junction points at TRS sites just upstream of their ORFs. In contrast, there are very few junctions and no TRS sites immediately upstream of ORF7b and ORF10 suggesting that, if they are expressed, they are likely expressed from other subgenomic RNAs. Interestingly, the closest major junction point to ORF6 is located at a TRS site within the M coding sequence. This indicates the existence of subgenomic RNAs that contain a distinct 3' region of M ("variant M") close to the 5' end of the subgenomic RNA, followed closely by

ORF6. Notably, all three independent dRNAseq datasets support these major junction points.

To independently assess the possibility that junction locations are an artifact from dRNAseq, we conducted a similar junction analysis using short read Illumina datasets from SARS-CoV-2 infected cells generated by Blanco-Melo et al. [10]. This analysis yielded similar results (**Supplemental Figure 1**). Of note, the three dRNAseq datasets and the short-read datasets used four different virus isolates collected in four different continents, suggesting that these findings broadly represent SARS-CoV-2 biology. However, since these studies were performed in cell lines, future studies will be required to carefully assess SARS-CoV-2 transcriptional activity *in vivo*.

Based on these junction analyses and the locations of identified TRS body sites, predicted subgenomic RNAs and their most 5' gene or genes are shown in **Figure 1E**. In addition to the viral genome, we predict 8 major subgenomic RNAs based both on identification of the TRS core sequence and presence of a major junction peak.

Coordinate-derived transcripts contain variant ORFs.

We next sought to determine if there were any variant or noncanonical ORFs present in SARS-CoV-2 transcripts. To address this question, we first filtered out coordinate-derived transcripts that did not contain the 5' leader sequence. ORFs were predicted directly from the remaining coordinate-derived transcripts and aligned to SARS-CoV-2 genes as detailed in the methods. This analysis revealed that the ORF count greatly increases for ORFs closer to the 3' end of the genome (**Figure 2A**), reflecting both the 3' bias inherent to dRNAseq and the tiered nature of the subgenomic RNAs. Interestingly, we found that alignments from a subset of predicted ORFs were "variant" and contained only a fraction of the expected protein sequence (**Figure 2A**). We paid particular attention to variant ORF1a, S, and M ORFs due to their abundance relative to their canonical variants.

The start-site distribution of canonical and variant ORFs on their transcripts revealed variant ORF1a, S, and M are generally closer to the 5' end of their respective subgenomic RNAs than canonical ORF1a, S, and M (**Figure 2B**). While there is evidence for leaky scanning of some CoV RNAs [12], this positioning may raise the possibility that they are expressed [13] and provides the opportunity for these variant ORFs to regulate downstream genes [12, 14].

Significant proportions of ORFs encoding M, S, and ORF1a are variant.

To elucidate the nature of these variant ORFs, we investigated the stop and start points of each variant ORF relative to their canonical counterparts (**Figure 2C**). This analysis revealed that all M variants were predicted to initiate at a TTG start codon 42 amino acids before the end of the M gene, ending at the canonical stop codon. While the efficacy of translation initiation from this start codon is unknown, it has an acceptable Kozak consensus site with an A at the -3 position and a G at the +4 position. Variant M ORFs make up between 7.9% and 19% of M ORFs in the three dRNAseq datasets and are likely created by the major junction point at the TRS sequence within the M reading frame (**Figure 1D**). If translated, variant M will lack the N-terminal transmembrane domains and contain regions predicted to be on the inside of the virion or in the cytosol.

S variants make up between 5.1% and 24.2% of all S ORFs identified in these datasets (**Figure 2C**). There is a more diffuse pattern of S start and stop points, indicating their formation may be stochastic and not tied to a specific junctional event. The majority of S variants are predicted to have a truncated N terminus and resultant lack of their N-terminal domain and receptor-binding domain [15], while maintaining their C-terminal transmembrane domains. Whether variant S ORFs are translated and packaged into virions remains to be determined.

Surprisingly, we found that nearly all identified ORF1a ORFs are variant (**Figure 2C**). These variant ORFs tend to consist of the N-terminal 200 to 800 amino acids of ORF1a, encoding Nsp1 and Nsp2. Nsp1 has 84.4% identity to the nsp1 of SARS-CoV-1, which is thought to inhibit host gene expression and inhibit the type I interferon response [16]. In addition, a recent proteomics analysis by Gordon et al suggests Nsp1 binds the DNA polymerase Alpha complex, raising the possibility that Nsp1 modulates DNA replication [17]. Nsp2 has 68% identity to Nsp2 of SARS-CoV-1; the function of Nsp2 is less clear, but it is thought to disrupt host signaling [18]. Viral protein-host protein interaction data by Gordon et al. suggest Nsp2 interacts with proteins involved in endosomal transport and translation repression.

Canonical junctions are hypothesized to be driven by homology between body TRS sites and the leader TRS site [6, 19]. To address the possibility that a specific TRS-like sequence is driving the observed non-canonical junctions in ORF1a and S, we assessed homology between the sequences flanking the 5' and 3' ends of each junction. Variations of the canonical TRS core sequence ACGAAC were the most common homologous sequences near the 5' and 3' ends of junctions with a 3' end near canonical TRS sites (**Figure 2D**). In contrast, neither junctions with 5' ends within ORF1a or junctions with 3' ends within S were flanked by a common homologous sequence. This observation suggests there is no specific TRS-like homologous sequence driving non-canonical junctions in these genes. In contrast, homologous sequences of junctions with 3' ends landing in M are dominated by the canonical TRS core sequence, suggesting that the TRS sequence within M is driving intra-M junctions.

The subgenomic RNAs predicted to encode variant ORF1a, S and M are depicted in **Figure 3A**. In addition, multiple sequence alignments of these variants and canonical ORF1a, S, and M from SARS-CoV-2 and the *Alphacoronavirus* CoV-229E are available in **Figures 3B-D**.

ORF1a variants are supported by coverage and junction analysis of long and short read datasets.

Analysis of read coverage in the three dRNAseq datasets revealed a “plateau” of elevated coverage at the 5’ end of the genome that supports the existence of a population of RNAs containing the 5’ fraction of the ORF1a. This elevated coverage decreases precipitously around genome position 1800, and again around genome position 6500 (**Figure 4A**). We found that a rise in the cumulative number of junctions is concomitant with these decreases in read coverage, supporting a model where this elevated coverage results from reads containing noncanonical junctions originating in ORF1a. Notably, the transition around genome position 1800 correlates well with the peak of variant ORF1a stop positions around residue position ~500 in ORF1a (**Figure 2C**).

It remained unclear if variant ORF1a is a major percentage of total ORF1a in the cell, or if this finding was due to a systematic bias favoring shorter reads in nanopore sequencing. If this finding was an artifact of dRNAseq, we would expect that this plateau should be elongated due to the existence of the longer canonical ORF1a ORFs that were selected against by dRNAseq.

To address this question, we investigated the coverage and junctions in three Illumina short-read datasets that do not have a bias against long transcripts [20, 21]. Analysis of these short-read datasets revealed a similar elevated coverage at the 5' side of ORF1a, with the expected decrease in sequencing coverage and increase in junctions around position 1800 (**Figure 4A**). Similar to the dRNAseq samples, a second decrease was observed around position 6500 (**Supplementary Figure 2A**).

We next investigated the distribution of ORF1a-originating junctions and found that these junctions increase in frequency towards the 3' end of the genome, with the highest abundance in N (**Figure 4B)**. This finding is consistent across dRNAseq and short read, paired end, datasets. To address the possibility that these junctions lead to fusions between ORF1a and downstream ORFs, we characterized the fusion status of the ORF1a variants and found that approximately 31% of ORF1a variants in all three dRNAseq datasets are C-terminal fusions, the majority containing N (**Figure 4C**). These fusions may be stochastic, as 1/3 of junctions would be expected to land in-frame of the downstream ORF.

To determine if non-canonical junctions within ORF1a are generated by other Coronaviruses, we investigated a dRNAseq transcriptome from cells infected with CoV-229E. In contrast to SARS-CoV-2, a *Betacoronavirus*, CoV-229E is an *Alphacoronavirus* that causes seasonal colds and is in active circulation. We found that there is a similar pattern of junctions and coverage of CoV-229E in infected cells, with two distinct drops in coverage at the 5' side of ORF1a concomitant with proportionate rises in junctions (**Supplementary Figure 2B**). While the major and minor inflection points occur in SARS-CoV-2 within Nsp2 and Nsp3 respectively, in CoV-229E they occur at the ends of Nsp1 (around genome position 620) and Nsp2 (around genome position 3200).

All together, these finding support the hypothesis that SARS-CoV-2 and other human CoVs generate truncated ORF1a ORFs at the stages of infection catalogued by these datasets.

Unannotated ORFs are abundant in SARS-CoV-2 subgenomic RNAs

Finally, we sought to determine if there was transcriptome support for ORFs that are currently unannotated. We conducted de-novo ORF prediction directly on the SARS-CoV-2 genome and searched for the presence of these ORFs in the three dRNAseq datasets. We found that numerous unannotated ORFs were as abundant as annotated ORFs including N and ORF10 (**Figure 5A**). The abundance of these ORFs partially reflects their location in the genome, with the majority of these unannotated ORFs starting 3' to position 28000. Counting the 5'-proximal ORF on each transcript revealed that only a small minority of transcripts contain an unannotated ORF as the most 5'-proximal ORF. N is the 5' proximal ORF on the most transcripts, followed by M, ORF7a, and ORF8. In support of conclusions reached by Taiaroa et al., Kim et al., and Davidson et al. [7-9], we also find very few transcripts that contain ORF10 as the most 5' ORF. However, translation of this ORF from other subgenomic RNAs cannot be ruled out.

The unannotated ORF 28284-28577 is particularly notable since it contains 70% nucleotide identity to SARS-CoV ORF9b. Furthermore, despite being the 5'-proximal ORF on a very small number of transcripts, expression of this ORF has strong proteomic support [22]. This likely reflects translation through leaky scanning of the N subgenomic RNA that has been observed in SARS-CoV-1 [12, 23]. With this in mind, lack of 5*'* proximity of other ORFs cannot rule out the probability they are expressed.

Discussion

We used dRNAseq and short-read datasets from cells infected with four independent SARS-CoV-2 isolates to identify unexpected, noncanonical RNA species. We show that SARS-CoV-2 produces RNAs that encode ORFs that may yield truncated variants of viral proteins. We identify a strong splice junction point within the M ORF, resulting in an N-terminal truncation in up to 19% of all M ORFs. We describe the identification of subgenomic RNAs harboring N-terminally truncated S ORFs which may lack the N-terminal and receptor-binding domains. Furthermore, we show that nearly all ORF1a ORFs detected in the three dRNAseq datasets contain only the N-terminal region of ORF1a. We support this finding by showing unexpected differences in read coverage of the 5' and 3' ends of ORF1a and show that this correlates with an increase in noncanonical junctions in dRNAseq and short-read datasets.

By identifying the 5' and 3' locations of junction sites in SARS-CoV-2 RNAs, we found that the vast majority of junctions were canonical junctions between the leader and body TRS sequences. Furthermore, our analysis reveals underlying levels of noncanonical junctions across the viral genome. The junction distribution in Figure 1B reveals common vertical lines at 3' TRS sites that have continuous 5' ends across the length of the viral genome (**Figure 1B-D**), suggesting that some noncanonical junctions are mediated by the 3' TRS.

By predicting ORFs directly from coordinate-derived transcripts, we found that large proportions of M, S, and ORF1a are variant. In the case of M, these variants are generated from a canonical junction between the leader sequence TRS and a TRS sequence located within the M open reading frame. We show that there is a TTG start codon within the M ORF that may make a truncated M variant. This TTG start codon has an optimal Kozak consensus site. While the expression of this M variant must be proven experimentally, the location of this ORF on resultant subgenomic RNAs gives it the potential to regulate the expression of ORF6 which is immediately downstream.

We found that up to 24% of all S ORFs are variant. Unlike the M and ORF1a variants, S variants have distributed start and stop positions, suggesting that S variants are created by various junctional events. The majority of these variants are expected to have lost their N-terminal domain and possibly the receptor-binding domain [15]. Notably, subgenomic RNAs encoding variant S ORFs have been reported for SARS-CoV-1 [24]. If translated, if will be important to determine if these S variants are incorporated into emergent virions and if they effect downstream virion entry.

Interestingly, we found that the vast majority of identified ORF1a ORFs in the three independent SARS-CoV-2 dRNAseq datasets are variant. These variant ORFs generally contain a consistent canonical start position but end prematurely around residue position 500, ultimately encoding Nsp1 and fragments of Nsp2. While the entirety of ORF1a is canonically transcribed as a large polypeptide with or without ORF1b [12], this may allow the virus to differentially regulate the level of different ORF1a substituents. We show that the presence of ORF1a variants is supported by elevated coverage at the 5' end of ORF1a in three SARS-CoV-2 long read datasets and three short read datasets, as well as in a CoV-229E dRNAseq dataset. Read coverage decreases in steps concomitantly with an increase in ORF1a-originating noncanonical junctions. We find that about 1/3 of these ORF1a variants are fused in frame with downstream ORFs, mostly N. Interestingly, some of the downstream ORF fusion partners are currently unannotated.

The fact that nearly all identified ORFs encoding ORF1a were variant raises the question of why nonvariant ORF1a was either undetected or at extremely low abundance. There are likely two reasons. First, nonvariant ORF1a would be found on the RNA genome, which would be at lower abundance that the subgenomic RNAs [25]. Second, given that genome-length transcripts are required to identify nonvariant ORF1a, the technical bias of dRNAseq against such large transcripts likely reduces their observed abundance [26].

While we find transcriptome evidence of variant ORFs in three independent dRNAseq datasets, and junctional support in three additional short-read datasets, it remains to be determined if they are expressed. Furthermore, while the studied datasets used four independent viral isolates, all studies utilized cells infected *in vitro*. It is important to investigate the transcriptional dynamics of SARS-CoV-2 *in vivo*.

Unlike junctions landing at canonical TRS body sites, we found that non-canonical junctions with a 5' end in ORF1a or 3' end in S did not have a single predominant homologous sequence near their 5' and 3' junction sites. In contrast, junctions with a 3' end in M often had the TRS core sequence near both their 5' and 3' junction points. This suggests that the non-canonical junctions in ORF1a and S are not driven by TRS-like homology and raises the possibility that these junctions originate stochastically or through an unknown mechanism.

After predicting unannotated ORFs directly from the SARS-CoV-2 genome sequence, we found transcriptome evidence of unannotated ORFs. There is a particular abundance of unannotated ORFs closer to the 3' end of the genome, likely reflecting the tiered nature of the subgenomic RNAs. However, we found that very few transcripts contain an unannotated ORF as the most 5' proximal ORF. We argue that this does not preclude the expression of these ORFs, and that this is an important area of future study. While the identity of subgenomic RNAs is often attributed to the most 5' ORF, there are many examples of polycistronic translation of CoV subgenomic RNAs through leaky scanning of upstream ORFs and through internal ribosome entry [12, 13, 23, 27-30].

In conclusion, we show that SARS-CoV-2 has higher coding potential than expected through two probable mechanisms. First, it can use noncanonical junction events to make variants or fusions of known coding sequences. Second, there may be a number of currently-uncharacterized open reading frames that generate protein products. Future studies are necessary to carefully delineate how these mechanisms influence SARS-CoV-2 pathogenesis. In the case of unannotated ORFs, antibody generation and the profiling of patient sera will be helpful in understanding if they are expressed, and if they elicit humoral immunity and pathogenesis. Finally, careful manipulations of viral ORFs using reverse genetics will be important to understand the role of unannotated ORFs and noncanonical junction events in the SARS-CoV-2 lifecycle.

Materials and Methods

Sequencing data

The three SARS-CoV-2 dRNAseq datasets, the isolate used, and the collection time after infection are listed below:

Taiaroa et al. [7]	Australia/VIC01/2020	Unknown
Kim et al. [8]	Korea/KCDC02/2020	24 hours post infection
Davidson et al. [9]	England/VE6-T/2020	5-7 days post infection

All three dRNAseq datasets utilized infected VERO cells.

All three SARS-CoV-2 short-read Illumina datasets were generated by Blanco-Melo et al. [10] using SARS-CoV-2 isolate USA/WA1/2020. All three samples utilized infected A549 cells ectopically overexpressing human ACE2, with RNA collected 24 hours post infection. SRA accessions: SRR11517741, SRR11517742, SRR11517743.

The CoV-229E long-read dataset was generated by Viehweger et al. [31]. In this experiment, Huh7 cells were infected with CoV-229E and RNA was collected 24 hours post infection.

Coordinate-derived transcript generation from long reads

Reads were first mapped to the SARS-CoV-2 genome (accession NC_045512.2) using minimap2 [32] and custom parameters detailed in minimap_sars2.sh and inspired by Kim et al. [8], and unmapped reads were discarded.

Next, coordinate-derived transcripts were generated using generate_synthetic_transcripts.py. In brief, this script reads a BAM file and the SARS-CoV-2 genome fasta and, by parsing the BAM alignment CIGAR and start position, determines the start, end, and junction coordinates of the alignment. Any minimap2-called deletion or intron of at least 1000 bases in length was considered a junction. Sequence between the calculated start and end positions, but not including sequence within junctions, were extracted from the genome fasta and output as coordinate-derived transcripts. A tab-delimited file containing the 5' and 3' coordinates of each junction was also produced.

All coordinate-derived transcripts were then aligned against a minimal leader sequence fragment (5' – CTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTCTC – 3') using blastn [33]. This minimal sequence was used instead of the full leader sequence because inspection of 5' coverage in the three long-read datasets revealed a significant drop in coverage at the end of this sequence. Transcripts with an alignment of at least 45 nucleotides in length and a percent identity of at least 85%, and that contain the start of this sequence within the first 50 nucleotides of the transcript, were kept. Transcripts not containing this leader sequence were discarded.

Analysis of junction coordinates from long reads

The 5' and 3' coordinates of each junction were processed in R [34], and figures generated with ggplot2 [35]. Gene-maps were generated using the gggenes R package [36]. Junctions are presented as one- or two-dimensional histograms, with the bin-size of each plot noted in the figure legend. A bin size of 10 bases, for example, indicates that one bar contains the total number of junctions that occurred in a given 10 base span.

The TRS sites in Figure 1D were labeled based on the approximate position of the TRS core sequence, 5' – ACGAAC – 3'.

Junction analysis of Illumina short reads

The short-read datasets were processed through minimap2 and generate_synthetic_transcripts.py in a similar manner as the long-read datasets and junction coordinates were input into R for analysis and plotting.

Assessment of homologous sequences proximal to 5' and 3' junction points

To determine if there was sequence homology enrichment near 5' and 3' junction points we identified the 15 nucleotides on either side of each point based on the SARS-CoV-2 reference genome sequence, resulting in 30 bases of context at both the 5' and 3' sites of each junction. We then determined the longest homologous match between these sequences. For plotting, we separated the junctions into categories. If a junction had a

3' end within 15 nucleotides of a canonical core TRS site that is to the 5' of a gene, it was considered a canonical junction of that gene. If a junction had a 3' end landing within the coding sequence of an ORF, it was considered internal. The exception to this rule is ORF1a – any junctions with a 5' location inside ORF1a were considered ORF1a internal junctions. ORF6, ORF7b, and ORF10 were not included in this analysis because they do not have a 5' canonical TRS sequence. The percentage of the junctions that consist of the most common homologous sequence were plotted.

De-novo ORF prediction and analysis from long reads

ORFs were called directly from each transcript using Prodigal [37], using parameters listed in prodigal.sh. Each ORF was translated into amino acid sequence using prodigal_to_orf_direct.py. If multiple Prodigal-predicted ORFs contained an overlapping amino acid sequence, the longest ORF was output. Each ORF was output in fasta format and labeled with transcript of origin and coordinates of the ORF on the transcript.

The amino acid sequence of each ORF was then mapped against canonical SARS-CoV-2 proteins (from accession NC_045512.2) as well as Prodigal-predicted unannotated ORFs from the SARS-CoV-2 genome using DIAMOND [38]. DIAMOND parameters are described in diamond.sh.

Various statistics of each ORF, including alignment information, the start position of the ORF on the transcript, and fusion information, were generated from the DIAMOND alignment file using parse_orf_assignments.py. An ORF was considered variant if it was not the same length as the canonical protein, with a notable exception. Because there are alternative start codons upstream of some annotated genes, and Prodigal would call ORFs using these start codons if they are present on a transcript, ORFs were allowed to have up to 20 amino acids of additional N-terminal amino acids while still being classified as canonical if they end at the expected position. If an ORF had matches against multiple SARS-CoV-2 proteins, it was given a primary assignment of the protein with the longest alignment length.

Assessment of genome coverage

Coverage at each position was assessed from minimap2-generated BAMs using the depth subcommand of samtools [39]. Command:

```
samtools depth -aa -d0 file.bam > file.cov
```

Multiple sequence alignment

Alignment was conducted using ClustalW [40] alignment through the MSA R package [41].

Code availability

The reproducible computational pipeline used to generate all results is publicly available at https://github.com/jnoms/virORF_direct. Furthermore, all R scripts used for plotting are present in this repository.

Author Contributions

J.N. analyzed the data; J.N., M.M., J.A.D. designed the study and wrote the manuscript.

Acknowledgements

The authors thank Taiaroa et al., Kim et al., and Davidson et al. for collecting the direct RNA sequencing data and making these data publicly available. We also thank Blanco-Melo et al. for their generation and release of the Illumina short read data. Portions of this research were conducted by using the O2 High Performance Compute Cluster, supported by the Research Computing Group at Harvard Medical School.

Conflicts of Interest

M.M. receives research support from Bayer, Ono, and Janssen, has patents licensed to Bayer and Labcorp, and is a consultant for OrigiMed. J.A.D. received research support from Constellation Pharmaceuticals and is a consultant for EMD Serono, Inc. and Merck & Co. Inc.

Funding

This work was supported in part by the US Public Health Service grants R35CA232128 and P01CA203655 to J.A.D.

Figure legends

Figure 1. Non-canonical junctions in the SARS-CoV-2 transcriptome.

A) Error-prone direct RNAseq reads from three independent datasets were mapped against the SARS-CoV-2 genome using minimap2. Using the resultant mapping information, genomic coordinates corresponding to the start and end of each transcript was determined. Deletions or introns with a length greater than 1000 detected by minimap2 were considered junctions and the genomic coordinates of the 5' and 3' ends of these junctions were collected. Synthetic transcripts were generated by extracting genomic sequences using the start, junction, and end coordinates.

B) Junctions were plotted based on their 5' and 3' coordinates over the entire SARS-CoV-2 genome. Darker color indicates a higher number of transcripts with similar 5' and 3' junctions. Bin-size is 100 bases. Columns are labeled by virus isolate: Australia/VIC01/2020 is from Taiaroa et al. [7]; Korea/KCDC03/2020 is from Kim et al. [8]; and England/VE6-T/2020 is from Davidson et al [9]. These three datasets were generated by dRNAseq.

C) Cut-out from figure 1B including only junctions with a 5' end before position 100 and 3' end after position 21000 in the genome are plotted. Dashed lines indicate the start coordinates of annotated viral genes. Bin-size is 1 base in the 5' direction and 100 bases in the 3' direction. Note that junction locations are 0-indexed, i.e. position 74 corresponds to base 75.

D) A histogram of 3' junctions past position 21000 with a 5' end before position 100 are plotted. Dashed lines indicate the start coordinates of annotated viral genes. Bin-size is 20 bases. The red arrow notes the major junction point within the M ORF.

E) Based on junction analysis, the predicted species of virus-produced RNAs are represented. The most 5' gene or genes on each subgenomic RNA are listed.

Figure 2. There are significant levels of variant M, S, and ORF1a.

A) Open reading frames were predicted directly from dRNAseq synthetic transcripts from three independent datasets. The plot represents the total number of transcripts containing each ORF. Canonical ORFs are colored blue, while variant ORFs are colored red.

B) A boxplot of the distribution of start-sites for the indicated ORF on each transcript, where a more 5' ORF will have a start-site closer to 1.

C) (Top) Schematics of M, N, and ORF1ab are displayed with the approximate location of predicted transmembrane domains labeled in red. (Bottom) A histogram of the start and end sites of variant M, N, and ORF1ab ORFs are displayed. Blue bars represent the start sites, and red represent the end sites of each variant ORF. The percentage to the right of each graph is the total percentage of each predicted protein that is variant. Histogram bin-sizes: M, 1; N, 10 ; ORF1ab, 20. NTD: N-terminal domain. RBD: Receptor binding domain. CD: Connector domain. NSP: Nonstructural protein.

D) Homologous sequences present in the 15 bases on either side of the 5' and 3' ends of each junction. Junctions are classified by the location of their 3' end – if this is within 15 bases of the canonical TRS site or if it falls within a gene it is assigned accordingly. The only exception is ORF1a – junctions with a 5' end originating in ORF1a are assigned to ORF1a. Labels represent the most common homologous sequence between the ends of each junction. The core TRS sequence ACGAAC is underlined.

Figure 3. Representative ORF1a, S, and M variants.

A) Schematic depicting representative structures of the subgenomic RNAs encoding identified variant proteins. ORF length is not to scale.

B) Alignment of a representative SARS-Cov-2 ORF1a variant against the canonical ORF1a and N from CoV-229E and SARS-CoV-2. The N-terminal 185 residues of the variant were aligned with the ORF1a protein sequences, while the C-terminal 32 residues were aligned with the N protein sequences using ClustalW [40] alignment through the MSA R package [41].

C) Alignment of a representative SARS-Cov-2 S variant against the canonical S from CoV-229E and SARS-CoV-2. The S variant is missing the N-terminal 344 residues of S. The initial alignment is present, but the additional lines of alignment are not displayed.

D) Alignment of a representative SARS-Cov-2 M variant against the canonical M from CoV-229E and SARS-CoV-2. This variant consists of the C-terminal 42 residues of M and is the most common variety of M variant.

Figure 4. Coverage and junctions from dRNAseq and short-read datasets support the identified ORF1a variants.

A) Read coverage of the first 25384 bases of the SARS-CoV-2 genome is plotted in black. The cumulative number of junctions, starting at position 240 and ending at position 25384, is plotted in red. The early inflection point of coverage and junctions is marked with a red arrow. A second inflection can be viewed near position 6500. USA/WA1/2020 samples are short-read Illumina datasets.
B) A histogram of the 3’ locations of junctions that have a 5’ end within ORF1a. Bin size is 30 bases.
C) The identity of ORF1a-fusion partners is plotted on the Y axis, with the count of such fusions on the X axis. The top 10 fusion partners for each sample are represented. Color indicates if the fusion partner is on the N and C terminus of ORF1a, if the terminus is ambiguous, or if the fusion is a “self” fusion between an upstream and downstream region of ORF1a. The percentage notes the percent of ORFs assigned to ORF1a that have a fusion. ORF1a fragments are generally 200-800 residues in length, as indicated by figure 2C.

Figure 5. Transcriptome evidence for unannotated ORFs.

A) All ORFs in the SARS-CoV-2 genome starting with NTG were determined. Total ORF counts are plotted. Only the top 15 most abundant predicted ORFs from the Taiaroa et al. dataset are plotted.
B) These counts represent the number of transcripts that contain each ORF as the most 5-proximal ORF.

Supplementary Figure 1. Short-read Illumina datasets support non-canonical junctions.

A) Junctions were predicted from three short-read RNA sequencing samples generated by Blanco-Melo et al., and analysis was conducted identically to the three long-read dRNAseq datasets. Junctions were plotted based on their 5’ and 3’ coordinates over the entire SARS-CoV-2 genome. Darker color indicates a higher number of transcripts with similar 5’ and 3’ junctions. Dashed lines indicate the start coordinates of annotated viral genes. Bin-size is 100 bases.
B) Junctions with a 5’ end before 100nt and 3’ end after position 21000 in the genome are plotted. Dashed lines indicate the start coordinates of annotated viral genes. Bin-size is 1 base in the 5’ direction and 100 bases in the 3’ direction.
F) A histogram of 3’ junctions past position 21000 that have a 5’ end before position 100 are plotted. Dashed lines indicate the start coordinates of viral ORFs. The red arrow notes the major junction point within the M ORF. Bin-size is 20 bases.

Supplementary Figure 2. Similar to dRNAseq datasets, the short-read datasets have primary and secondary inflection points in the coverage and junctions ORF1a.

A) Coverage (black) and cumulative junctions (red) of the three Illumina short-read datasets. These are the same data plotted in Figure 3A, but zoomed in. The first inflection point of junctions and coverage (around position 1800) and the second (around position 6500) are labeled with a red arrow.

B) Coverage (black) and cumulative junctions (red) of CoV-229E-infected Huh7 cells. The first inflection point of junctions and coverage (around position 620) and the second (around position 3200) are labeled with a red arrow.

References

1. Wu, F., et al., *A new coronavirus associated with human respiratory disease in China.* Nature, 2020. **579**(7798): p. 265-269.
2. Plant, E.P. and J.D. Dinman, *The role of programmed-1 ribosomal frameshifting in coronavirus propagation.* Frontiers in bioscience: a journal and virtual library, 2008. **13**: p. 4873.
3. Narayanan, K., C. Huang, and S. Makino, *SARS coronavirus accessory proteins.* Virus research, 2008. **133**(1): p. 113-121.
4. de Haan, C.A., et al., *The group-specific murine coronavirus genes are not essential, but their deletion, by reverse genetics, is attenuating in the natural host.* Virology, 2002. **296**(1): p. 177-189.
5. Yount, B., et al., *Severe acute respiratory syndrome coronavirus group-specific open reading frames encode nonessential functions for replication in cell cultures and mice.* Journal of virology, 2005. **79**(23): p. 14909-14922.
6. Sawicki, S.G., D.L. Sawicki, and S.G. Siddell, *A contemporary view of coronavirus transcription.* Journal of virology, 2007. **81**(1): p. 20-29.
7. Taiaroa, G., et al., *Direct RNA sequencing and early evolution of SARS-CoV-2.* bioRxiv, 2020: p. 2020.03.05.976167.
8. Kim, D., et al., *The architecture of SARS-CoV-2 transcriptome.* bioRxiv, 2020: p. 2020.03.12.988865.
9. Davidson, A.D., et al., *Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site.* bioRxiv, 2020: p. 2020.03.22.002204.
10. Blanco-Melo, D., et al., *SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems.* bioRxiv, 2020: p. 2020.03.24.004655.
11. Rang, F.J., W.P. Kloosterman, and J. de Ridder, *From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy.* Genome biology, 2018. **19**(1): p. 90.
12. Nakagawa, K., K. Lokugamage, and S. Makino, *Viral and cellular mRNA translation in coronavirus-infected cells*, in *Advances in virus research*. 2016, Elsevier. p. 165-192.
13. Firth, A.E. and I. Brierley, *Non-canonical translation in RNA viruses.* The Journal of general virology, 2012. **93**(Pt 7): p. 1385.
14. Morris, D.R. and A.P. Geballe, *Upstream open reading frames as regulators of mRNA translation.* Molecular and cellular biology, 2000. **20**(23): p. 8635-8642.
15. Wrapp, D., et al., *Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation.* Science, 2020. **367**(6483): p. 1260-1263.

16. Narayanan, K., et al., *Severe acute respiratory syndrome coronavirus nsp1 suppresses host gene expression, including that of type I interferon, in infected cells.* Journal of virology, 2008. **82**(9): p. 4471-4479.
17. Gordon, D.E., et al., *A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug-Repurposing.* bioRxiv, 2020: p. 2020.03.22.002386.
18. Cornillez-Ty, C.T., et al., *Severe acute respiratory syndrome coronavirus nonstructural protein 2 interacts with a host protein complex involved in mitochondrial biogenesis and intracellular signaling.* Journal of virology, 2009. **83**(19): p. 10314-10318.
19. Di, H., et al., *Expanded subgenomic mRNA transcriptome and coding capacity of a nidovirus.* Proceedings of the National Academy of Sciences, 2017. **114**(42): p. E8895-E8904.
20. Finotello, F. and B. Di Camillo, *Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis.* Briefings in functional genomics, 2015. **14**(2): p. 130-142.
21. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature methods, 2008. **5**(7): p. 621.
22. Cinatl, J., et al., *SARS-CoV-2 infected host cell proteomics reveal potential therapy targets.* 2020.
23. Shi, C.-S., et al., *SARS-coronavirus open reading frame-9b suppresses innate immunity by targeting mitochondria and the MAVS/TRAF3/TRAF6 signalosome.* The Journal of Immunology, 2014. **193**(6): p. 3080-3089.
24. Hussain, S., et al., *Identification of novel subgenomic RNAs and noncanonical transcription initiation signals of severe acute respiratory syndrome coronavirus.* Journal of virology, 2005. **79**(9): p. 5288-5295.
25. Hofmann, M.A., P.B. Sethna, and D.A. Brian, *Bovine coronavirus mRNA replication continues throughout persistent infection in cell culture.* Journal of Virology, 1990. **64**(9): p. 4108-4114.
26. Byrne, A., et al., *Realizing the potential of full-length transcriptome sequencing.* Philosophical Transactions of the Royal Society B, 2019. **374**(1786): p. 20190097.
27. O'Connor, J.B. and D.A. Brian, *Downstream ribosomal entry for translation of coronavirus TGEV gene 3b.* Virology, 2000. **269**(1): p. 172-182.
28. Liu, D. and S. Inglis, *Internal entry of ribosomes on a tricistronic mRNA encoded by infectious bronchitis virus.* Journal of virology, 1992. **66**(10): p. 6143-6154.
29. Jendrach, M., V. Thiel, and S. Siddell, *Characterization of an internal ribosome entry site within mRNA 5 of murine hepatitis virus.* Archives of virology, 1999. **144**(5): p. 921-933.
30. Thiel, V. and S.G. Siddell, *Internal ribosome entry in the coding region of murine hepatitis virus mRNA 5.* Journal of General Virology, 1994. **75**(11): p. 3041-3046.
31. Viehweger, A., et al., *Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis.* Genome research, 2019. **29**(9): p. 1545-1554.
32. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.

33. Camacho, C., et al., *BLAST+: architecture and applications.* BMC bioinformatics, 2009. **10**(1): p. 421.

34. Team, R.C., *R: A language and environment for statistical computing.* 2013.

35. Wickham, H., *ggplot2.* Wiley Interdisciplinary Reviews: Computational Statistics, 2011. **3**(2): p. 180-185.

36. Wilkins, D., *gggenes: draw gene arrow maps in 'ggplot2'. R package version 0.4. 0*. 2019.

37. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC bioinformatics, 2010. **11**(1): p. 119.

38. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using DIAMOND.* Nature methods, 2015. **12**(1): p. 59.

39. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.

40. Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* bioinformatics, 2007. **23**(21): p. 2947-2948.

41. Bodenhofer, U., et al., *msa: an R package for multiple sequence alignment.* Bioinformatics, 2015. **31**(24): p. 3997-3999.

Figure 1

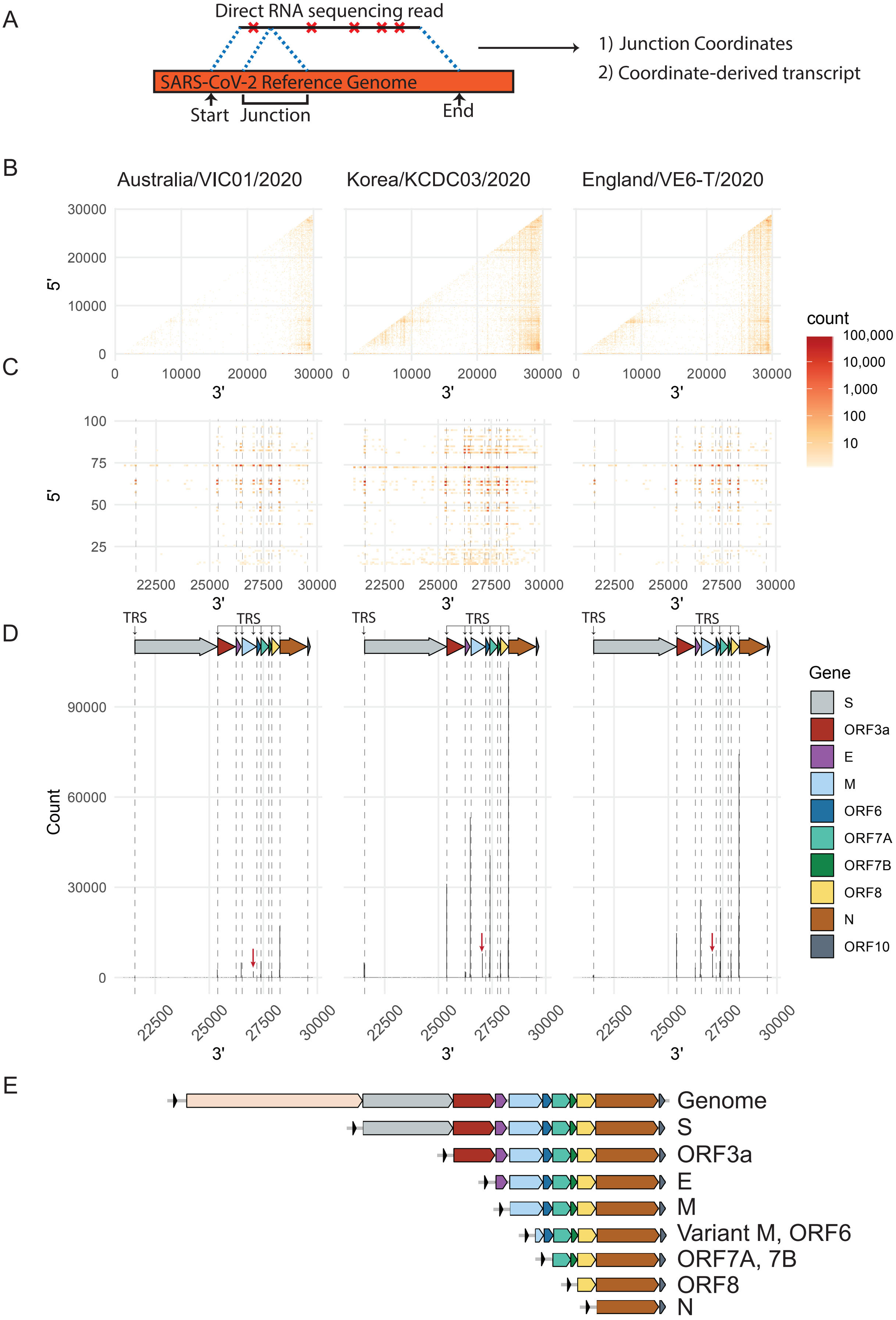

A
Direct RNA sequencing read
SARS-CoV-2 Reference Genome
Start
Junction
End
1) Junction Coordinates
2) Coordinate-derived transcript
B
Australia/VIC01/2020
Korea/KCDC03/2020
England/VE6-T/2020
5'
30000
20000
10000
0
0
10000
20000
30000
3'
count
100,000
10,000
1,000
100
10
C
5'
100
75
50
25
22500
25000
27500
30000
3'
D
TRS
TRS
Count
90000
60000
30000
0
22500
25000
27500
30000
3'
Gene
S
ORF3a
E
M
ORF6
ORF7A
ORF7B
ORF8
N
ORF10
E
Genome
S
ORF3a
E
M
Variant M, ORF6
ORF7A, 7B
ORF8
N

Figure 2

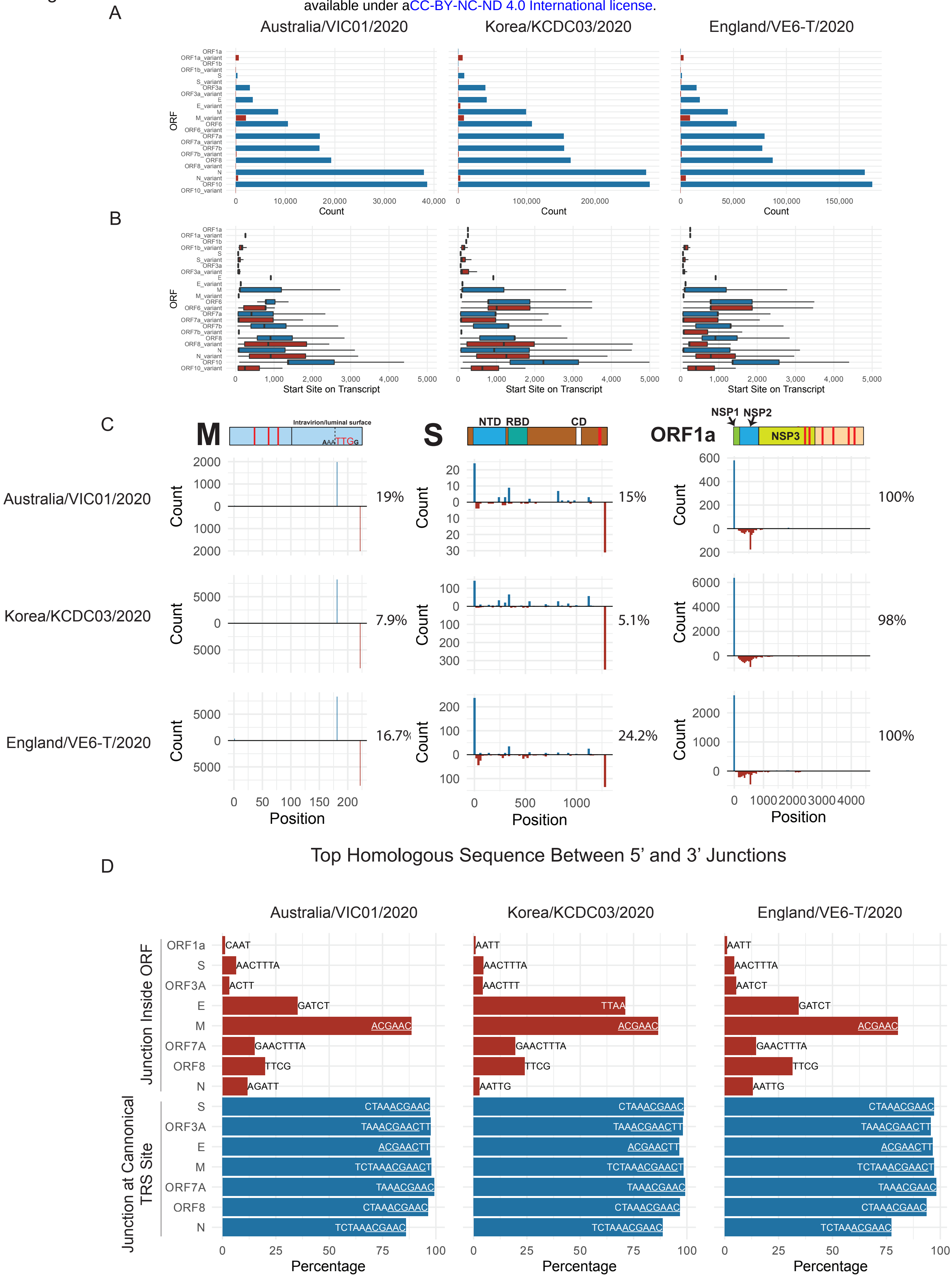
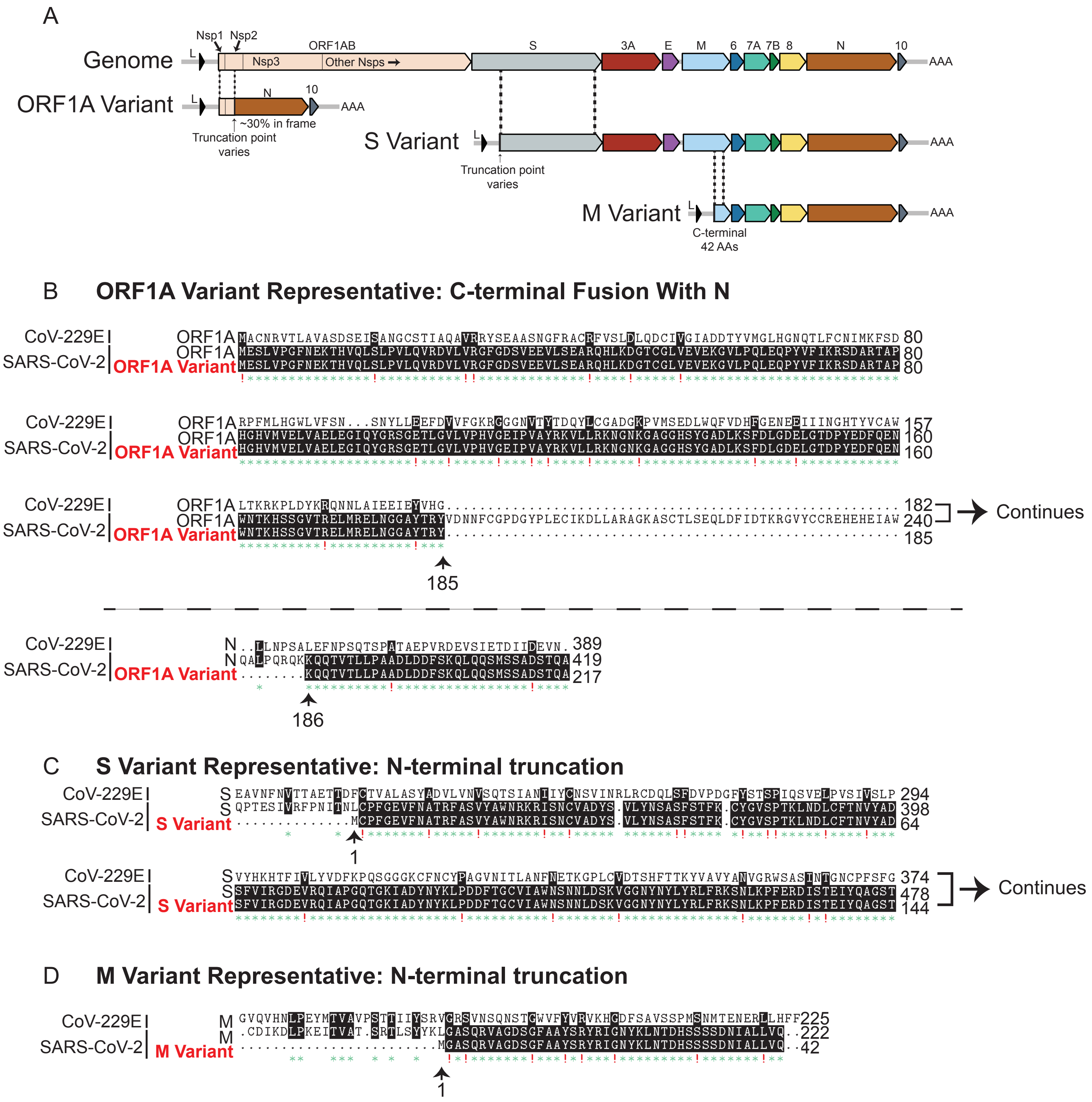

A
Australia/VIC01/2020
Korea/KCDC03/2020
England/VE6-T/2020
ORF
ORF1a
ORF1a_variant
ORF1b
ORF1b_variant
S
S_variant
ORF3a
ORF3a_variant
E
E_variant
M
M_variant
ORF6
ORF6_variant
ORF7a
ORF7a_variant
ORF7b
ORF7b_variant
ORF8
ORF8_variant
N
N_variant
ORF10
ORF10_variant
0
10,000
20,000
30,000
40,000
100,000
200,000
50,000
150,000
Count
B
1,000
2,000
3,000
4,000
5,000
Start Site on Transcript
C
M
Intravirion/luminal surface
AAATTGG
S
NTD
RBD
CD
ORF1a
NSP1
NSP2
NSP3
Count
Position
19%
15%
100%
7.9%
5.1%
98%
16.7%
24.2%
100%
2000
1000
0
5000
20
10
30
100
200
300
600
400
6000
4000
0
50
100
150
200
500
1000
2000
3000
4000
D
Top Homologous Sequence Between 5' and 3' Junctions
Junction Inside ORF
Junction at Cannonical TRS Site
ORF1a
S
ORF3A
E
M
ORF7A
ORF8
N
CAAT
AACTTTA
ACTT
GATCT
ACGAAC
GAACTTTA
TTCG
AGATT
AATT
AACTTT
TTAA
AATTG
AATCT
CTAAACGAAC
TAAACGAACTT
ACGAACTT
TCTAAACGAACT
TAAACGAAC
CTAAACGAAC
TCTAAACGAAC
0
25
50
75
100
Percentage

Figure 3

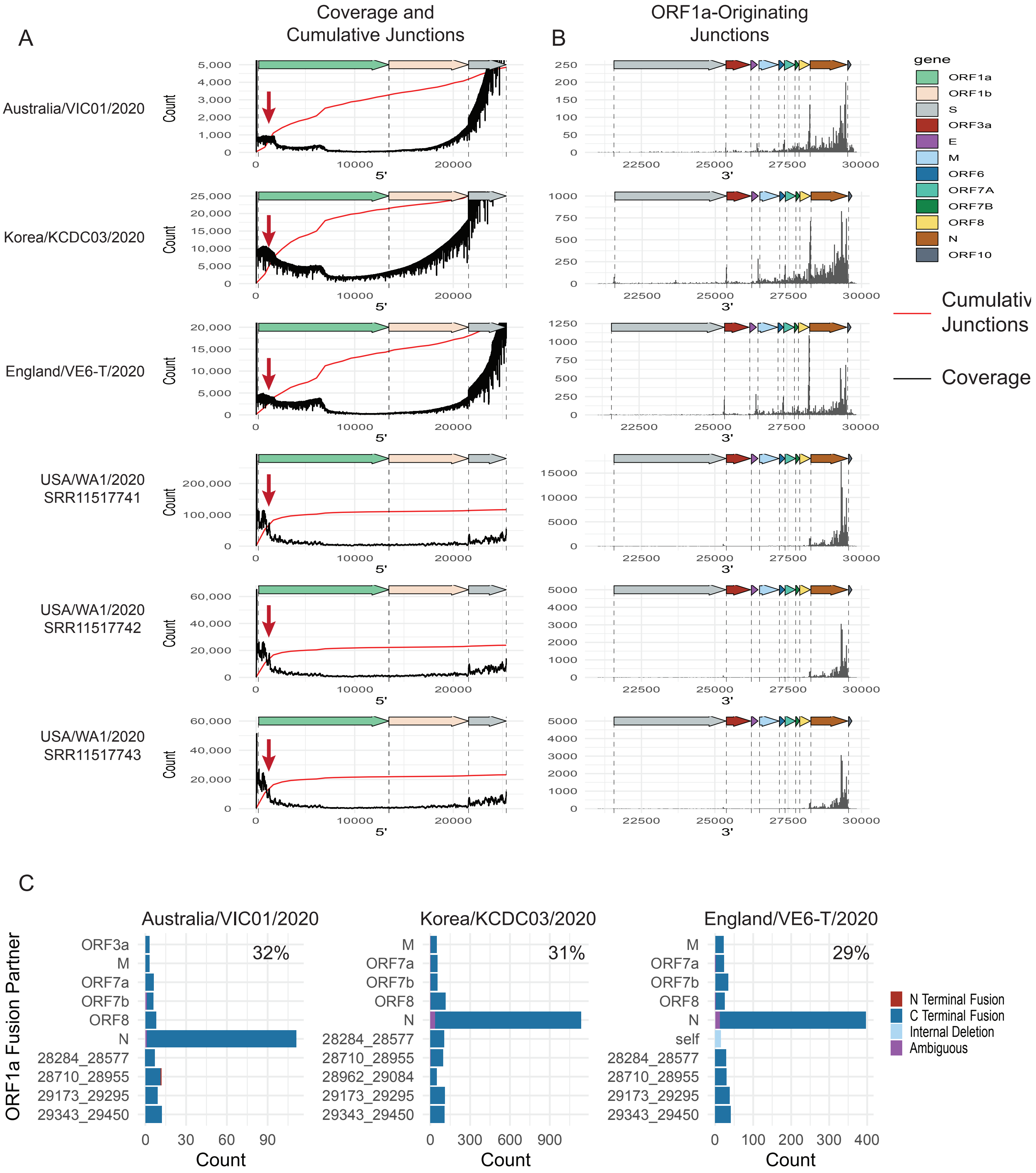

A
Genome
L
Nsp1
Nsp2
ORF1AB
Nsp3
Other Nsps →
S
3A
E
M
6
7A
7B
8
N
10
AAA
ORF1A Variant
L
N
10
AAA
~30% in frame
Truncation point varies
S Variant
L
AAA
Truncation point varies
M Variant
L
AAA
C-terminal 42 AAs
B
CoV-229E
SARS-CoV-2
ORF1A MACNRVTLAVASDSEISANGCSTIAQAVRRYSEAASNGFRACRFVSLDLQDCIVGIADDTYVMGLHGNQTLFCNIMKFSD 80
ORF1A MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAP 80
ORF1A Variant MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVEKGVLPQLEQPYVFIKRSDARTAP 80
ORF1A RPFMLHGWLVFSN...SNYLLEEFDVVFGKRGGGNVTYTDQYLCGADGKPVMSEDLWQFVDHFGENEEIIINGHTYVCAW 157
ORF1A HGHVMVELVAELEGIQYGRSGETLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQEN 160
ORF1A Variant HGHVMVELVAELEGIQYGRSGETLGVLVPHVGEIPVAYRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQEN 160
ORF1A LTKRKPLDYKRQNNLAIEEIEYVHG... 182
ORF1A WNTKHSSGVTRELMRELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIAW 240
ORF1A Variant WNTKHSSGVTRELMRELNGGAYTRY... 185
Continues
185
N ..LLNPSALEFNPSQTSPATAEPVRDEVSIETDIIDEVN. 389
N QALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADSTQA 419
ORF1A Variant ........KQQTVTLLPAADLDDFSKQLQQSMSSADSTQA 217
186
C
S EAVNFNVTTAETTDFCTVALASYADVLVNVSQTSIANIIYCNSVINRLRCDQLSFDVPDGFYSTSPIQSVELPVSIVSLP 294
S QPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYS.VLYNSASFSTFK.CYGVSPTKLNDLCFTNVYAD 398
S Variant ...............MCPFGEVFNATRFASVYAWNRKRISNCVADYS.VLYNSASFSTFK.CYGVSPTKLNDLCFTNVYAD 64
1
S VYHKHTFIVLYVDFKPQSGGGKCFNCYPAGVNITLANFNETKGPLCVDTSHFTTKYVAVYANVGRWSASINTGNCPFSFG 374
S SFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGST 478
S Variant SFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGST 144
Continues
D
M GVQVHNLPEYMTVAVPSTTIIYSRVGRSVNSQNSTGWVFYVRVKHGDFSAVSSPMSNMTENERLLHFF225
M .CDIKDLPKEITVAT.SRTLSYYKLGASQRVAGDSGFAAYSRYRIGNYKLNTDHSSSSDNIALLVQ..222
M Variant .........................MGASQRVAGDSGFAAYSRYRIGNYKLNTDHSSSSDNIALLVQ..42
1

B ORF1A Variant Representative: C-terminal Fusion With N

C S Variant Representative: N-terminal truncation

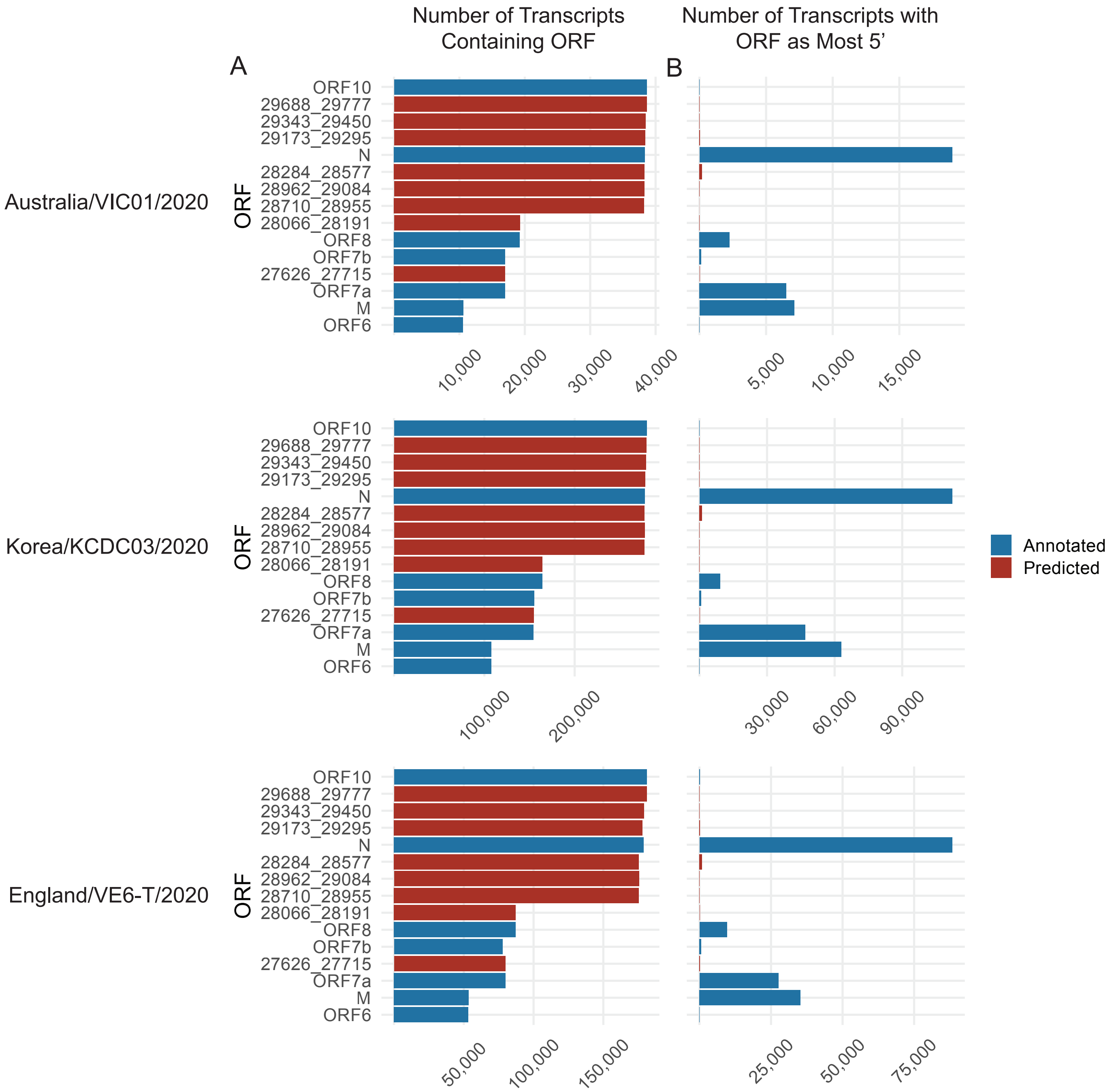
D M Variant Representative: N-terminal truncation

Figure 4

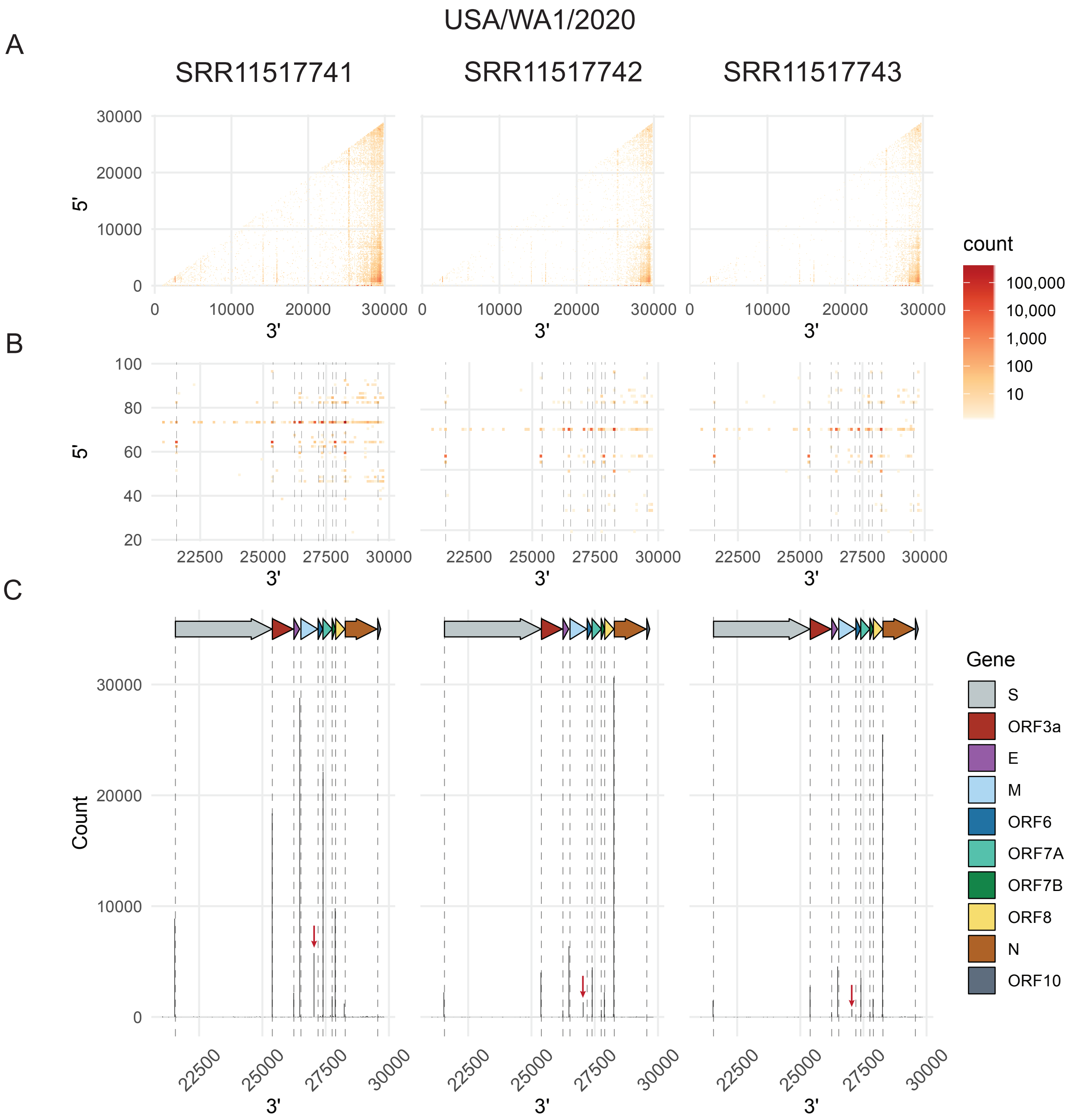


A
Coverage and Cumulative Junctions
B
ORF1a-Originating Junctions
Australia/VIC01/2020
Korea/KCDC03/2020
England/VE6-T/2020
USA/WA1/2020 SRR11517741
USA/WA1/2020 SRR11517742
USA/WA1/2020 SRR11517743
Count
5,000
4,000
3,000
2,000
1,000
0
25,000
20,000
15,000
10,000
200,000
100,000
60,000
40,000
20,000
0
10000
20000
5'
3'
250
200
150
100
50
1000
750
500
1250
15000
10000
5000
4000
3000
2000
22500
25000
27500
30000
gene
ORF1a
ORF1b
S
ORF3a
E
M
ORF6
ORF7A
ORF7B
ORF8
N
ORF10
Cumulative Junctions
Coverage
C
ORF1a Fusion Partner
Australia/VIC01/2020
Korea/KCDC03/2020
England/VE6-T/2020
32%
31%
29%
ORF3a
M
ORF7a
ORF7b
ORF8
N
self
28284_28577
28710_28955
28962_29084
29173_29295
29343_29450
0
30
60
90
300
600
900
100
200
300
400
Count
N Terminal Fusion
C Terminal Fusion
Internal Deletion
Ambiguous

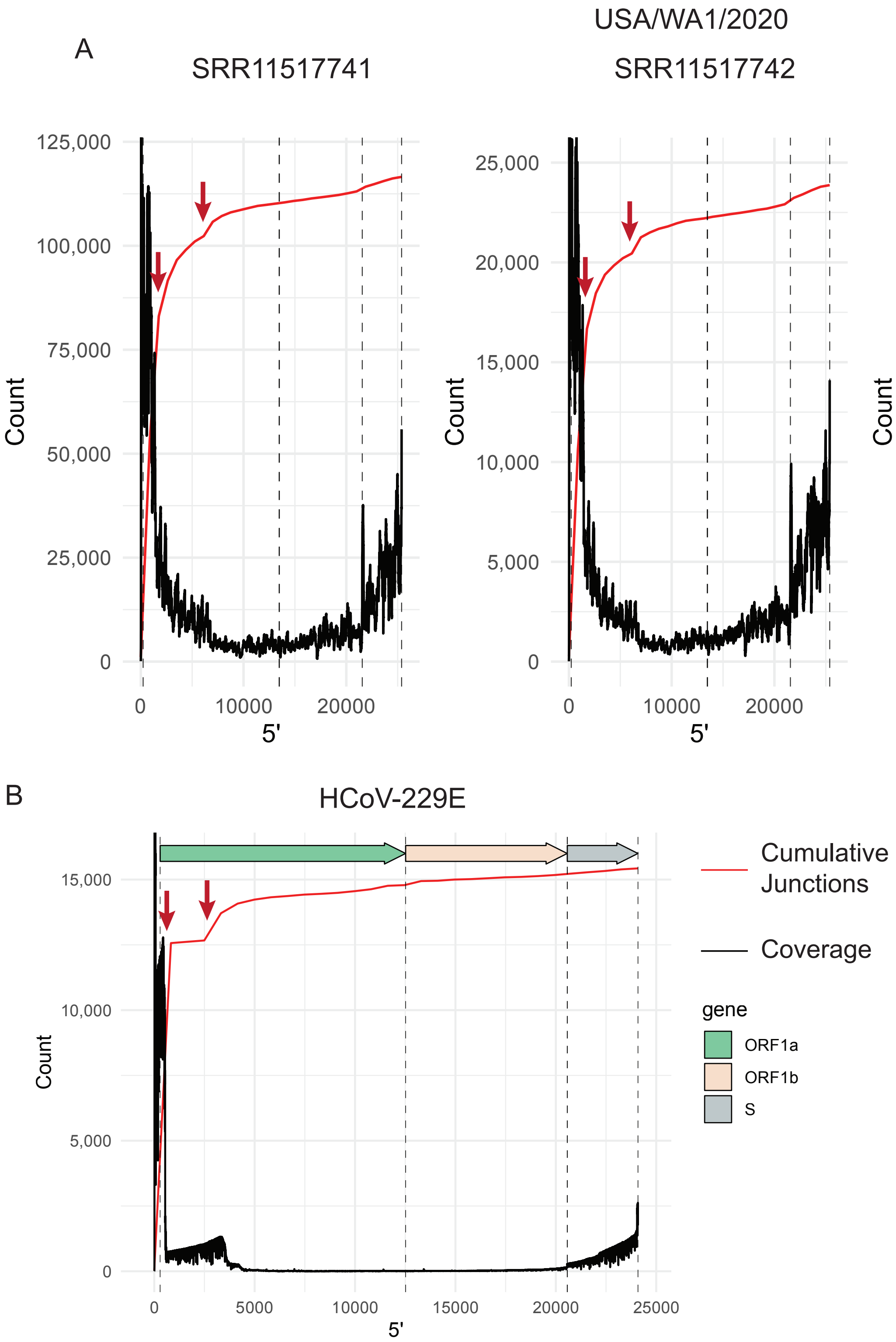
Figure 5


Number of Transcripts Containing ORF
Number of Transcripts with ORF as Most 5'
A
B
Australia/VIC01/2020
Korea/KCDC03/2020
England/VE6-T/2020
ORF
ORF10
29688_29777
29343_29450
29173_29295
N
28284_28577
28962_29084
28710_28955
28066_28191
ORF8
ORF7b
27626_27715
ORF7a
M
ORF6
10,000
20,000
30,000
40,000
5,000
10,000
15,000
100,000
200,000
30,000
60,000
90,000
50,000
100,000
150,000
25,000
50,000
75,000
Annotated
Predicted

Supplemental Figure 1



Supplemental Figure 2


USA/WA1/2020
A
SRR11517741
SRR11517742
SRR11517743
Count
5'
B
HCoV-229E
Cumulative Junctions
Coverage
gene
ORF1a
ORF1b
S