The SARS-CoV-2 Subgenome Landscape and its Novel Regulatory Features

Dehe Wang, Ao Jiang, Jiangpeng Feng, Guangnan Li, Dong Guo, Muhammad Sajid, Kai Wu, Qiuhan Zhang, Yann Ponty, Sebastian Will, et al.

The SARS-CoV-2 Subgenome Landscape and its Novel Regulatory Features

Dehe Wang[1, 2, 5] Ao Jiang[1, 5] Jiangpeng Feng[1, 5] Guangnan Li[1, 2, 5]
Dong Guo[1, 5] Muhammad Sajid[1] Kai Wu[1, 2] Qiuhan Zhang[1] Yann Ponty[3]
Sebastian Will[3] Feiyan Liu[1, 2] Xinghai Yu[1, 2] Shaopeng Li[1, 2] Qianyun Liu[1]
Xing-Lou Yang[4] Ming Guo[1] Xingqiao Li[1, 2] Mingzhou Chen[1]
Zheng-Li Shi[4] Ke Lan[1, 2, *] Yu Chen[1, *]
and Yu Zhou[1, 2, 6, *]

[1] State Key Laboratory of Virology, Modern Virology Research Center, College of Life Sciences, Wuhan University, Wuhan, China
[2] Frontier Science Center for Immunology and Metabolism, Wuhan University, Wuhan, China
[3] CNRS UMR 7161 LIX, Ecole Polytechnique, Institut Polytechnique de Paris, France
[4] CAS Key Laboratory of Special Pathogens, Wuhan Institute of Virology, Center for Biosafety Mega-Science, Chinese Academy of Sciences, Wuhan, China
[5] These authors contributed equally
[6] Lead Contact
[*]Correspondence: klan@whu.edu.cn (K.L.), chenyu@whu.edu.cn (Y.C.),yu.zhou@whu.edu.cn (Y.Z.

Abstract

COVID-19, caused by Coronavirus SARS-CoV-2,is now in global pandemic. Coronaviruses are known to generate negative subgenomes (sgRNAs)through Transcription-Regulating Sequence (TRS)-dependent template switch, but the global dynamic landscapes of coronaviral subgenomes and regulatory rules remain unclear. Here, using NGS short-read and Nanopore long-read poly(A) RNA sequencing in two cell types at multiple time points post-infection of SARS-CoV-2, we identified hundreds of template switches and constructed the dynamic landscapes of SARS-CoV-2 subgenomes. Interestingly, template switch could occur in bidirectional manner, with diverse SARS-CoV-2 subgenomes generated from successive template switching events. The majority of template switches result from RNA-RNA interactions,including seed and compensatory modes, with terminal pairing status as a key determinant. Moreover, two TRS-independent template switch modes are also responsible for subgenome biogenesis. Collectively, our findings reveal the subgenome landscape of SARS-CoV-2 and its regulatory features, providing a molecular basis for understanding subgenome biogenesis and developing novel anti-viral strategies.

1 INTRODUCTION

The recent outbreak of COVID-19 caused by SARS-CoV-2 (also referred to as HCoV-19) (**???**) has turned into a global pandemic, causing more than a million deaths as of October 2020 (**?**). SARS-CoV-2 is an enveloped RNA virus with $\tilde{3}$0 kb long positive-sense genome and belongs to the genus betacoronavirus, which shows 50% and 77.5% genome identity with Middle East respiratory syndrome coronavirus (MERS-CoV) and SARS-CoV, respectively (**??**). The genomic RNAs (gRNAs) of coronaviruses have a 5' cap structure and 3' poly(A) tail, at the 5'-end of which two large open reading frames (ORF1a/1b) encode 16 viral nonstructural proteins (nsps) occupying two-thirds of the genome. The polyprotein 1a/1ab (pp1a/1ab) are translated directly from the genomic RNA through

-1 ribosomal frameshifting (**?**). The 3'- end of CoV genome (one-third of the genome size) contains the genes encoding several main structural proteins including spike protein (S), envelope protein (E), membrane protein (M), nucleocapsid protein (N) and various accessory proteins (**??**).

The CoV genomes have a hallmark process of replication and transcription facilitated by the replication-transcription complex (RTC) with RNA-dependent RNA polymerase (RdRP) activity (**?**), which is more complicated than other types of RNA viruses. The negative strand RNAs are synthesized by RdRP starting from the 3'-end of positive (+)gRNAs, from which continuous synthesis generates full-length complementary negative (-)gRNAs, while discontinuous jumping produces (-)subgenomic RNAs (sgRNAs) with common 5'- and 3'-ends (**?**). Positive-sense progeny gRNAs and sgRNAs are synthesized by using these negative-strand RNA intermediates as templates (**?**). The discontinuous jumping step, called "template switch", is mediated by transcription-regulating sequence (TRS) in the genome body (TRS-B) and in 5'-leader sequence (TRS-L) upstream ORF1ab (**?**), resulting in the fusion of leader-body sequences. Thiel et al. identified eight sgRNAs of SARS- CoV (**?**), while our subsequent study identified ten sgRNAs including two novel sgRNAs (**?**). Recently, more sgRNA variants of HCoV-229E were reported and remained to be characterized (**?**). Kim et al. reported a high-resolution map of the transcriptome and RNA modifications of SARS-CoV-2 in Vero E6 cells (**?**). However, the dynamic landscapes of subgenomes from template switching are unclear for CoV genomes including SARS-CoV-2, and whether the jumping events happen in positive strand synthesis is largely unknown.

TRSs comprise a conserved 6-7 nt core sequence surrounded by variable sequences. Different core TRSs were previously reported including CUAAAC for coronavirus TGEV (**?**), ACGAAC for SARS-CoV (**??**), and CUUUAGA for equine torovirus (**?**). It is hypothesized that the formation of a duplex between TRS-L and downstream TRS-B core sequences determines the template switches (**?**). However, the regulatory features of TRS-like elements in CoVs, including SARS-CoV-2, are not defined yet.

Previous studies, including our work **?**, mainly used RT-PCR coupled with clone-sequencing to characterize the template switch junction, which is low-throughput and unable to detect novel events. Northern blot is generally used to validate specific sizes of sgRNAs, but the detailed sequences are unknown and the resolution is limited. Using next-generation sequencing (NGS) technologies, we have detected the SARS-CoV-2 virus in patients (**?**), assembled the SARS-CoV-2 genome sequence (**?**), and characterized the transcriptomes of patient samples (**?**). NGS provides high-throughput short reads with the capacity to quantify gene expression and characterize splicing junctions, but it is difficult to assemble multiple full-length RNAs. Recently, Viehweger et al. used Nanopore direct RNA sequencing to sequence the full-genome of HCoV-229 without amplification (**?**). We have devised an integrative approach with multi-strategic RNA-seq including NGS and PacBio long read sequencing to construct the high-resolution transcriptional landscape (**?**).

In this study, we employed both NGS and Nanopore direct RNA sequencing techniques to systematically characterize 1) the global and dynamic profiles of template switching events; and 2) the full-length subgenomes of SARS-CoV-2 in two host cell types post-infection at different time points. We further investigated the pairing rules between the upstream and downstream junction sites in those template switches, and found two major modes of RNA-RNA interactions. Moreover, we found that template switch also exists during positive strand synthesis, and thus identified two other large classes of subgenomes. Collectively, our findings provide a global view of the SARS-CoV-2 subgenomes and uncover the molecular basis governing their biogenesis.

2 RESULTS

2.1 Quantitative landscape of template switches in SARS-CoV-2

To explore the global dynamic landscapes of coronaviral subgenomes, we first verified the presence of subgenomes (sgRNAs) in SARS-CoV-2 infected Vero E6 cells using Northern blot (Figure S1). To characterize high-resolution SARS-CoV-2 sgRNAs, we used hybrid poly(A) selected RNA sequencing technologies to analyze local template switching events and construct full sgRNAs simultaneously. Using Poly(A) RNAs enriched from total RNAs ex-

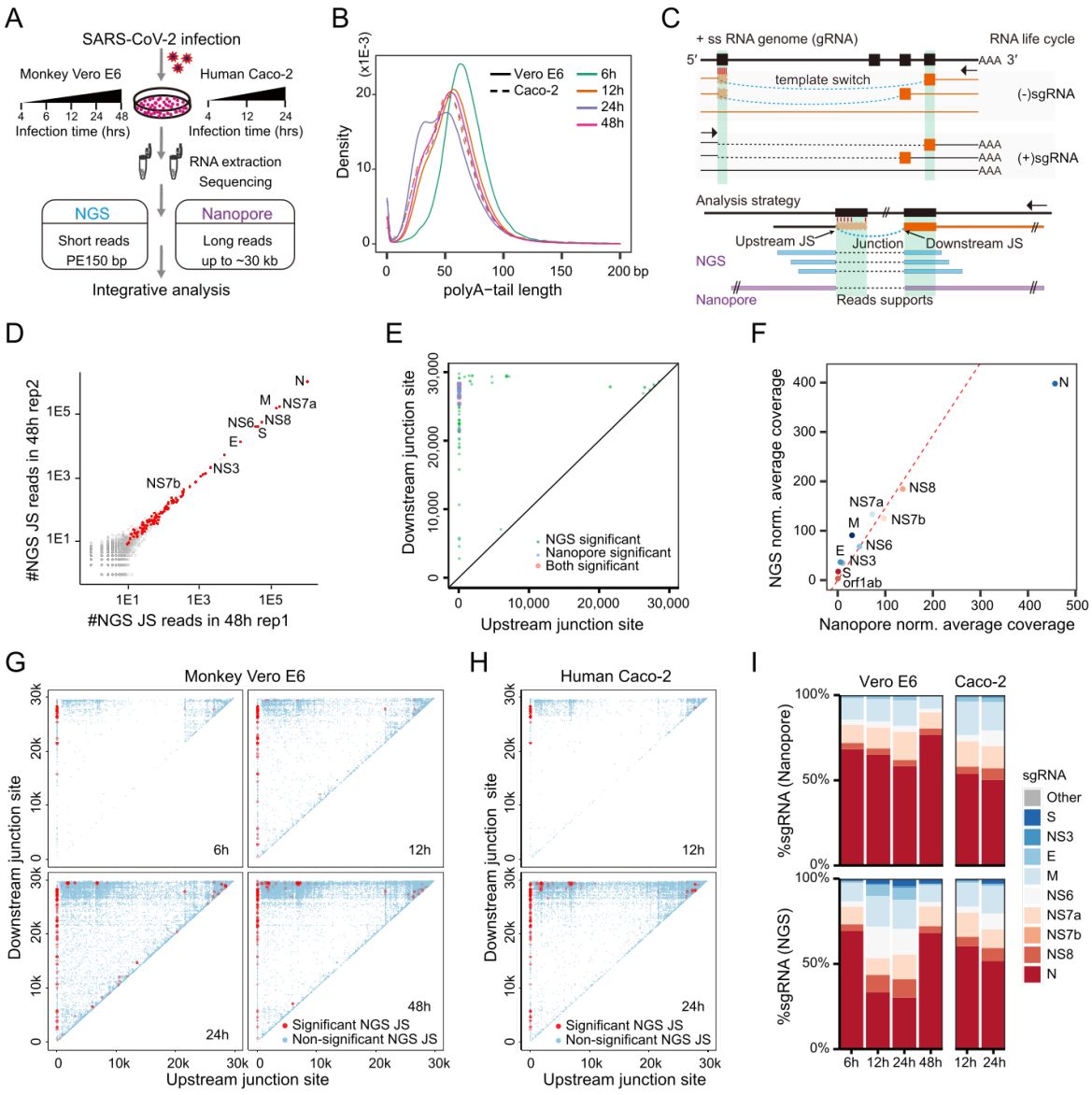

Figure 1: Experimental strategy and analysis for global mapping of template switches

(A) Experimental design for decoding SARS-CoV-2 subgenome dynamics at different time points post-infection in Vero E6 and Caco-2 cells.

(B) Distribution of poly(A)-tail length in Nanopore reads in different samples.

(C) SARS-CoV-2 RNA genome life cycle and the analysis strategy. The template switches are represented by curved dashed lines, and identified by junctions in NGS and Nanopore reads.

(D) Reproducibility between two replicates of NGS data. Each dot represents the reads counts of one junction in replicates 1 (x-axis) and 2 (y-axis). Red points represent the significant junctions identified from statistical analysis.

(E) Global view of NGS-consistent and Nanopore-consistent junction sites in Vero E6 cells 48 h post-infection. Each dot represents a junction linking from the start (x-axis) to the end genomic position (y-axis). NGS-only, Nanopore-only, and both consistent junction sites are represented in green, blue, and red colors, respectively.

(F) Comparison of the signal coverage for each type of sgRNAs between Nanopore and NGS platforms in Vero E6 cells 48 h post-infection.

(G) Global view of NGS-derived junction sites in VeroE6 cells infected with SARS-CoV-2 at 6 h, 12 h, 24 h, and 48 h. Red points represent the statistically significant junction sites.

(H) Same as (G) for Caco-2 data at 12 h and 24 h.

(I) Statistics of sgRNA composition in different samples based on Nanopore (top) and NGS (bottom) reads.

See also Figure S1 and S2.

tracted from SARS-CoV-2 (WIV04, IVCAS 6.7512) infected monkey Vero E6 cells (ATCC number: CRL- 1586) and human Caco-2 cells, we constructed RNA sequencing libraries with duplicates for NGS Illumina and Nanopore MinION platforms, respectively. NGS libraries were sequenced in Pair- End 150 bp mode, while Nanopore libraries were sequenced by direct RNA sequencing (Figure 1A).

To investigate the dynamic landscapes of viral RNAs, we performed the assays at multiple time points post-infection of SARS-CoV-2 in Vero E6 and Caco-2 cells. We chose the two cell types to explore potential differences on sgRNA biogenesis in different hosts from different species with different tissue origins. The ratio of viral reads between Nanopore and NGS are relatively consistent, around 0.1%, 7%, 50% for 4 h, 6 h, and 12 h, respectively. The ratio at 24 h reaches the same level as that at 48 h, 80%-90% in Vero E6 cells (Table S1). The fractions of SARS-CoV-2 reads in Caco-2 samples are 0.02%, 1.2%, and 21% for 4 h, 12 h, and 24 h, smaller than those in Vero E6 samples, respectively, consistent with known low infection rate in Caco-2 cells (Table S1). We found that the Nanopore long reads contain poly(A) tails with median poly(A) length around 50 nt (Figure 1B), similar to that at other time points and consistent with a recent report (**?**). We found that all the Nanopore reads are mapped in the (+) genome, suggesting that (-)gRNA or (-)sgRNAs do not contain poly(A) tails. In Vero E6 cells at 48 h post- infection, the median length of Nanopore reads is 1,248 nt, and 22 reads are mapped to genome with a length close to 30 kb, covering the whole viral genome.

We next developed a toolkit to identify robust template switching events from NGS short reads across the SARS-CoV-2 genome. Template switches generate "jumping" junctions (Figure 1C), similar to the exon-exon junction reads resulting from pre-mRNA splicing. To ensure the correct localization of junctions, we required the sequences in reads flanking the junctions at both sides to have at least 20 nt exact matches to the genome. We first analyzed the Vero E6 samples at 48 h post-infection. To verify the robustness of the junctions, we compared the counts of reads for all observed junctions in replicates 1 and 2, and found that the significant junctions to be highly reproducible (Figure 1D). In total, we identified 45,343 junctions with counts in combined data from two replicates. To further remove potential noises from uneven RNA abundance, we used statistical scoring to remove the effect of local background around both junction sites (Figure S2A- D and STAR Methods), and obtained 100 significant junctions (Figure 1D, red points, and Table S2). To further verify the junctions, we identified 141 significant junctions embedded in the Nanopore long reads using similar statistical methods (Figure S2E and Table S2), 31 junctions of which are significantly overexpressed in our NGS reads (Figure 1E), suggesting the reliability of the predicted template switching events.

The above SARS-CoV-2 junctions that we detected in NGS or Nanopore data have included all of the ten previously identified leader-body fusions in SARS-CoV. We further quantified the expression levels of the junctions with NGS and Nanopore reads, and found that N protein showed the highest level while the expression levels of the genes increased from the 5' to 3' direction of the positive genome (Figure 1F). This indicates that intermediate sgRNAs can serve as templates to generate shorter sgRNAs by further template switch (see multi-switches below). These canonical junctions are highly abundant and represent 57.8% (99,548/172,107) of the total junction counts in full-length Nanopore reads. Beyond canonical junctions, we found many novel leader- body junctions, as well as other types of body-body junctions (Table S2).

We performed a similar global analysis for earlier time points of SARS-CoV-2 infections, and observed an increasing number of junction sites in Vero E6 cells (Figure 1G and S2F) and Caco-2 cells (Figure 1H and S2G), respectively. For each sample, we counted the numbers of leader-body junctions for the sgRNAs termed based on the first annotated gene downstream of the junction. Consistently, the N sgRNA has the highest expression level of junction sites in both Nanopore and NGS data (Figure 1I).

2.2 Diverse types of full-length subgenomes

To identify complete subgenomes, we took the advantages of Nanopore long reads and only considered those reads covering the 5'-end leader sequence and extending to the 3'-end of SARS- CoV-2 genome. Moreover, internal junctions from template switching must be found in all 4 samples of NGS and Nanopore data set from Vero E6 cells 48 h post-infection (Table S2). By this definition, we identified 433 different subgenomes (sgRNAs) in 208

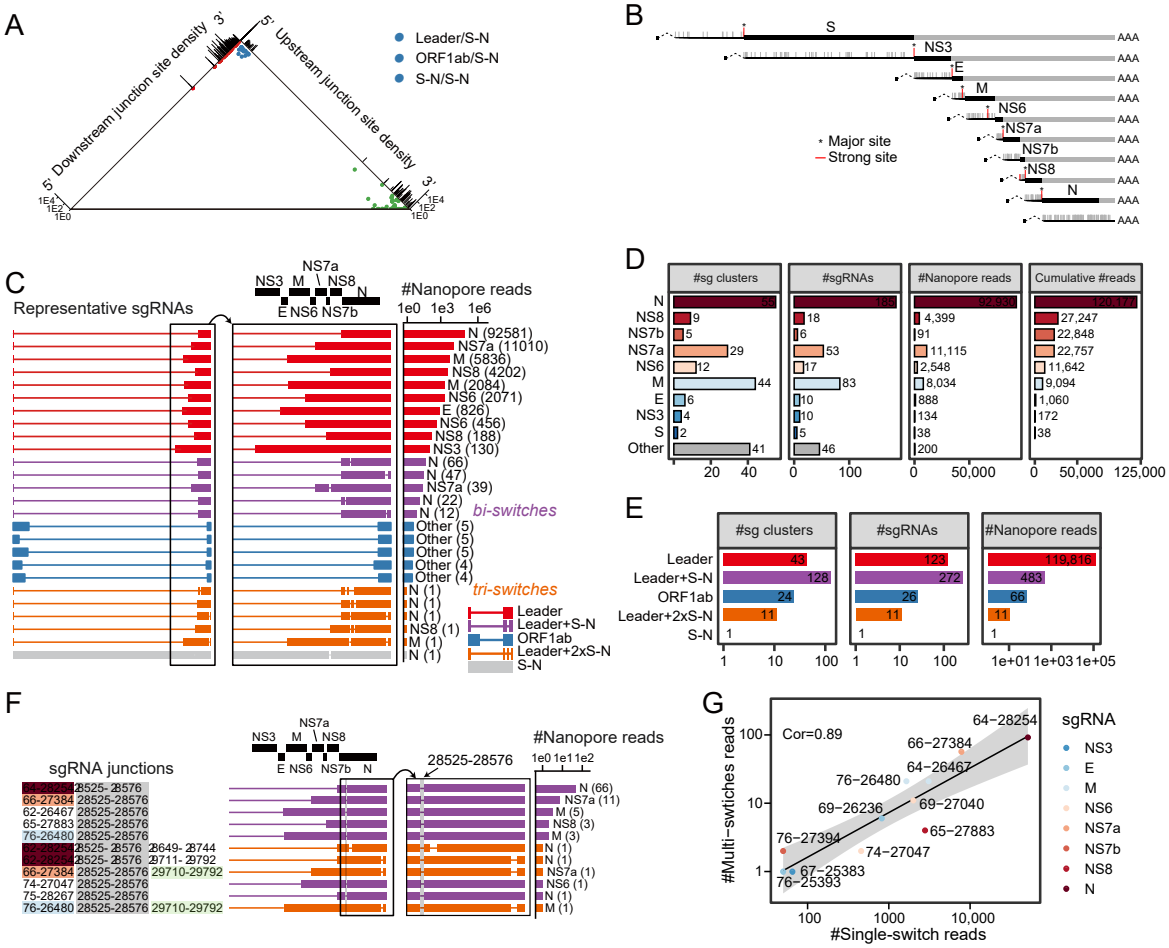

Figure 2: Global landscape of SARS-CoV-2 subgenomes

(A) Global view of consistent template switches in both NGS and Nanopore data. Each template switch is represented as a point by the genomic positions of its upstream and downstream junction sites in the genome. Three types of template switches are shown in different colors (Leader/S-N, red; ORF1ab/S-N, blue; S-N/S-N, green). The densities of upstream and downstream junction sites are shown in the upper right and upper left bar graphs, respectively.

(B) Distribution of the downstream junction sites for Leader-group sgRNAs (48 h Vero E6). The sgRNA names were assigned based on the first annotated gene downstream of the junction. The strong sites (with more than 100 NGS reads support) were marked as red lines, of which the major site (with the largest number) in each sgRNA group was marked with a star symbol.

(C) Subgenome clusters reconstructed from Nanopore long reads. Representative examples for five different types of subgenomes (colored legend) are shown by row in global (left) and zoom-in (middle) views with the number of supporting reads (right). Box and line represent transcribed and skipped regions due to template switches, respectively. Top 10 of Leader-type and top 5 for other types of subgenomes are shown. The label of subgenome was assigned by the first ORF after the template switch.

(D) Statistics for ten subgenome types classified by the first complete ORF in subgenome (Vero E6, 48 h). For each type of sgRNA, the number of clusters, sgRNAs, Nanopore reads, and cumulative count of Nanopore reads containing the ORF are shown respectively. Because S sgRNAs are the longest canonical sgRNAs, they might be less efficiently sequenced by Nanopore.

(E) The number of subgenome (sg) clusters, subgenomes, and subgenome reads for five types of subgenomes.

(F) Examples of multi-switches sgRNAs with common junction (28525-28576). There are 7 bi- switches and 4 tri-switches sgRNAs (numbers of supporting Nanopore reads shown on the right).

(G) Comparison between the number of multi-switches reads vs. that of single-switch reads with specific junction for Leader-type sgRNAs. The Spearman correlation coefficient was labeled. See also Figure S3 and S4.

clusters by merging neighboring upstream or downstream sites within 5 nt, which were subsequently classified into 3 groups (Figure 2A and Table S3): the 1st group is named as Leader/S-N, representing canonical template switches joining leader sequence with downstream genes from S to N; the 2nd group, named as ORF1ab/S-N, represents novel sgRNAs with template switches linking from positions inside ORF1ab with downstream genes from S to N; the 3rd group, named as S-N/S-N, contains novel sgRNAs with internal template switches inside S to N regions. The latter two groups of non- canonical events were also observed in a recent report with different classification rules (Kim, 2020).

The sgRNAs in Leader/S-N group were termed based on the first annotated gene downstream of the junction. The junction sites associated with those sgRNAs are shown in Figure 2B, in which the strong sites (with more than 100 NGS reads support) and their major site (with the largest number) for each sgRNA are marked as red line and star symbol, respectively.

The overall structures of full-length sgRNAs are illustrated in Figure 2C, and the complete map of the core subgenomes for all clusters is shown in Figure S3. The expression level of each subgenome is quantified by the number of corresponding Nanopore long reads. The numbers of sgRNA clusters, events, and Nanopore reads in the Vero E6 48 h sample are shown in Figure 2D, and the numbers for earlier time-points in Vero E6 and Caco-2 cells are shown in Figure S4A, suggesting different expression requirements for these sgRNAs.

Interestingly, we noticed that some sgRNAs have two or more gaps resulting from template switching. This suggests that template switches may occur simultaneously with or independently from other switching events. We identified two groups of multi-switches junctions, Leader+S-N and Leader+2xS-N, which contain Leader/S-N junction and 1 or 2 S-N/S-N junction(s), respectively (Figure 2C, E). The reads numbers for single-switch sgRNAs are much larger than those for multi-switches sgRNAs, and the reads numbers of Leader+S-N sgRNAs are larger than those of Leader+2xS-N sgRNAs, which also points out that the Leader/S-N group is the dominant form of sgRNAs (Figure 2E).

Moreover, different multi-switches sgRNAs share common junctions. As shown in Figure 2F, seven bi-switches and four tri-switches sgRNAs share junction site 28525-28576 with more Nanopore reads support for the bi-switches sgRNAs. Generally, the higher the expression of one parental sgRNA, the larger the counts of multi-switches sgRNAs originating from it. The correlation between the number of multi-switches and that of corresponding single-switch reads is 0.89 (Figure 2G). Consistently, this positive correlation is also evident at different time points post SARS-CoV-2 infection in both Vero E6 and Caco-2 cells (Figure S4B). These data support that sgRNAs resulting from template switching can function as templates for additional template switching events. Collectively, these results showed the complete landscapes of the subgenome structures and their expression levels, providing a useful resource for studying their functions and regulation mechanisms.

2.3 RNA-RNA interaction patterns for bidirectional template switches

To explore the potential rules in governing template switches, we first examined the RNA- RNA base-pairings between potential TRS-L and TRS-B for the 9 canonical sgRNAs observed in almost all the samples (Figure 3A, left). As expected, we found previously known TRS motif (ACGAAC) in the leader sequence (TRS-L), and (ACGAAC/AAGAAC) in the body sequences (anti-TRS-B). Surprisingly, we observed extensive base-pairings with 7-12 consecutive base-pairs beyond the 6 pairs between TRS-L and anti-TRS-B (Figure 3A, middle).

To analyze the base-pairing in a general manner, for one specific sgRNA with a template switch joining upstream and downstream junction sites (JSs), we denote the left and right 20 nt segments flanking upstream JS as UL and UR, and those flanking downstream JS as DL and DR, respectively. We used the RNAhybrid program (**?**) to find the optimal base- pairings with minimum free energy (MFE) between flanking segments, UR with anti-DR (containing TRS-L and anti-TRS-B) compared to UL with anti-DL. As expected, we found that the base-pairings between UR and anti-DR to be stronger than those between UL and anti-DL (Figure 3A, right). Intriguingly, the pairings have a strong tendency to be closer to the JSs for all the 9 sgRNAs (Figure 3A). In analogy to the miRNA-mRNA base-pairing rules (**?**), we defined two base-pairing patterns: seed mode (6 base-pairs in 1-7 nt flanking JSs) and compensatory mode (with additional base-pairs outside seed region), as marked for S and N sgRNAs in Figure 3A.

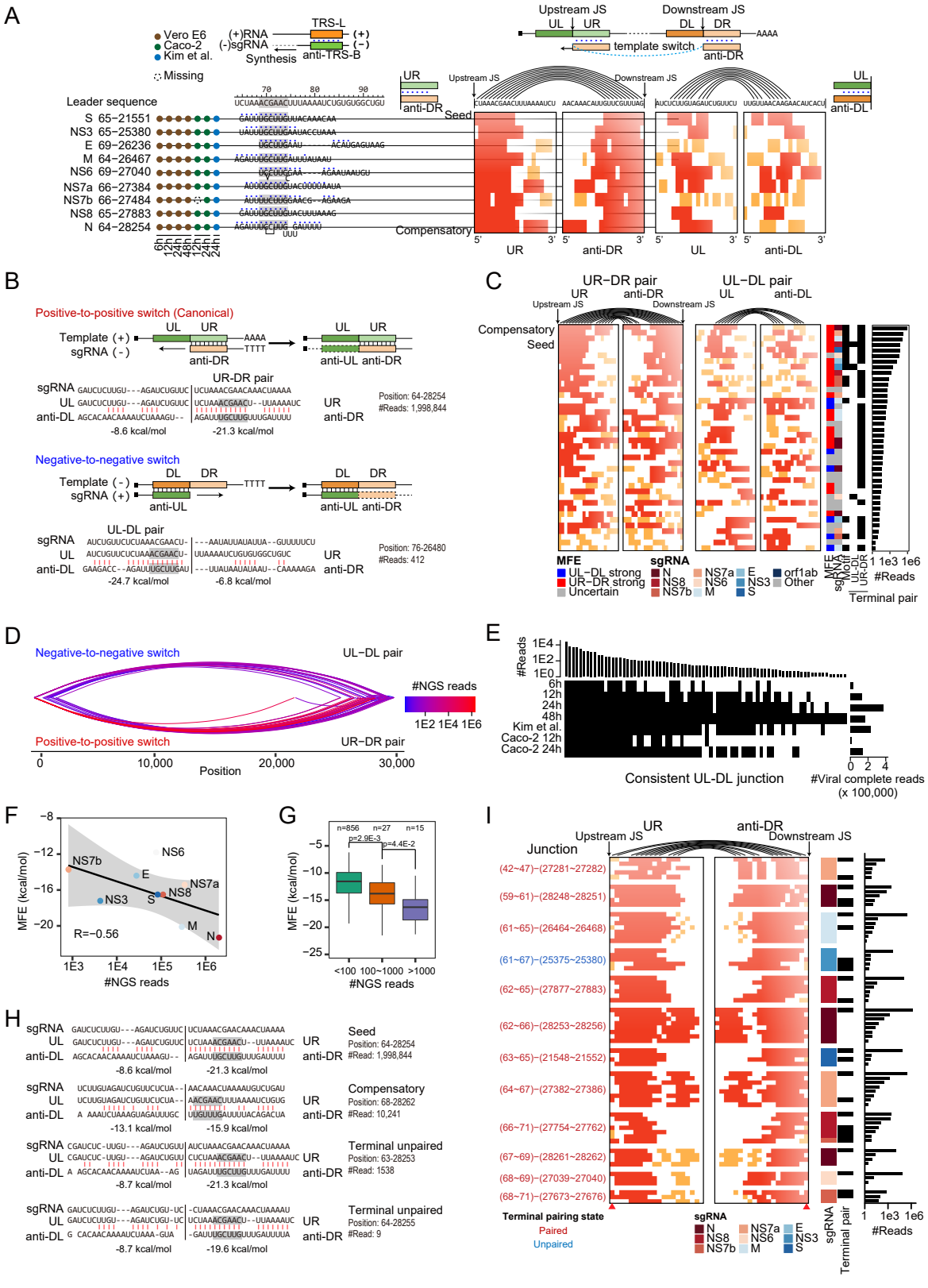

Figure 3: RNA-RNA pairing determinants in template switch efficacy. See subcaptions on next page.

Figure 3: RNA-RNA pairing determinants in template switch efficacy (continued)

(A) The RNA-RNA base-pairing patterns for the 9 canonical SARS-CoV-2 sgRNAs. The presence/absence of sgRNAs in 7 Nanopore samples (by column) are shown on the left with filled circles or empty circle (NS7b sgRNA). Base-pairings between the TRS-L and anti-TRS-B segments are represented by the blue dots, and the TRS motifs are highlighted in gray color. Heatmaps on the right representing the base-pairings between the UR-DR pair (UR and anti- DR) and UL-DL pair (UL and anti-DL). The red or orange square represents paired state, whereas the white square represents unpaired state for the base-pairings between two specific segments flanking the upstream and downstream junction sites for template switching events by row, as illustrated by the arcs linking the predicted base-pairs for the first row of template switch (S sgRNA). Red color indicates consecutive paired state in a 6 nt segment with at least 5 base-pairs.

(B) Illustrations and examples for the positive-to-positive (top, UR-DR pair, canonical) and negative-to-negative (bottom, UL-DL pair) template switch modes. Known TRS motifs are highlighted in gray box. The number of NGS reads in 48 h Vero E6 data and the MFEs between different pairing segments are shown.

(C) Heatmaps as (A) representing the RNA-RNA base-pairings in two modes (UR-DR pair and UL-DL pair) for consistent and core template switch junctions from Vero E6 cells 48 h data. The junctions shown by row are detected in both NGS and Nanopore reads with the largest number of read support in 5 nt windows from the Leader-type sgRNAs. The rows are sorted by the number of supporting NGS reads.

(D) Global view of negative-to-negative (up) and positive-to-positive (bottom) template switches for consistent junctions in both NGS and Nanopore data 48 h post-infection. The number of supporting NGS reads are shown by color-scaled lines.

(E) Consistent UL-DL junctions observed in Nanopore reads from different samples. The junctions are ordered by column according to their total number of reads in all 7 samples (top). The presence of junctions in each sample (by row) are represented by black rectangles. The total numbers of complete Nanopore reads for all samples are shown on the right.

(F) The relationship between the MFE and the number of NGS reads for 9 major Leader-group sgRNA junctions (48 h Vero E6). The Spearman correlation coefficient was labeled.

(G) Boxplots of MFEs for Leader-group sgRNA junctions sub-grouped by the number of NGS reads (48 h Vero E6). The number of junctions in each group and the p-values from one-sided t-test are shown on the top.

(H) Representative examples showing RNA-RNA interaction features affect template switch efficacy. RNA base-pairing pattern, MFE, terminal paired/unpaired status, and number of observed reads are shown for each example.

(I) RNA-RNA base-pairing visualization as (A) between UR and anti-DR segments flanking template switch sites. The columns indicating pairing states of the two terminal bases are marked by the red arrows. Neighboring junctions with similar pairing pattern were grouped together, and the terminal pairing state for the major junctions in each group was marked by color (red for Paired and blue for Unpaired). The corresponding sgRNA, terminal pair state, and the reads numbers (48 h Vero E6) for each junction were shown by row on the right.

See also Figure S5 and S6.

The canonical negative subgenomes are generated through positive-to-positive template switches, and it is assumed that the (+)sgRNAs are then copied from those (-)sgRNAs without template switches. We investigated whether there also would exist negative-to-negative template switch, occurring while generating (+)sgRNAs. In order to discriminate between the two modes of events, we considered that the two different processes are mediated by two different sets of sequences flanking the junction, either UR-DR pair (UR::anti-DR for positive-to-positive mode) or UL-DL pair (anti-UL::DL for negative-to-negative mode), respectively. We thus posited that the relative strengths of RNA-RNA pairing between UR-DR over UL-DL could indicate the correct mode. We tested this hypothesis, as expected, the two modes can be discriminated by the minimum free energy (MFE) between UR-DR and UL-DL pairs (see Method Details). As shown in Figure 3B, for the top example on junction 64-28254, the MFE of UR-DR is much lower than the UL-DL pair, suggesting that the template is switched during (-)sgRNA synthesis. In contrast, for the bottom example on junction 76-26480, the UL-DL pair is much stronger than the UR-DR pair, supporting that the template switch occurs during (+)sgRNA synthesis.

We further checked the two modes on above consistent junction sites found in all samples 48 h post-infection. We found that the junctions with a high-expression level (top junctions) have much lower MFE either in UR-DR or UL-DL mode than those with a low-expression level (bottom junctions) or random junctions w/o known TRS motif pair (Figure S5A). These data constitute further evidence that both modes exist, and can be differentiated by considering the relative positioning of the MFEs for the two pairs.

Moreover, we investigated the detailed RNA-RNA base-pairing patterns that mediate template switch, involving sequences UR and anti-DR, or UL and anti-DL, respectively (20 nt flanking the junction sites (JSs), as shown in Figure 3A-B). We observed that many junction-reads have close junction sites with small shifts, and we used a method based on maximal connected subgraph to group those within 5 nt to clusters, assigning the one with the highest count as the core junction. We then classified the pairings of the Leader-type sgRNAs into 3 groups: UR-DR strong, UL-DL strong, or Uncertain based on the difference of MFEs between the two modes (Figure 3C). There

are about 50%, 17.5%, and 32.5% cases for UR-DR strong, UL-DL strong, and Uncertain groups, respectively. We sorted the sgRNAs by their number of supporting NGS reads, and observed that the base-pairings flanking JSs are stronger for high-expression sgRNAs, especially the top sgRNAs with compensatory or seed pairing mode. The UR-DR and UL-DL pairing patterns for all consistent junctions are shown in Figure S5B.

For the sample in Vero E6 48 h post-infection, we identified 78 and 134 consistent junctions for negative-to-negative and positive-to-positive template switch modes, respectively (Figure 3D). Although the negative-to-negative template switches have lower expression than positive-to- positive ones, they are frequently detected in multiple SARS-CoV-2 infected samples, especially for those with high expression (Figure 3E). The junctions resulting from the two different modes are supported in both NGS RNA-seq and Nanopore-seq data, as shown by the positive correlations between NGS and Nanopore read numbers for two modes of template switches, respectively (Figure S5C). In some cases, the read mapping result may influence the inferred mode (Figure S5D). Empirically, we found that the minimap2 mapping program (**?**) prefers the UR-DR setting, and the number of positive-to-positive template switches during (-)sgRNA synthesis may be overestimated.

We observed several pairing features determining the efficacy of the template switch. The strength of pairing is a key factor, as the MFEs between UR-DR or UL-DL pairs have evident negative correlation with the counts of junction reads supporting template switches for canonical sgRNAs (Figure 3F) and sgRNAs with different expression levels (Figure 3G). As shown in the examples of Figure 3H, the top one with minimum MFE has the largest number of reads. Another key effect is the terminal pairing status, such as the bottom two examples compared to the upper two in Figure 3H. The change from paired to unpaired status in the terminal base decreases the observed number of reads at least 6-fold. We classified Leader-type junctions into subgroups with comparable base-pairing patterns, and checked the impact of the terminal pairing state on read numbers within each subgroup independently. As shown in Figure 3I, the junction with the largest number of reads has paired state for the terminal pairing in 11 out of 12 subgroups.

The global RNA base-pairing flanking JSs have similar patterns for SARS-CoV-2 infected Caco-2 cells (Figure S6A), SARS and MERS infected Calu-3 cells (Figure S6B). The features on MFE and terminal pairing status are also observed in SARS and MERS (Figure S6C-D).

These results provided strong evidence to support that template switch also exists during forward (+)sgRNA synthesis (not only copying from (-)sgRNA), and showed that the RNA-RNA interaction (RRI) strength and pattern are key determinants for the frequency of template switch.

2.4 TRS motif independent RNA-RNA interaction mediated template switch

Previous studies found the transcription-regulating sequence (TRS) is important in the biogenesis of subgenomes, and the same TRS motif in leader and body sequences can form strong base pairings during (-)sgRNA synthesis and mediate template switch (**????**). Can the same TRS motif be found in the SARS- CoV-2 genome? We searched the canonical TRS motif AAGAAC and ACGAAC across the positive and negative SARS-CoV-2 genome, and found 30 and 12 occurrences, respectively (Figure 4A). Unsurprisingly, we did find the TRS motif in 158 out of 7,499 unique junction sites found in two NGS replicates from Vero E6 cells 48 h post SARS-CoV-2 infection, especially the canonical template switching events between TRS-L (top row, upstream JS position) and TRS-B sequences (bottom rows, downstream JS position). Both positive and negative TRS motifs can participate in template switches. However, some TRS motifs do not have template switching events around them (Figure 4A), indicating that the motifs may not be the determinant, or that other features block their effects.

Next, we performed statistical analysis of the TRS occurrences, based on above categorization into 3 subgroups of consistent junctions. We observed that the ORF1ab and S-N groups do not have TRS motif pairs involved in (Figure 4B left). Even for the leader sequence group, only about half of template switching events contain the TRS motif pair (Figure 4B right). As exemplified in Figure 4C, the efficacy of TRS motif ACGAAC depends on the pairing context in comparison with case #1 vs. #5 junctions associated with the NS8 gene, in which the stronger the pairing, the larger the number of reads. In contrast, cases #2-4 have a larger number of reads than #5 although they do not have the TRS motif. Again, we observe that the terminal base-pairing status has a strong impact on template

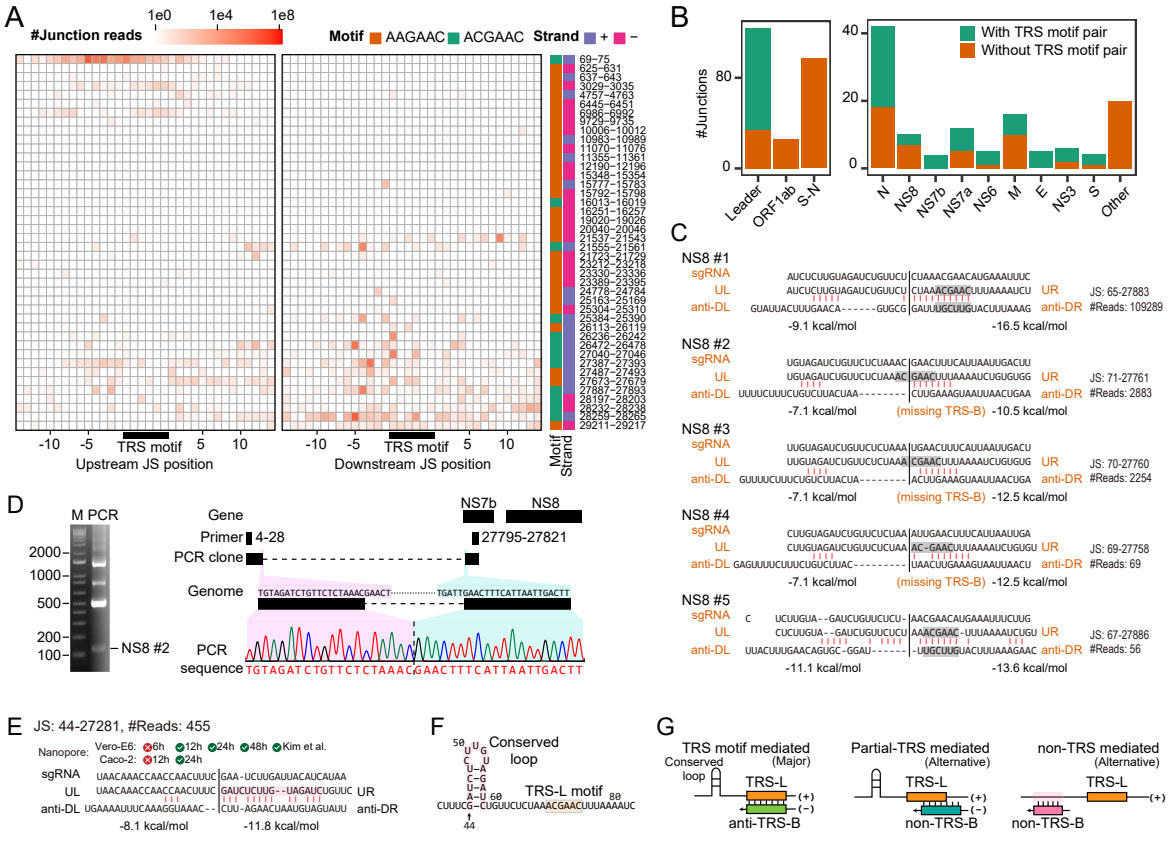

Figure 4: TRS motif independent RNA-RNA interaction in template switch

(A) Heatmap showing the usage of TRS motifs in donor and acceptor sites of template switches. The motif sequences, strand, and genomic positions are annotated on the right side.

(B) Proportion of TRS-mediated template switches in different types of subgenomes.

(C) Representative template switch examples w/o TRS motifs. The TRS motif is marked in gray and number of the junction reads detected for each class of NS8 is shown.

(D) RT-PCR validation for NS8 #2 sgRNA by clone sequencing. The locations of primers, genome sequence, and cloned sequence are shown on the right.

(E) An example of upstream non-TRS mediated Leader-type sgRNA. Number of NGS reads for Vero E6 48 h was shown. The presence of this sgRNA in Nanopore data from different samples are marked.

(F) Illustration of the non-TRS sequence (pink) in a conserved loop for non-TRS mediated Leader-type sgRNA in (E).

(G) Illustration of three types of pairing models mediating template switches for Leader-type sgRNAs.

switch efficacy. We further validated the NS8 #2 junction by RT-PCR and clone sequencing (Figure 4D), resulting from RNA base-pairing without TRS motif around downstream junction site.

Interestingly, we detected one special Leader-type sgRNA without TRS motif around neither upstream nor downstream junction sites, although it has low expression but is found in multiple samples based on our and published Nanopore data sets (Figure 4E). The bases in UR segment pairing with anti-DR are in the conserved loop (Rfam v14.2 RF03117) (**?**) upstream of the TRS-L motif (Figure 4F), and we hypothesize the RNA secondary structure of this loop may lead to the low frequency of this type of sgRNA. For the Leader-type sgRNAs, the major mode of template switches involves TRS motif pair plus additional base-pairs, while alternative modes are partial-TRS or non-TRS mediated base-pairings (Figure 4G).

In sum, many template switches do not need the TRS motif, and RNA-RNA interactions mediated proximity of

two long-range junction sites may be the key determinants in the coronaviral subgenome biogenesis.

2.5 Weak but extensive fusions between ORF1ab and N ORF RNA regions

In addition to the leader sequence participating template switches, the other group of long- range template switches unexpectedly occurs between upstream site in ORF1ab and the downstream site inside the N ORF RNA region. Contrasting with the case of the leader sequence pattern, where the junction start position is concentrated in one site around position 64 in the SARS-CoV-2 genome, the junction start position in ORF1ab group has a broad distribution and two peaks towards the 5'-end of ORF1ab in Vero E6 cells at different time-points (Figure 5A). We further analyzed the viral RNA profile in Vero E6 cells from published data (**?**), showing a similar trend (Figure S7A). Moreover, we also observed similar patterns including Leader- and ORF1ab-type sgRNAs in Caco-2 cells at two time-points (Figure S7B). Interestingly, the ORF1ab- type sgRNAs were detected in another coronavirus HCoV-229E infection (WT), but not in its mutant strain infection (SL2) by analyzing the published Nanopore data (**?**), as shown in Figure S7C. The different profiles between WT and SL2 may be caused by the loss of the conserved loop in SL2 (Figure S7D). We validated the presence of ORF1ab sgRNAs by RT- PCR and clone sequencing. The broad band indicates diverse types of ORF1b-type junctions between the designed primers (Figure 5B), consistent with the Nanopore reads (Figure 5C). These data together suggest the biogenesis of ORF1ab-type sgRNAs is widespread and regulated.

Interestingly, the ending position also has a broad distribution inside N RNA region, as exemplified by the Vero E6 data 48 h post-infection (Figure S7E). As summarized by the number of Nanopore long reads, the junctions starting in leader sequence have 168,208 reads over 2,339 unique junctions, while the junctions starting in ORF1ab have 15,774 reads but over 14,325 different variable junctions (Figure S7F). The sgRNA types were assigned according to the last protein upstream of the junction in ORF1ab (Figure 5D), and the nsp1 and nsp2 sgRNAs are the major ones in ORF1ab-type (Figure 5E).

To investigate the potential functions of these non-canonical sgRNAs, we analyzed a recently published Ribo-seq data using ribosome footprints to infer SARS-CoV-2 coding capacity (**?**). We counted the number of junction sites with Ribo-seq read support (Figure S7G) by sgRNA groups we defined in Figure 2A. There are 153 non-canonical sgRNA junctions showing ribosome binding evidence, including 11 ORF1ab derived JSs (Figure 5F). Furthermore, we have downloaded and analyzed recently published Mass Spectrometry (MS) data from the ProteomeXchange database using a stringent pipeline (Figure S7H). We found that two kinds of ORF1ab-type sgRNA generated peptides that span the template switching junction site in two or more samples (Figure 5F-G and Figure S7I). These results suggest that some of the ORF1ab-type sgRNAs have the ability to generate new peptides/proteins, or occupy host translational machinery. Complete sets of predicted ORFs for these non-canonical sgRNAs are included in the Table S3, which motivates further studies to validate their biogenesis and biological functions.

3 DISCUSSION

Understanding the process and mechanism of SARS-CoV-2 sgRNA biogenesis is crucial for identifying potential anti-viral drug targets. SARS-CoV genomes are known to generate several subgenomes through template switch during negative genome synthesis mediated by interactions between leader and body TRS elements. Our results revealed diverse modes of subgenome biogenesis (Figure S7J), and discovered several key determinants of RNA-RNA interactions affecting the template switching efficacy (Figure S7K).

Wu and Brian have shown that transfected BCoV subgenome can function as template for negative-strand synthesis (**?**), and our data further confirm this mode by identifying many sgRNAs with multi-switches events in SARS-CoV-2 infected cells. Besides known positive-to-positive template switch, we discovered negative-to-negative template switch during (+)-strand synthesis. Interestingly, Wu and Brain reported a special sgRNA as ambisense chimeras resulting from in trans positive-to-negative-strand template switching in bovine coronavirus (**?**), and more complex modes of sgRNA synthesis merit further investigation.

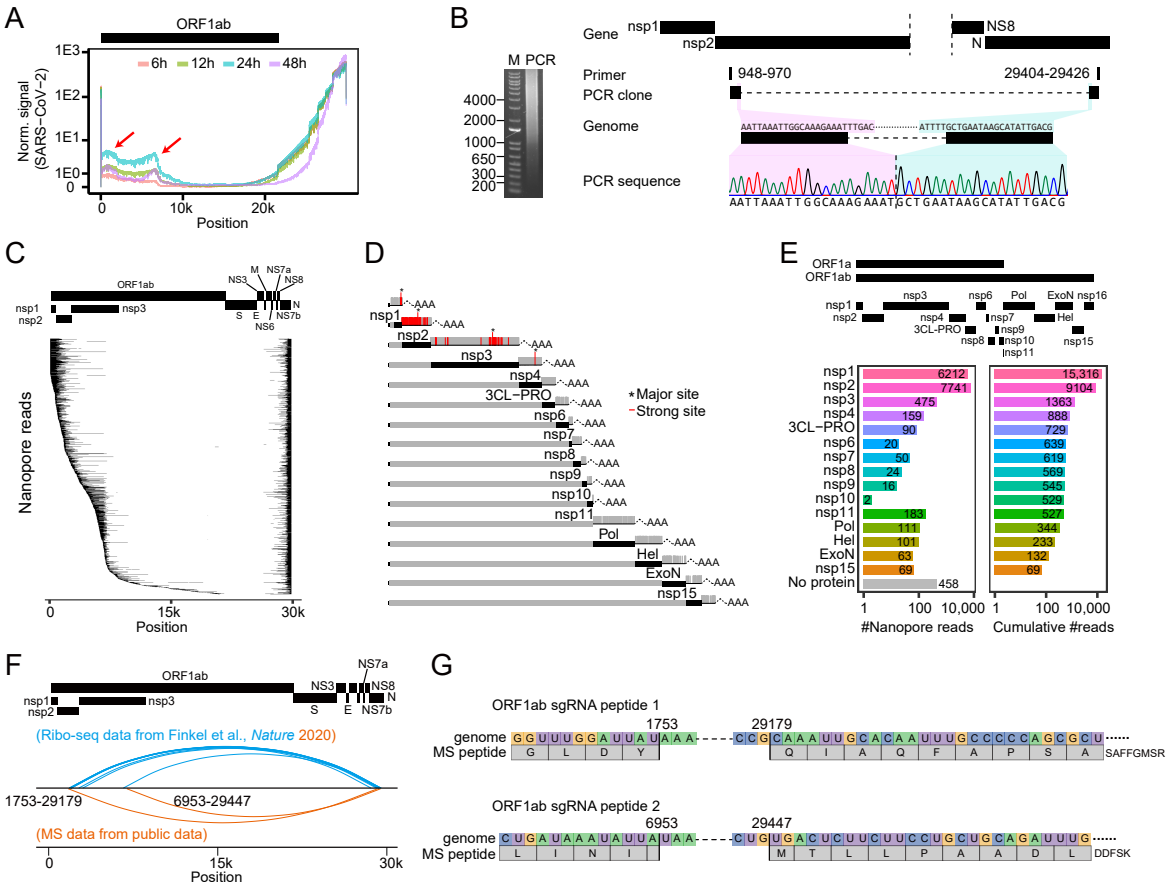

Figure 5: Extensive fusion subgenomes between ORF1ab and N RNA region

(A) Nanopore reads coverage profiles across the SARS-CoV-2 genome for different time-points post-infection in Vero E6 cells. The arrows in ORF1ab region mark two obvious loci with abrupt changes indicating template switch junctions.

(B) RT-PCR validation for ORF1ab sgRNA by clone sequencing. The diffusing band indicates diverse types of junctions between the designed primers. The locations of primers, genome sequence, and cloned sequence are shown on the right.

(C) Global view of Nanopore reads associated with ORF1ab-mediated long-range template switches (Vero E6, 48 h).

(D) Distribution of the upstream junction site for ORF1ab sgRNAs (Vero E6, 48 h). The sgRNA types were assigned according to the last protein upstream of the junction in ORF1ab. The strong sites were labeled as a red line, whereas the major sites were marked with star symbol.

(E) The counts (left) and cumulative counts (right) of nanopore reads assigned to the 16 types of ORF1ab sgRNAs (Vero E6, 48 h).

(F) Graphic illustration of the ORF1ab-type JSs covered by the SARS-CoV-2 Ribo-seq reads (blue curves) and MS peptides (orange curves).

(G) Examples of ORF1ab-type sgRNA derived peptides spanning the sgRNA junctions from MS data.

See also Figure S7.

Previous studies hypothesized a three-step model of coronavirus transcription, including initiation pre-complex formation, base-pairing scanning by the pre-complex, and template switch (**?**). Our results provide detailed features of the base-pairing scanning for efficient template switching. It is also reported that the formation of local secondary structures or high- order RNA structures downstream of switching sites are important to pause continuous transcription, and thus to promote switching (**??**). Co-variational mutation analysis of multiple genomes found conserved structural RNA elements in the terminal regions of Alphacoronavirus genomes (**?**). However, at the moment, similar analyses are hindered by the low mutation rate observed in SAR-CoV-2, or would require the consideration of distantly related genomes, at the risk of overlooking specific features of SAR-CoV-2. High-throughput RNA structural profiling methods such as SHAPE-MaP (**?**) and PARIS (**?**) together would be useful for decoding the mechanism from the perspective of the RNA structure and RNA-RNA interaction network. Meanwhile, several RNA binding proteins (RBPs) have been proposed to participate in the biogenesis of subgenomes (**?**). Candidate RBPs can be systematically identified from specific RNA-capture assays followed by mass spectrometry. Functional assays such as RBP knockout, and in vivo binding assays such as CLIP-seq (**?**), could be used to validate the roles and mechanisms of regulatory RBPs. The mechanisms for why viral RdRP pauses and jumps are still unclear and need further investigation.

It is worth noting that we did not capture the negative strand intermediates of CoV genomic and subgenomic RNAs, as they may lack the poly(A) tails, on which our purification or sequencing methods depend. To characterize the ratios between negative and positive subgenomes, we need to use a poly(A)- RNA sequencing method to sequence and quantify the negative subgenomes. Furthermore, time-course nascent RNA-seq is a promising strategy and could be the object of future studies to depict the dynamic maps of RNAs during viral genome replication and transcription.

Our results provide a quantitative high-resolution map of subgenome structures. What parts of these subgenomes are translated? We may characterize the coding regions at the sgRNA level by using Nanopore direct polysome profiling to obtain ribosome footprints. We also discovered many non-canonical template switching events, including potential defective ones generating truncated mRNAs of the N protein. Those subgenomic RNAs are similar to the defective interfering RNAs (DI-RNAs) found in some RNA viruses, including HCoV-229E as recently reported (**??**).

3.1 Limitations

The findings in this study are based on our data from cell lines (human Caco-2 and monkey Vero E6) and need further confirmation in infected primary cells to investigate the sgRNA biogenesis in tissues under physiological conditions. The regulatory features we reported to govern template switches are predicted by computational RNA-RNA base-pairing analysis and could be verified through further experimental studies. Unfortunately, the state biosafety law, along with obvious ethical and safety concerns, prevents us from performing mutation-rescue experiments on live viruses, due to the potential risk of creating artificial highly pathogenic coronavirus.

ACKNOWLEDGEMENTS

The authors thank the help from Dr. Zhihong Hu's lab in the Wuhan Institute of Virology. This study was supported by grants from the China NSFC projects (32041007, 31922039, 31871316, and 81672008), the National Key R&D Program of China (2017YFA0504400), National Science and Technology Major Project (2018YFA0900801), and Special Fund for COVID-19 Research of Wuhan University. We are grateful to Beijing Taikang Yicai Foundation for their support. Part of computation in this work was done in the Supercomputing Center of Wuhan University.

AUTHOR CONTRIBUTIONS

Y.Z., Y.C., and K.L. conceived the study. A.J., G.N.L., J.P.F., D.G., M.S., F.Y.L., Q.H.Z., and M.G. performed the experiments. D.H.W., Y.P., S.W., Y.Z., H.X.Y., S.P.L., and X.Q.L. analyzed the sequencing data. D.H.W. and K.W. analyzed the MS data. Q.Y.L. and J.P.F. designed primers and probes. X.L.Y. and Z.L.S. provided virus infected

cell lines. Y.Z., Y.C., K.L., D.H.W., and M.Z.C wrote the manuscript with the input from all authors. All authors discussed the results and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

REFERENCES

Bartel, D.P. (2018). Metazoan MicroRNAs. Cell 173, 20–51.

Chen, L., Liu, W., Zhang, Q., Xu, K., Ye, G., Wu, W., Sun, Z., Liu, F., Wu, K., Zhong, B., et al. (2020a). RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg. Microbes Infect. 9, 313–319.

Chen, Y., Liu, Q., and Guo, D. (2020b). Emerging coronaviruses: genome structure, replication, and pathogenesis. J. Med. Virol. 92, 418–423.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423.

Davidson, A.D., Williamson, M.K., Lewis, S., Shoemark, D., Carroll, M.W., Heesom, K.J., Zambon, M., Ellis, J., Lewis, P.A., Hiscox, J.A., et al. (2020). Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. Genome Med. 12, 68.

De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., and Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34, 2666–2669.

Deutsch, E.W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., Campbell, D.S., Bernal-Llinares, M., Okuda, S., Kawano, S., et al. (2017). The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic Acids Res. 45, D1100–D1106.

Di, H., Jr, J.C.M., Morantz, E.K., Tang, H.-Y., Graham, R.L., Baric, R.S., and Brinton, M.A. (2017). Expanded subgenomic mRNA transcriptome and coding capacity of a nidovirus. Proc. Natl. Acad. Sci. E8895–E8904.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID- 19 in real time. Lancet Infect. Dis.

Finkel, Y., Mizrahi, O., Nachshon, A., Weingarten-Gabbay, S., Morgenstern, D., Yahalom-Ronen, Y., Tamir, H., Achdout, H., Stein, D., Israeli, O., et al. (2020). The coding capacity of SARS-CoV- 2. Nature.

Grenga, L., Gallais, F., Pible, O., Gaillard, J.-C., Gouveia, D., Batina, H., Bazaline, N., Ruat, S., Culotta, K., Miotello, G., et al. (2020). Shotgun proteomics analysis of SARS-CoV-2-infected cells and how it can optimize whole viral particle antigen production for vaccines. Emerg. Microbes Infect. 9, 1712–1721.

Hussain, S., Pan, J., Chen, Y., Yang, Y., Xu, J., Peng, Y., Wu, Y., Li, Z., Zhu, Y., and Tien, P. (2005). Identification of Novel Subgenomic RNAs and Noncanonical Transcription Initiation Signals of Severe Acute Respiratory Syndrome Coronavirus. J. Virol. 79, 5288–5295.

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., and Petrov, A.I. (2018). Rfam 13.0: shifting to a genome-centric resource for non- coding RNA families. Nucleic Acids Res. 46, D335–D342.

Kim, D. (2020). The Architecture of SARS-CoV-2 Transcriptome. Cell 181, 914–921.

Köster, J., and Rahmann, S. (2018). Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 34, 3600.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094– 3100.

Liu, W., Zhang, Q., Chen, J., Xiang, R., Song, H., Shu, S., Chen, L., Liang, L., Zhou, J., You, L., et al. (2020). Detection of Covid-19 in Children in Early January 2020 in Wuhan, China. N. Engl. J. Med. 382, 1370–1371.

Loman, N.J., Quick, J., and Simpson, J.T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat. Methods 12, 733–735.

Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., Mattick, J.S., et al. (2016). RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell 165, 1267–1279.

Madhugiri, R., Fricke, M., Marz, M., and Ziebuhr, J. (2014). RNA structure analysis of alphacoronavirus terminal genome regions. Virus Res. 194, 76–89.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal 17, 10.

Mateos-Gomez, P.A., Morales, L., Zuniga, S., Enjuanes, L., and Sola, I. (2013). Long-Distance RNA-RNA Interactions in the Coronavirus Genome Form High-Order Structures Promoting Discontinuous RNA Synthesis during Transcription. J. Virol. 87, 177–186.

National Genomics Data Center Members and Partners (2020). Database Resources of the National Genomics Data Center in 2020. Nucleic Acids Res. 48, D24–D33.

Nicholson, B.L., and White, K.A. (2014). Functional long-range RNA–RNA interactions in positive-strand RNA viruses. Nat. Rev. Microbiol. 12, 493–504.

Pasternak, A.O. (2001). Sequence requirements for RNA strand transfer during nidovirus discontinuous subgenomic RNA synthesis. EMBO J. 20, 7220–7228.

Pathak, K.B., and Nagy, P.D. (2009). Defective Interfering RNAs: Foes of Viruses and Friends of Virologists. Viruses 1, 895–919.

Perlman, S., and Netland, J. (2009). Coronaviruses post-SARS: update on replication and pathogenesis. Nat. Rev. Microbiol. 7, 439–450.

Rehmsmeier, M. (2004). Fast and effective prediction of microRNA/target duplexes. RNA 10, 1507–1517.

Smola, M.J., Rice, G.M., Busan, S., Siegfried, N.A., and Weeks, K.M. (2015). Selective 2'- hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nat. Protoc. 10, 1643–1669.

Snijder, E.J., Decroly, E., and Ziebuhr, J. (2016). The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. Adv. Virus Res. 96, 59.

Sola, I., Mateos-Gomez, P.A., Almazan, F., Zuñiga, S., and Enjuanes, L. (2011). RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. RNA Biol. 8, 237–248.

Sola, I., Almazán, F., Zúniga, S., and Enjuanes, L. (2015). Continuous and Discontinuous RNA Synthesis in Coronaviruses. Annu Rev Virol 2, 265–288.

Stewart, H., Brown, K., Dinan, A.M., Irigoyen, N., Snijder, E.J., and Firth, A.E. (2018). Transcriptional and Translational Landscape of Equine Torovirus. J. Virol. 92, 24.

Thiel, V., Ivanov, K.A., Putics, A., Hertzig, T., and Ziebuhr, J. (2003). Mechanisms and enzymes involved in SARS coronavirus genome expression. J. Gen. Virol. 84, 2305–2315.

Tyanova, S., Temu, T., and Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat. Protoc. 11, 2301–2319.

Viehweger, A., Krautwurst, S., Lamkiewicz, K., Madhugiri, R., Ziebuhr, J., Hölzer, M., and Marz, M. (2019). Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. Genome Res. 29, 1545–1554.

Wang, K., Wang, D., and Zheng, X. (2019). Multi-strategic RNA-seq analysis reveals a high-resolution transcriptional landscape in cotton. Nat. Commun. 10, 4714.

Wu, H.-Y., and Brian, D.A. (2007). 5'-Proximal Hot Spot for an Inducible Positive-to-Negative- Strand Template Switch by Coronavirus RNA-Dependent RNA Polymerase. J. Virol. 81, 3206– 3215.

Wu, H.Y., and Brian, D.A. (2010). Subgenomic messenger RNA amplification in coronaviruses. Proc. Natl. Acad. Sci. U. S. A. 107, 12257–12262.

Xiong, Y., Liu, Y., Cao, L., Wang, D., Guo, M., Jiang, A., Guo, D., Hu, W., Yang, J., Tang, Z., et al. (2020). Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in COVID-19 patients. Emerg. Microbes Infect. 9, 761–770.

Xue, Y., Zhou, Y., Wu, T., Zhu, T., Ji, X., Kwon, Y.-S., Zhang, C., Yeo, G., Black, D.L., Sun, H., et al. (2009). Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping. Mol. Cell 36, 996–1006.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270–273.

Zuniga, S., Sola, I., Alonso, S., and Enjuanes, L. (2004). Sequence Motifs Involved in the Regulation of Discontinuous Coronavirus Subgenomic RNA Synthesis. J. Virol. 78, 980–994.

4 METHODS

4.1 RESOURCE AVAILABILITY

4.1.1 Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Yu Zhou (yu.zhou@whu.edu.cn).

4.1.2 Materials Availability

This study did not generate unique reagents.

4.1.3 Data and Code Availability

The raw sequencing data from this study are deposited in the Genome Sequence Archive in BIG Data Center (https://bigd.big.ac.cn/) (National Genomics Data Center Members and Partners, 2020), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under the accession number: CRA002508 for Vero E6 data and HRA000412 for Caco-2 data under project PRJCA002477. The source codes for all the analysis including workflows in Snakemake (**?**) and scripts in Python and R are available on GitHub at

https://github.com/zhouyulab/cov2sg/

Additional Supplemental Items are available from Mendeley Data at

http://dx.doi.org/10.17632/mt8vm22bbj.1

4.2 EXPERIMENTAL MODEL AND SUBJECT DETAILS

4.2.1 Cell Culture and Virus Infections

African green monkey kidney (Vero E6) cells and human colorectal adenocarcinoma (Caco- 2) cells were grown and maintained in Dulbecco's modified Eagle medium (Gibco Invitrogen Corp.) supplemented with 10% heat-inactivated fetal bovine serum (Gibco Invitrogen Corp.) and 1% of penicillin and streptomycin (Gibco Invitrogen Corp.) at 37°C in an incubator with 5% CO2. Cells were infected at a multiplicity of infection of 0.1 with plaque purified SARS-CoV-2 virus 27 from Vero E6 cells (WIV04, IVCAS 6.7512) provided by Dr. Zhengli-Li Shi's lab in the Wuhan Institute of Virology (**?**).

4.3 METHOD DETAILS

4.3.1 Northern blotting and simulation

The total RNAs from SARS-CoV-2 infected Vero E6 cells were extracted by using Trizol (Invitrogen) according to the manufacturer's instructions. 5 μg of extracted RNA and an RNA ladder (Millennium Markers-Formamide, Ambion) were fractionated in a 1.2% denaturing agarose gel containing 2.2 M formaldehyde with 1x MOPS (3-(N-morpholino) propanesulfonic acid) buffer for 3 hours at 100 V. After overnight capillary transfer to an Hybond-N+ membrane (Amersham, GE Healthcare) and UV cross-linking of the transferred RNA to the membrane, the membrane was prehybridized in DIG Easy Hyb buffer (Roche) at 68 °C for 1 hour, then probed at 68°C for 9 hours with DIG-labeled strand-specific denatured RNA probes according to the protocol of the manufacturer (Roche). The membrane was then washed with a low-stringency buffer containing 2× SSC plus 0.1% SDS at room temperature followed by a wash with a high- stringency buffer containing 0.1× SSC plus 0.1% SDS at 68 °C. Then, the membrane was incubated with block buffer for 30 min at RT, with shaking, and then incubated with the DIG- antibody diluted in block buffer (1:10000) for 30 min, with shaking. The signals were detected with NBT/BCIP stock solution (Roche) using Fujifilm LAS-4000 Super CCD Remote Control Science Imaging System. Furthermore, the RNA ladder on the exposed membrane was stained with methylene blue (wash the membrane for 10 min in 3% HAc, stain for 30sec - 1 min with 0.04% methyleneblue/0.5 M Na-acetate pH5.2, and distain with nuclease-free water until the background is nearly white) to compare the sizes of the target bands. The negative probe, complementary to the 3' end (positions 29090 to 29870) of SARS-CoV-2 positive genome, was used to detect positive-strand subgenomic RNAs (sgRNAs). The positive probe, complementary to the negative probe, was used to detect minus-strand sgRNAs.

In order to compare the Nanopore sequencing data with Northern blot, we simulated Northern image according to the sequence lengths of Nanopore long-reads based on the following logarithmic relationship between molecular weight and mobility: $lg(M) = -b\,m + k$. Where M represents molecular weight (RNA length) and m represents mobility. The bands are generated based on the counts for specific lengths using the density function in R with parameters: $n = 20$ and $bw = 0.01$. The scale_alpha_continuous function in ggplot2 package was used to simulate the low exposure (LE) and over exposure (OE) conditions.

4.3.2 Reverse transcription

Total RNA from SARS-CoV-2-infected Vero E6 and Caco-2 cells was extracted by using TRIzol (Invitrogen) followed by DNaseI (Takara) treatment. Reverse transcription (SuperScript III Reverse Transcriptase [Invitrogen]) was done with virus-specific RT primers.

4.3.3 Poly(A) RNA sequencing

PolyA RNAs were isolated from 1 μg total RNA by using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB) for rRNA depletion. Strand-specific RNA-seq libraries were performed using NEBNext Ultra II Directional RNA Library Prep Kit (NEB), and 150-300 bp insert size was selected following the manufacturer's instructions. PolyA RNA-seq was performed on NovaSeq 6000 System (Illumina).

4.3.4 Nanopore direct RNA sequencing

For Nanopore sequencing, 1 μg total RNA was used for the library construction following the manufacturer's instructions of the Oxford Nanopore Direct RNA Sequencing protocol (SQK- RNA002). PolyA RNAs were ligated to double-strand RT adaptor (RTA) with oligo(dT) sticky end by T4 Quick DNA ligase (NEB) followed by SuperScript III (Invitrogen) mediated reverse transcription for 30 min. RNA/DNA hybrids were recovered by Agencourt RNAClean XP beads and ligated to Nanopore sequencing RNA adaptor (RMX). 1 μl SUPERase-In RNase inhibitor (Invitrogen, 20 U/μl) was added to both ligation steps. The Direct RNA-seq library was recovered by Agencourt RNAClean XP beads and loaded on FLO-MIN106D flow cell after priming followed by a 48-hour sequencing run on MinION device (Oxford Nanopore Technologies).

4.4 QUANTIFICATION AND STATISTICAL ANALYSIS

4.4.1 Published data collection

Public RNA-seq data were downloaded from NCBI SRA database, and public Nanopore data were downloaded from Open Science Framework (OSF). The accession numbers and data source are described in Supplementary Table 1. The conserved 5'UTR secondary structures for SARS- CoV-2 and HCoV-229E viruses were retrieved from the Rfam database (**?**) with accession number RF03117 (Betacoronavirus) and RF03116 (Alphacoronavirus), respectively. The information of Ribo-seq reads covering sgRNA JSs was downloaded from the Supplementary Table 1 of recent study (**?**). The published Mass Spectrometry data sets were downloaded from ProteomeXchange database (**?**) under identifiers PXD018241 (**?**), PXD018594 (**?**), and PXD021120.

4.4.2 Mapping of NGS RNA-seq data

The adaptors in raw reads were removed using cutadapt (v2.5) program (**?**). After filtering out potential ribosomal RNAs, the clean reads were firstly mapped to the host genome (Vero E6: Chlorocebus sabaeus Ensembl v99; Caco-2 and Calu-3: human hg38) using STAR (v2.7.2b) program (**?**) with parameters "–sjdbScore 1 –outFilterMultimapNmax 20 –outFilterMismatchNmax 999 –outFilterMismatchNoverReadLmax 0.04 –alignIntronMin 20 –alignIntronMax 1000000 –alignMatesGapMax 1000000 –alignSJoverhangMin 8 – alignSJDBoverhangMin 1". The unmapped reads were then mapped to the virus genome (SARS- CoV-2: WIV04, NCBI accession number MN996528; SARS: NCBI accession number NC_004718.3; MERS: NCBI accession number NC_038294.1) using STAR with customized parameters to alleviate the penalty on non-canonical splicing gaps (–outFilterMultimapNmax 1 - -alignSJoverhangMin 8 –outSJfilterOverhangMin 8 8 8 8 –outSJfilterCountUniqueMin 3 3 3 3 - -outSJfilterCountTotalMin 3 3 3 3 –outSJfilterDistToOtherSJmin 0 0 0 0 –scoreGap -4 – scoreGapNoncan -4 –scoreGapATAC -4 –alignIntronMax 30000 –alignMatesGapMax 30000 – alignSJstitchMismatchNmax -1 -1 -1 -1). The uniquely mapped reads were kept for further analysis.

4.4.3 Processing of Nanopore data

The base-calling of raw data was done by guppy v3.4.5 with dual Tesla V100 using the HAC model. The quality control was performed with NanoPlot v1.28.4 (**?**). Poly(A) 31 tails were detected by nanopolish (v0.12.3) (**?**). The reads were mapped to the reference genomes (host and virus genomes combined) using minimap2 (v2.17) with customized parameters (-ax splice -un -k14 –no-end-flt –secondary=no) (**?**). The reads belonging to the viral genome are used in further analysis.

4.4.4 Identification of significant junction sites

Junction site (JS) candidates from template switches were identified by finding the gaps in reads with flanking sequences exactly matched over 20 nt at each side. To get significant junction sites, three rules were used in filtering JS candidates. The top 2.5% of highest relative expressed junctions (#JS reads / #total JS reads) were selected firstly to remove junctions with a small number of supporting reads, as shown in Figure S2A. Considering the extremely unbalanced read depth across the virus genome, the local relative abundance score (S) was used to remove bias in high abundant regions. The S is defined as the geometric mean of Supstream and Sdownstream, which are computed as the ratio of number of reads over nearby signal (+/- 100 nt) for upstream and downstream positions, respectively. The normalized junction counts (N) is defined as

$$\mathrm{N} = \frac{\#\mathrm{JS}}{\sqrt{\mathrm{Supstream} \times \mathrm{Sdownstream}}}.$$

Due to big differences in the distribution of S between high- and low- expressed junction sites, the threshold of S was determined by treating the low-expression component as a control group and limiting the false positive rate (alpha) equals to 0.01 (Figure S2B). Finally, junctions with a gap shorter than 100 nt were removed. The JSs passing these filters were called significant in one data set. The junctions which are significant in both replicates and the merged datasets are defined as the set of significant junction sites in this study.

For Nanopore data derived junctions, similar methods were used except the supporting reads number was required to be larger than two (Figure S2C-D).

4.4.5 Reconstruction of subgenome

To obtain more comprehensive subgenomes, consistent junction sites, which are supported in all four Vero E6 samples (two replicates in both NGS and Nanopore sequencing), were used to reconstruct SARS-CoV-2 subgenomes. Nanopore long reads, requiring all junctions to be consistent junction sites and both the start and end positions are within 45 nt to the genome boundaries, were counted as reliable subgenomes. The subgenomes were merged into clusters if all of their upstream and downstream junction sites are within 5 nt.

The sgRNAs were classified into three groups (Leader, ORF1ab and S-N) according to the junction position (Fig. 2A). The names of Leader-type sgRNAs were assigned by the first complete ORF downstream of the junction. The ORF1ab-type sgRNAs were named by the last complete ORF upstream of the junction in ORF1ab region. The strong junctions were called by requiring 100 or more NGS reads for Leader-type sgRNAs and 10 or more Nanopore reads for ORF1ab- type sgRNAs, respectively. The strong site with the largest count of reads in each type of sgRNA was assigned as the major junction.

ORF prediction was done using BioPython package v1.73 (**?**) for reconstructed sgRNAs according to the following two rules. For each Leader-type sgRNA, the annotated AUG ORFs start with the first annotated AUG downstream of the long junction. Non- annotated AUG between the first annotated AUG and the junction site was used for ORF prediction as upstream AUG ORF. For each ORF1ab-type sgRNA, the start codon of the ORF1ab gene was used to predict potential ORFs.

4.4.6 Characterization of pairing rules

For each of the 7,499 consistent SARS-CoV-2 junctions from NGS data, two pairs of sequences (UR-DR and UL-DL) were analyzed according to the two possible modes of template switch (Figure 3B). The minimum free energy (MFE) and RNA-RNA interaction pattern of paired sequences (20 nt each) were computed using the RNAhybrid (v2.1.2) program (**?**) with default parameters. The pairing of terminal bases between 5'UR and 3'DR or between 3'UL and 5'DL were added if they could form A-U/G-C/G-U pairings, while MFE was not adjusted. The mode of the pair with minimum MFE is assigned to the junction site, if the difference of two MFEs is greater than 1 kcal/mol. Otherwise, the mode of the junction is assigned as "Uncertain". One-side t test was used to evaluate the MFE differences between different groups of sgRNA JSs by expression level.

The "random junctions" were randomly generated with BEDTools (random command) to sequentially sample two locations across the viral genome as control junctions, not considering whether they were observed or not. The flanking sequences of 20 nt each were used to calculate MFEs. We used the default smoothing setting of geom_density function in ggplot2 package to compare the energy of the strongest, weakest and random junctions. KS-test was used as a significance test.

To evaluate the effect of terminal pairing state on sgRNAs generation, we divided junctions with the same pairing pattern into different subgroups for comparison in order to exclude interference from other factors. For junctions in one JS cluster, a subgroup was defined as a connected graph formed by the junction pairs set that have minimum value of the Hamming distance in all switching pairing strings were less than or equal to 2 in both of the UL-DL and UR- DR pairing states. The junction site with the highest expression in a subgroup is considered as a major site for a junction subgroup.

For SARS-CoV and MERS-CoV viruses, the same method of analysis was used. Due to the missing Nanopore data, the junctions appearing within two NGS replicates, and supported by more than 10 reads were used.

4.4.7 Motif analysis

The canonical TRS motifs (AAGAAC/ACGAAC) were searched in both the forward and reverse strands of the SARS-CoV-2 genome. One junction site was considered as motif-mediated if any complete TRS motif is found within 20 nt of both ends of this junction.

4.4.8 Noncanonical template switch analysis

A junction site (JS) from the template switch was considered as a long-range JS if its start site is upstream of the end site of the ORF1ab gene and its end site is downstream of the end of ORF1ab. Long-range junctions were divided into two categories based on their start positions. Long-range junctions with a start site smaller than 100 were defined as canonical leader sequence- mediated junctions and the rest were called non-canonical junctions. There is a large number of noncanonical junctions starting within the ORF1ab gene in both NGS and Nanopore data (Figure 1G-H and Figure S2F-G). Considering that there are only about 1.1 Nanopore reads per junction on average but a huge number of types (Figure S7F), the non-canonical junctions were only filtered by position, regardless of the expression level.

4.4.9 RT-PCR of sgRNAs

Vero E6 cells were infected with SARS-CoV-2 for 24 or 48 hours and harvested for RNA extraction with Trizol (Invitrogen). The extracted RNA was reverse transcribed into cDNA with an oligo-dT15 primer and SuperScript III RTase (ThermoFisher). PCR reactions were done with KOD-Plus-Neo DNA polymerase (TOYOBO) using SARS-CoV-2 specific primers. The PCR products were subjected to electrophoresis in 1% agarose gels and visualized with ethidium bromide staining.

4.4.10 Mass Spectrometry data analysis

A stringent pipeline was built to analyze SARS-CoV-2 Mass Spectrometry data recently published to investigate the coding capacity of non-canonical sgRNAs (Figure S7H). The Open Reading Frames (ORFs) were predicted using the 3-frame translation in ORF1ab-type sgRNAs, and using the 6-frame translation in SARS-CoV-2 genome sequence. The ORFs with more than 20 aa were used in the further analysis. The green monkey proteome was download from Uniprot database (19525 entries in 20201022 version). All protein candidates from predicted ORFs and the green monkey proteome were merged together as database for searching proteins and peptides with Maxquant v1.5.2.8 (**?**) in label-free quantification (LFQ) mode with default parameters.

Peptides only found in sgRNA proteins were used to identify ORF1ab-type sgRNA peptides in requiring the peptide to cover the sgRNA JS with flanking length over 10 nt on either side in two or more different biological samples. Those peptides having b ion support for the amino acids upstream JS and y ion support for the amino acids downstream JS were reported as validated ORF1ab-type sgRNA peptides.

Supplementary Material

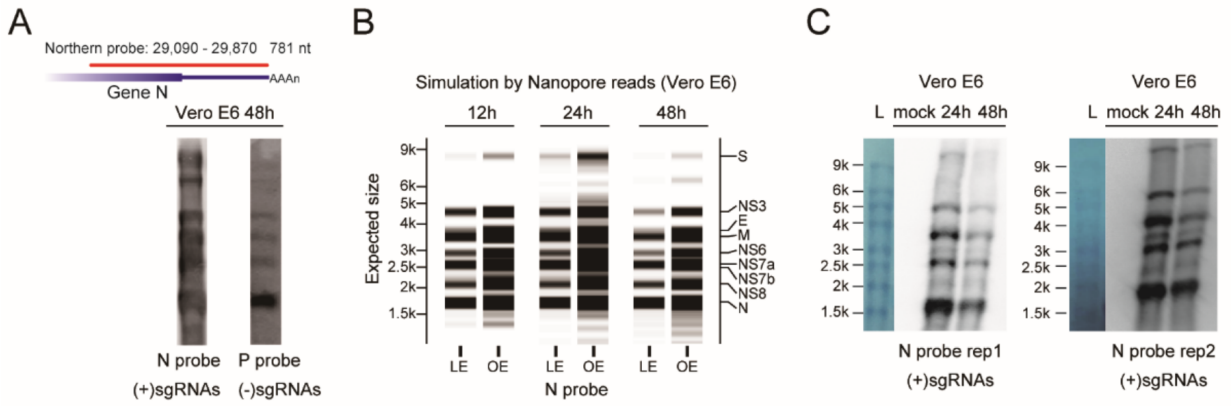

Figure S1: Northern results for SARS-CoV-2 subgenomes, related to Figure 1 (A) Northern results from RNA probes targeting (+)sgRNAs (N probe) and (-)sgRNAs (P probe) in Vero E6 48 h post-infection. The red line represents the position of probes relative to the genome. (B) Simulated Northern results for (+)sgRNAs by N probe with low- (LE) and over-exposure (OE) from Nanopore long reads. (C) Northern results by N probe for Vero E6 mock, 24 h, and 48 h post-infection samples with RNA ladders (L) in replicates.

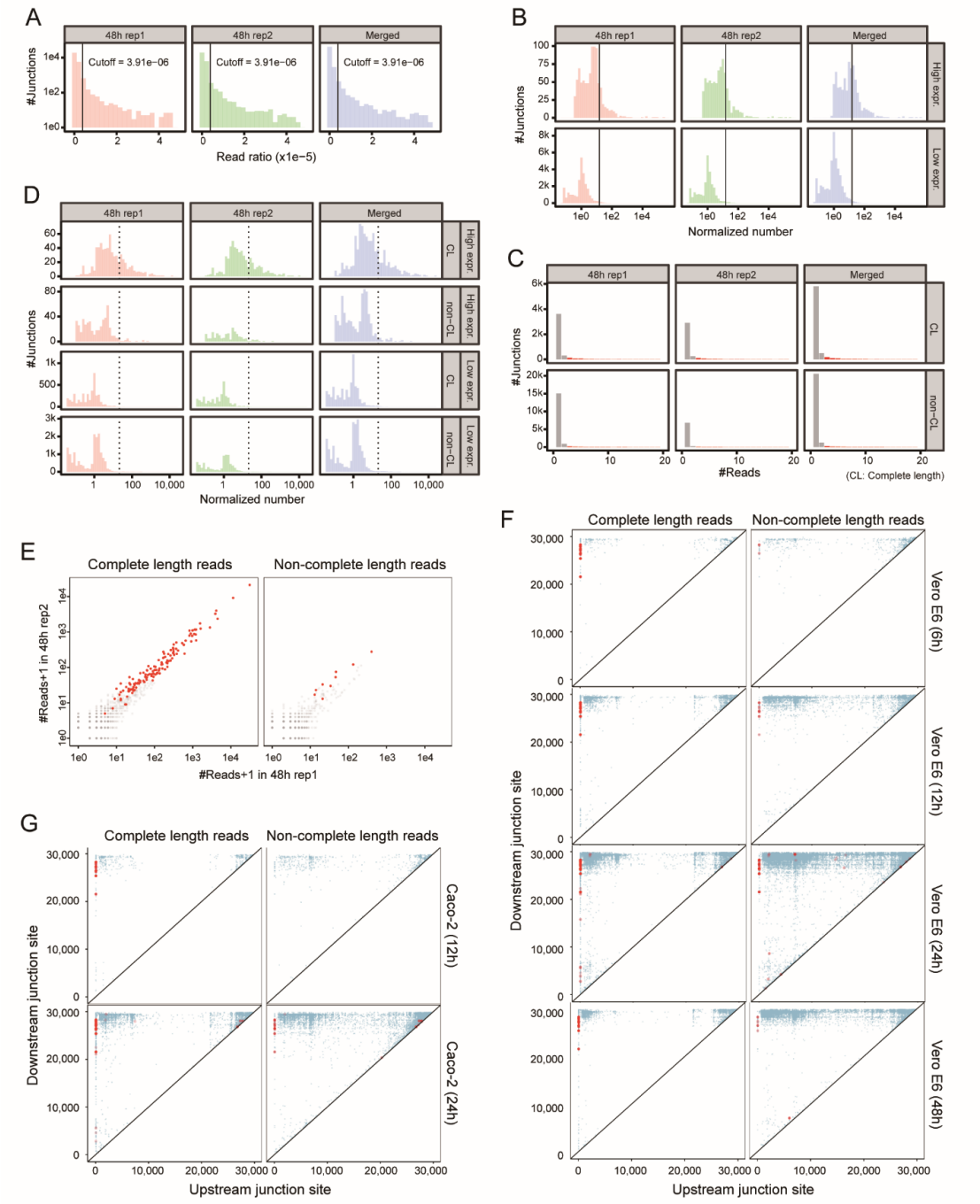

Figure S2: Global view of significant template switches, related to Figure 1 (A) Statistical threshold selection of normalized expression level for NGS data. (B) Threshold selection of local relative expression level for NGS data. (C) Threshold selection of normalized expression level for Nanopore data. (D) Threshold selection of local relative expression level for Nanopore data. (E) Reproducibility between two replicates of Nanopore full length and non-full length reads. The dot represents the reads counts per junction in replicate 1 (x-axis) and 2 (y-axis). Red points mark the significant junctions as in Figure 1D. (F) Global view of significant junction sites from Nanopore full length and non-full length reads in Vero E6 cells infected with SARS- CoV-2 at 6 h, 12 h, 24 h, and 48 h. Each dot represents a junction linking from the start position (x-axis) to the end position (y-axis). Red points represent the significant junctions identified from statistical analysis. (G) Global view of significant junction sites from Nanopore full length and non-full length reads in Caco-2 cells infected with SARS-CoV-2 at 12 h and 24 h. Red points represent the statistically significant junction sites.

Figure S3: Complete set of SARS-CoV-2 subgenomes, related to Figure 2 (A) Subgenome clusters reconstructed from Nanopore long reads (Vero E6, 48 h). Representative examples for different types of subgenomes (as in Figure 2C) are shown by row in a global view. (B) Subgenome zoom-in view with the number of Nanopore reads (right). Box and line represent transcribed and skipped regions due to template switches, respectively.

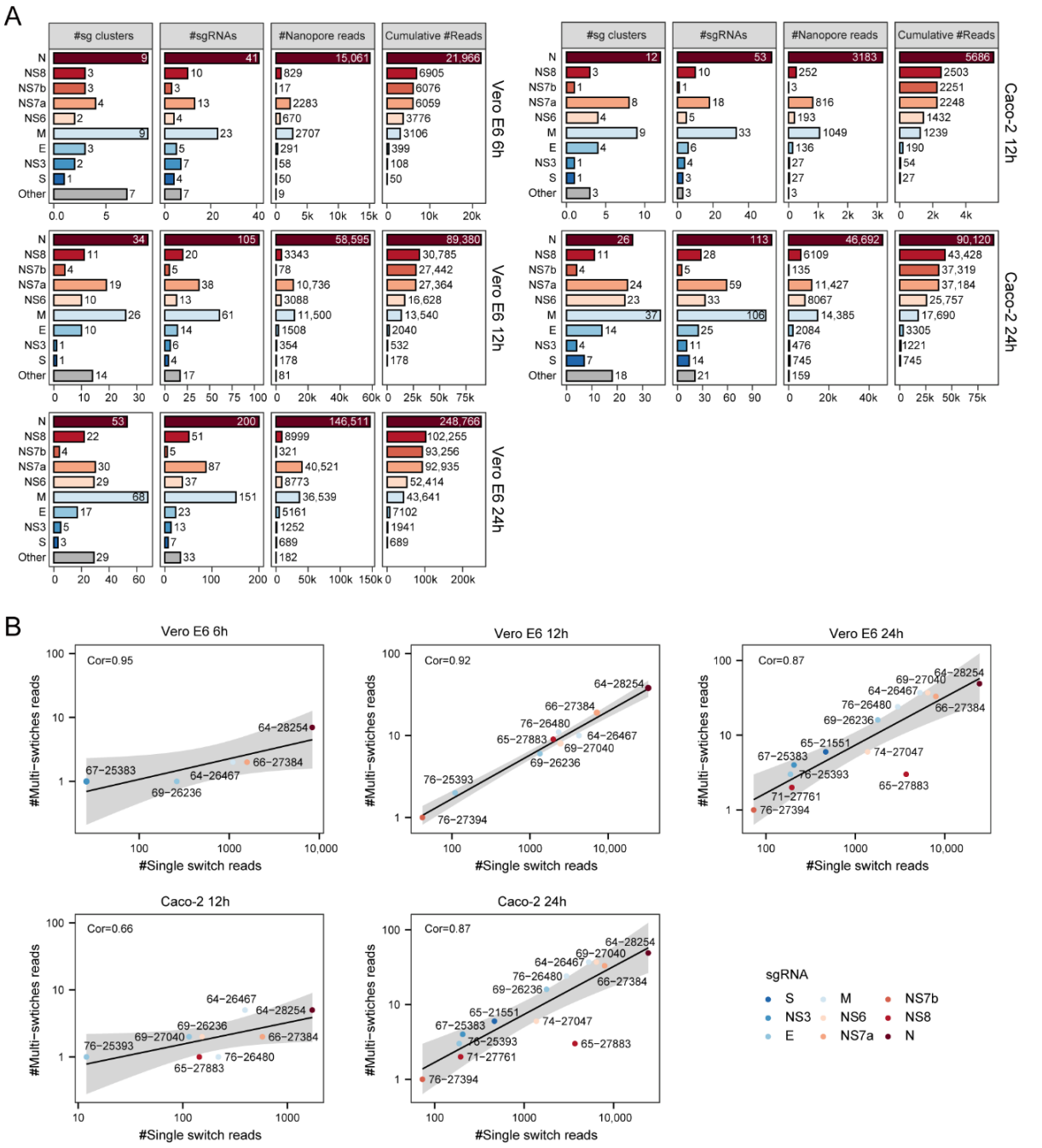

Figure S4: Statistics of canonical sgRNAs and multi-switches sgRNAs, related to Figure 2 (A) Statistics of 10 reconstructed Leader-type sgRNAs from Vero E6 cells post-infection at 6 h, 12 h, and 24 h (left) and from Caco-2 cells post-infection at 12 h and 24 h (right). (B) Comparison between the number of multi-switches reads vs. that of single-switch reads with specific junction for Leader-type sgRNAs in different samples. The Spearman correlation coefficient was labeled.

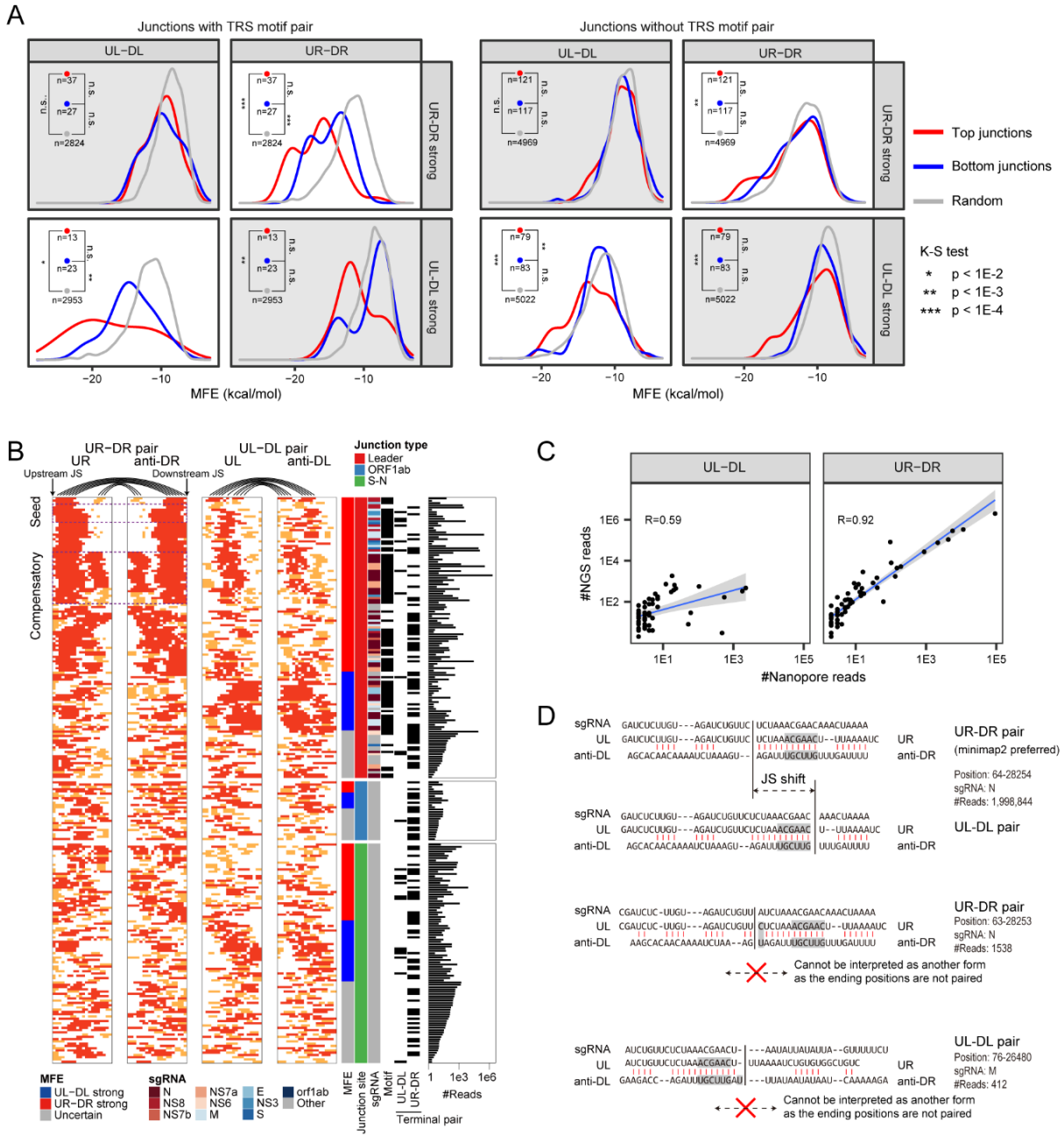

Figure S5: Switching modes and RNA-RNA interaction features for consistent template switches, related to Figure 3 (A) Minimum free energy (MFE) distributions for template switch junctions with (left) or without (right) TRS motif pair (Vero E6, 48 h). The junctions are split into two switch modes (UR-DR strong or UL-DL strong) to investigate the MFEs between UL-DL and UR-DR pairs (20 nt sequences) for highly expressed (red), lowly expressed (blue), and random (grey) junctions. The panels in white background highlight the MFE distributions from the potential causal junction pair for the corresponding switch mode. KS-test was used for significance testing. The two modes are defined as in Figure 3B. (B) Heatmap representation of base pairings between UR-DR and between UL-DL for consistent template switches. Each row represents two modes for one junction. The red or orange square represents paired state, whereas the white square represents unpaired state for the base-pairings between two specific segments flanking the upstream and downstream junction sites for template switching events by row, as illustrated by the arcs linking the predicted base-pairs for the first row of template switch. Energy group by MFE difference, subgenome type by first ORF, the occurrence of TRS motif, terminal pairing status, and NGS reads number are shown on the right side. Representative seed and compensatory pairing patterns are marked with dotted boxes in the UR-DR pair mode. (C) Comparison between the numbers of NGS and Nanopore reads for UL-DL (left, negative-to-negative) and UR-DR pair (right, positive-to-positive) junctions in Vero E6 cells 48 h post-infection. (D) Schematic examples illustrating the number of template switches with forward mode may be underestimated due to the mapping ambiguity and preference of a specific mapping program (top case). There are cases that can only be interpreted by either forward or reverse mode (bottom two cases).

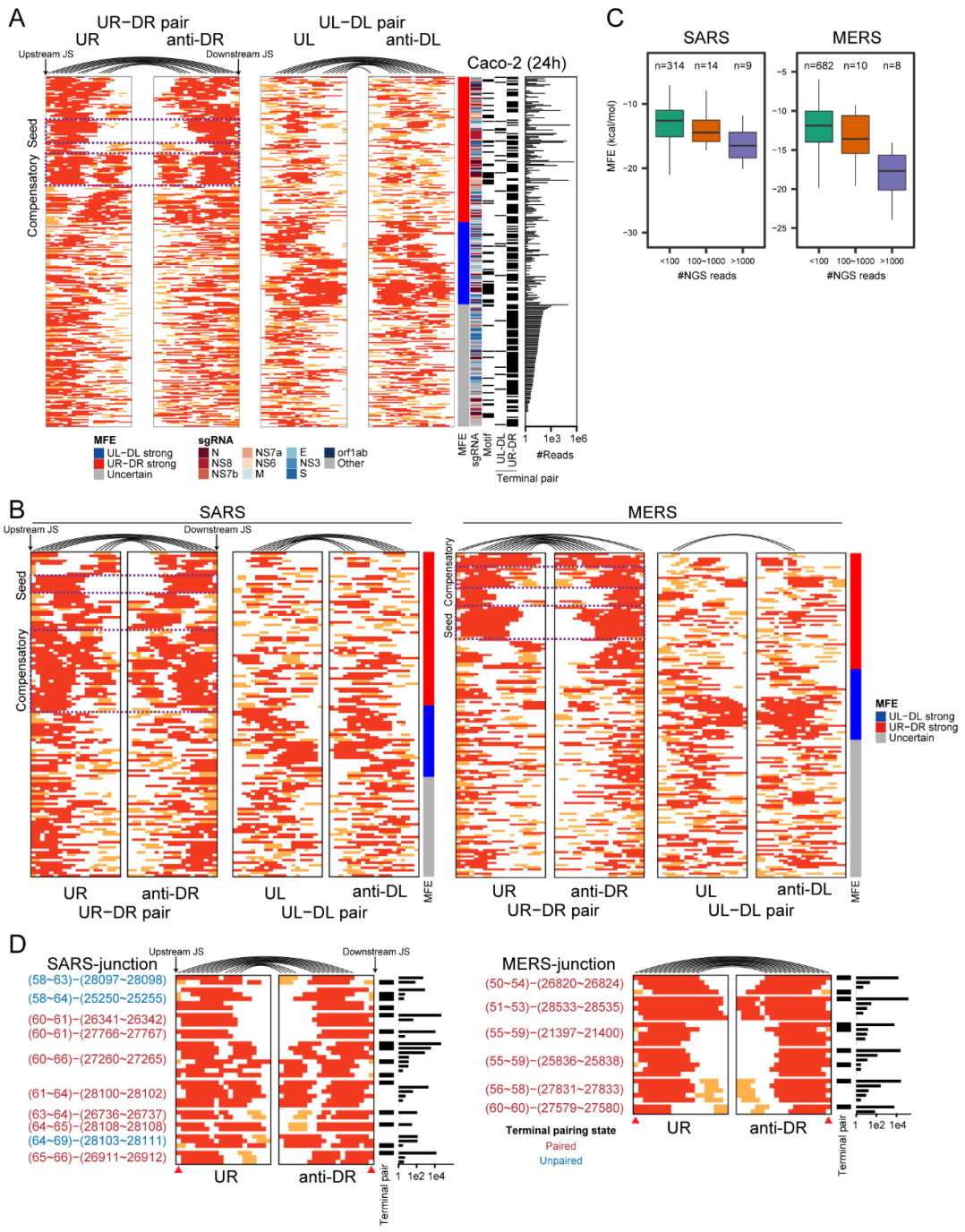

Figure S6: RNA-RNA interaction features of template switches in different coronaviruses, related to Figure 3 (A) Heatmap representation of base pairings between UR-DR and between UL-DL for consistent template switches in SARS-CoV-2 infected Caco-2 cells at 24 h, as Figure 3B. The red or orange square represents paired state, whereas the white square represents unpaired state for the base- pairings between two specific segments flanking the upstream and downstream junction sites for template switching events by row, as illustrated by the arcs linking the predicted base-pairs for the first row of template switch. (B) Heatmaps as (A) representing the RNA-RNA base-pairings in two modes (UR-DR pair and UL-DL pair) for SARS and MERS template switch junctions. The junctions (rows) are from Leader-group sgRNAs detected in public NGS data with at least 10 supporting reads from SARS (left) and MERS (right) infected Calu-3 cells. The MEF propensity is shown on the right for junctions by row. The dotted boxes highlight representative junctions with RNA-RNA base-pairings in seed and compensatory manners. (C) Boxplots of MFEs sub- grouped by the number of NGS reads for Leader-group sgRNA junctions in SARS (left) and MERS (right) infected samples. (D) RNA-RNA base-pairing visualization (same as Figure 3J) between UR and anti-DR segments flanking template switch sites for SARS (left) and MERS (right) infected samples. The columns indicating pairing states of the two terminal bases are marked by arrows at the bottom.

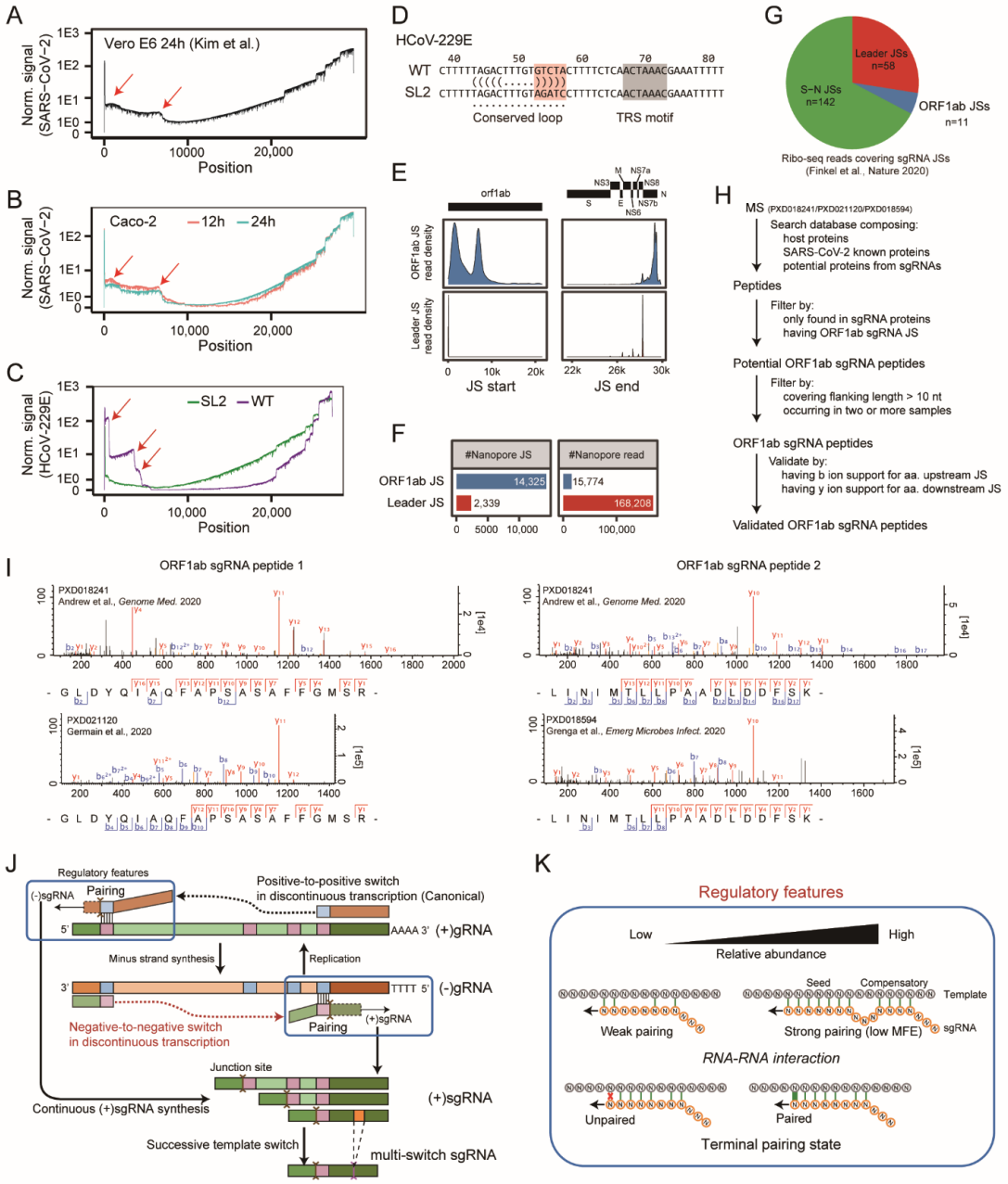

Figure S7: Non-canonical template switches in different coronaviruses, related to Figure 5 (A) Nanopore read coverage profile across the SARS-CoV-2 genome post-infection in Vero E6 cells from the Kim et al. study. The arrows in ORF1ab region mark obvious loci with abrupt changes indicating template switch junctions. (B) Nanopore read coverage profiles across the SARS-CoV-2 genome in Caco-2 cells at 12 h and 24 h post-infection. (C) Nanopore read coverage profiles across the HCoV-229E genome for WT and SL2 strains 24 h post-infection in Huh7 cells from the Viehweger et al. study. (D) Sequence difference between the WT and SL2 strain for HCoV-229E virus. (E) The distribution of junction boundaries (left for JS start and right for JS end) from Nanopore data for ORF1ab-mediated (top) and Leader sequence-mediated (canonical, bottom) long-range template switches in Vero E6 cells 48 h post SARS-CoV-2 infection. (F) The number of different junction sites (JS, left) and reads (right) for ORF1ab-mediated and Leader sequence-mediated (canonical) template switches supported by Nanopore data from Vero E6 cells 48 h post SARS-CoV-2 infection. (G) Statistics of different types of junction sites (JSs) covered by SARS-CoV-2 Ribo-seq reads from the Finkel et al. study. (H) Mass Spectrometry (MS) data processing pipeline to identify sgRNA derived peptides. (I) Mass spectrum of the ORF1ab-type peptides in Figure 5G with both y and b ion supports. (J) Graphical illustration for positive-to- positive (canonical) and negative-to-negative switching modes in generating single- and multi- switch sgRNAs from discontinuous transcription from gRNA. (K) Graphical illustration of regulatory features in governing sgRNA abundance. Strong RNA-RNA base-pairings (low MFE, Minimum Free Energy of base-pairings) and terminal pairing status would lead to efficient template switching and high abundance of corresponding sgRNA.