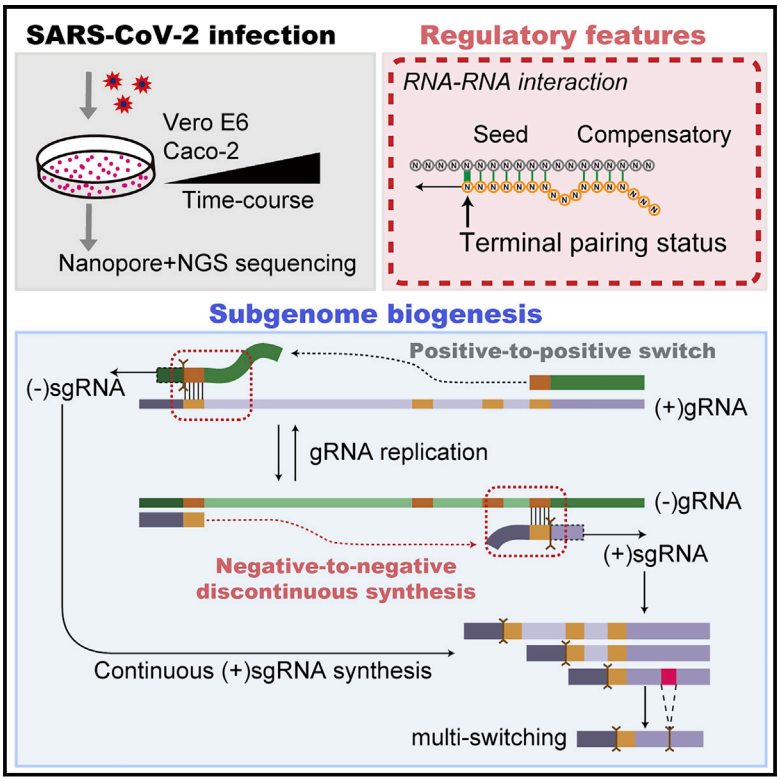
Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Molecular Cell

Article

The SARS-CoV-2 subgenome landscape and its novel regulatory features

Graphical Abstract



Highlights

- Dynamic subgenome landscapes of SARS-CoV-2 in two host cells are constructed
- Bidirectional and successive template switching diversify sgRNA biogenesis
- Several key determinants governing template switching efficacy are discovered
- Canonical TRS-independent RNA-RNA interaction mediates template switches

Authors

Dehe Wang, Ao Jiang, Jiangpeng Feng, ..., Ke Lan, Yu Chen, Yu Zhou

Correspondence

klan@whu.edu.cn (K.L.), chenyu@whu.edu.cn (Y.C.), yu.zhou@whu.edu.cn (Y.Z.)

In brief

Wang et al. construct dynamic landscapes of SARS-CoV-2 subgenomic RNAs (sgRNAs) by using integrated poly(A) RNA sequencing at various time points after infection. Computational analyses reveal novel modes of viral sgRNA biogenesis and decode regulatory features, including several key determinants governing the efficacy of template switching in coronaviral RNA transcription.

Wang et al., 2021, Molecular Cell *81*, 2135–2147
May 20, 2021
https://doi.org/10.1016/j.molcel.2021.02.036

CellPress

Molecular Cell

CellPress

Article

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, 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.)

https://doi.org/10.1016/j.molcel.2021.02.036

SUMMARY

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is currently a global pandemic. CoVs are known to generate negative subgenomes (subgenomic RNAs [sgRNAs]) through transcription-regulating sequence (TRS)-dependent template switching, but the global dynamic landscapes of coronaviral subgenomes and regulatory rules remain unclear. Here, using next-generation sequencing (NGS) short-read and Nanopore long-read poly(A) RNA sequencing in two cell types at multiple time points after infection with SARS-CoV-2, we identified hundreds of template switches and constructed the dynamic landscapes of SARS-CoV-2 subgenomes. Interestingly, template switching could occur in a 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. Two TRS-independent template switch modes are also responsible for subgenome biogenesis. 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.

INTRODUCTION

The recent outbreak of coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; also referred to as human coronavirus 2019 [HCoV-19]; Zhou et al., 2020; Liu et al., 2020; Chen et al., 2020a) has turned into a global pandemic, causing more than a million deaths as of October 2020 (Dong et al., 2020). SARS-CoV-2 is an enveloped RNA virus with ~30,000-nt-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 (Zhou et al., 2020; Chen et al., 2020a). The genomic RNAs (gRNAs) of CoVs have a 5′ cap structure and 3′ poly(A) tail, at whose 5′ end two large open reading frames (ORF1a/1b) encode 16 viral nonstructural proteins (nsps) occupying two-thirds of the genome. Polyproteins 1a/1ab (pp1a/1ab) are translated directly from the gRNA through −1 ribosomal frameshifting (Perlman and Netland, 2009). The 3′ end of the CoV genome (one-third of the genome size) contains genes encoding several main structural proteins, including the spike protein (S), envelope protein (E), membrane protein (M), nucleocapsid protein (N), and various accessory proteins (Perlman and Netland, 2009; Chen et al., 2020b).

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 (Snijder et al., 2016), which is more complicated than that of 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, whereas discontinuous jumping produces (−)subgenomic RNAs (sgRNAs) with common 5′ and 3′ ends (Hussain et al., 2005). Positive-sense progeny gRNAs and sgRNAs are synthesized using these negative-strand RNA intermediates as templates (Sola et al., 2015). The discontinuous jumping step, called "template switch," is mediated by the transcription-regulating sequence (TRS) in

the genome body (TRS-B) and in the 5′ leader sequence (TRS-L) upstream of ORF1ab (Sola et al., 2015), resulting in fusion of leader-body sequences. Thiel et al. (2003) identified eight sgRNAs of SARS-CoV (Thiel et al., 2003), whereas our subsequent study identified 10 sgRNAs, including two novel sgRNAs (Hussain et al., 2005). Recently, more sgRNA variants of HCoV-229E were reported and remain to be characterized (Viehweger et al., 2019). Kim et al. (2020) 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- to 7-nt core sequence surrounded by variable sequences. Different core TRSs have been reported previously, including CUAAAC for CoV transmissible gastroenteritis virus (TGEV) (Zúñiga et al., 2004), ACGAAC for SARS-CoV (Hussain et al., 2005; Thiel et al., 2003), and CUUUAGA for equine torovirus (Stewart et al., 2018). It is hypothesized that formation of a duplex between TRS-L and downstream TRS-B core sequences determines the template switches (Sola et al., 2015). However, the regulatory features of TRS-like elements in CoVs, including SARS-CoV-2, have not yet been defined.

Previous studies, including our work (Hussain et al., 2005), 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 blotting 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 detected the SARS-CoV-2 virus in people (Liu et al., 2020), assembled the SARS-CoV-2 genome sequence (Chen et al., 2020a), and characterized the transcriptomes of samples from affected individuals (Xiong et al., 2020). 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. (2019) 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 sequencing (RNA-seq), including NGS and Pacific Biosciences (PacBio) long-read sequencing, to construct the high-resolution transcriptional landscape (Wang et al., 2019).

In this study, we employed NGS and Nanopore direct RNA-seq 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 after 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 switching also exists during positive-strand synthesis and identified two other large classes of subgenomes. Our findings provide a global view of SARS-CoV-2 subgenomes and uncover the molecular basis governing their biogenesis.

RESULTS

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 blotting (Figure S1). To characterize high-resolution SARS-CoV-2 sgRNAs, we used hybrid poly(A)-selected RNA-seq technologies to analyze local template-switching events and simultaneously construct full sgRNAs. Using poly(A) RNAs enriched from total RNAs extracted from monkey Vero E6 cells (ATCC number CRL-1586) and human Caco-2 cells infected with SARS-CoV-2 (WIV04, IVCAS 6.7512), we constructed RNA-seq libraries with duplicates for the NGS Illumina and Nanopore MinION platforms, respectively. NGS libraries were sequenced in paired-end 150-bp mode, whereas Nanopore libraries were sequenced by direct RNA-seq (Figure 1A).

To investigate the dynamic landscapes of viral RNAs, we performed the assays at multiple time points after infection of Vero E6 and Caco-2 cells with SARS-CoV-2. We chose the two cell types to explore potential differences in 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%, and 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, respectively; smaller than those in Vero E6 samples and consistent with the known low infection rate in Caco-2 cells (Table S1). We found that the Nanopore long reads contain poly(A) tails with a median poly(A) length of around 50 nt (Figure 1B), similar to that at other time points and consistent with a recent report (Kim et al., 2020). We found that all Nanopore reads are mapped in the (+) genome, suggesting that (−)gRNA or (−)sgRNAs do not contain poly(A) tails. In Vero E6 cells 48 h after infection, the median length of Nanopore reads is 1,248 nt, and 22 reads are mapped to a genome with a length close to 30,000 nt, 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 correct localization of junctions, we required the sequences in reads flanking the junctions on both sides to have at least 20-nt exact matches with the genome. We first analyzed the Vero E6 samples 48 h after 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 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 noise from uneven RNA abundance, we used statistical scoring to remove the effect of local background around both junction sites (Figures S2A-S2D; STAR Methods) and obtained 100 significant junctions (Figure 1D, red points; Table S2). To further verify the junctions, we identified 141 significant junctions embedded in the Nanopore long reads using similar statistical

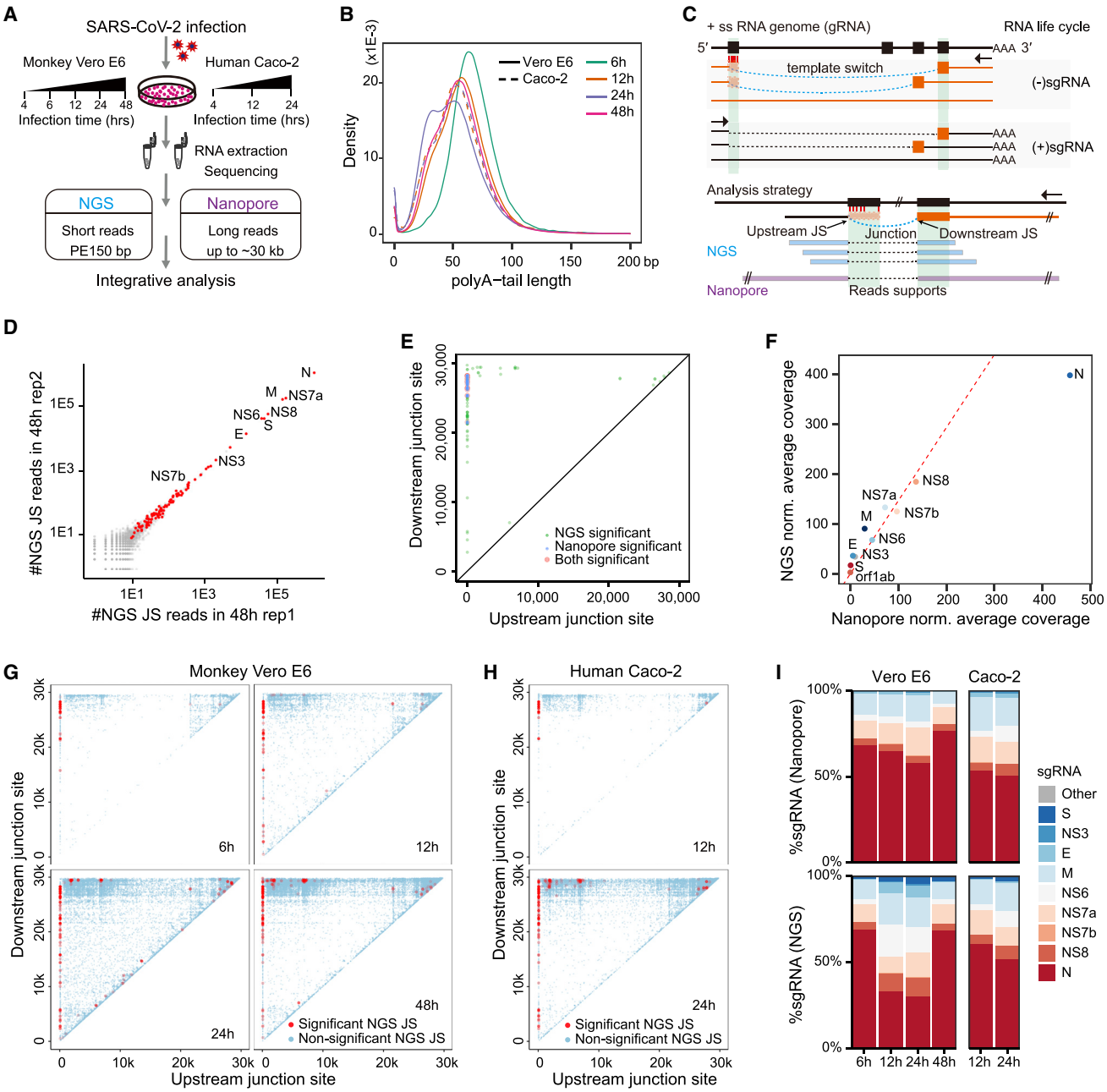


Figure 1. Experimental strategy and analysis for global mapping of template switches

(A) Experimental design for decoding SARS-CoV-2 subgenome dynamics at different time points after infection of 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 of two replicates of NGS data. Each dot represents the read counts of one junction in replicates 1 (x axis) and 2 (y axis). Red points represent the significant junctions identified by statistical analysis.

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

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

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

(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 Figures S1 and S2.

methods (Figure S2E; Table S2), 31 of which are overexpressed significantly in our NGS reads (Figure 1E), suggesting reliability of the predicted template-switching events.

The SARS-CoV-2 junctions we detected in NGS or Nanopore data included all 10 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 the N protein showed the highest level, whereas 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 switching (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 infection and observed an increasing number of junction sites in Vero E6 cells (Figures 1G and S2F) and Caco-2 cells (Figures 1H and S2G). 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 or as other group. The N sgRNA has the highest expression level of junction sites in Nanopore and NGS data (Figure 1I).

Diverse types of full-length subgenomes

To identify complete subgenomes, we took advantage of Nanopore long reads and only considered reads covering the 5′ end leader sequence and extending to the 3′ end of SARS-CoV-2 genome. Internal junctions from template switching must be found in all 4 samples of NGS and Nanopore datasets from Vero E6 cells 48 h after infection (Table S2). By this definition, we identified 433 different subgenomes (sgRNAs) in 208 clusters by merging neighboring upstream or downstream sites within 5 nt, which were subsequently classified into 3 groups (Figure 2A; Table S3). The first group is called leader/S-N, representing canonical template switches joining leader sequence with downstream genes from S to N; the second group, called ORF1ab/S-N, represents novel sgRNAs with template switches linking positions in ORF1ab with downstream genes from S to N; and the third group, called S-N/S-N, contains novel sgRNAs with internal template switches in S-to-N regions. The latter two groups of non-canonical events were also observed in a recent report with different classification rules (Kim et al., 2020).

The sgRNAs in the leader/S-N group were named based on the first annotated gene downstream of the junction. The distribution of downstream junction sites associated with those sgRNAs is shown in Figure 2B, in which the strong sites (with more than 100 NGS reads support) and the major site (with the largest number of NGS reads support) for each sgRNA are marked by red lines and asterisks, 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 had two or more gaps resulting from template switching. This suggests that template switches may occur simultaneously with or independent of other switching events. We identified two groups of multi-switch junctions, leader+S-N and leader+2×S-N, which contain the leader/S-N junction and 1 or 2 S-N/S-N junction(s), respectively (Figures 2C and 2E). The read numbers for single-switch sgRNAs are much larger than those for multi-switch sgRNAs, and the read numbers of leader+S-N sgRNAs are larger than those of leader+2×S-N sgRNAs, which also indicates that the Leader/S-N group is the dominant form of sgRNAs (Figure 2E).

Moreover, different multi-switch sgRNAs share common junctions. As shown in Figure 2F, seven bi-switch and four tri-switch sgRNAs share junction site 28,525–28,576, with more Nanopore read support for the bi-switch sgRNAs. Generally, the higher the expression of one parental sgRNA, the larger the counts of multi-switch sgRNAs originating from it. The correlation between the number of multi-switches and that of corresponding single-switch reads is 0.89 (Figure 2G). This positive correlation is also evident at different time points after SARS-CoV-2 infection in Vero E6 and Caco-2 cells (Figure S4B). These data support the theory that sgRNAs resulting from template switching can function as templates for additional template-switching events. These results show the complete landscapes of the subgenome structures and their expression levels, providing a useful resource for studying their functions and regulation mechanisms.

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 samples (Figure 3A, left). As expected, we found an already 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 base pairs between TRS-L and anti-TRS-B (Figure 3A, center).

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 JSs as UL (upstream left) and UR (upstream right), respectively, and those flanking downstream JSs as DL (downstream left) and DR (downstream right), respectively. We used the RNAhybrid program (Rehmsmeier et al., 2004) to find optimal base pairings with minimum free energy (MFE) between flanking segments, UR with anti-DR (containing TRS-L and anti-TRS-B) compared with UL with anti-DL. As expected, we found 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 9 sgRNAs (Figure 3A). In analogy to the microRNA (miRNA)-mRNA base-pairing rules (Bartel, 2018), we defined two base-pairing patterns: seed mode (6 bp in 1- to

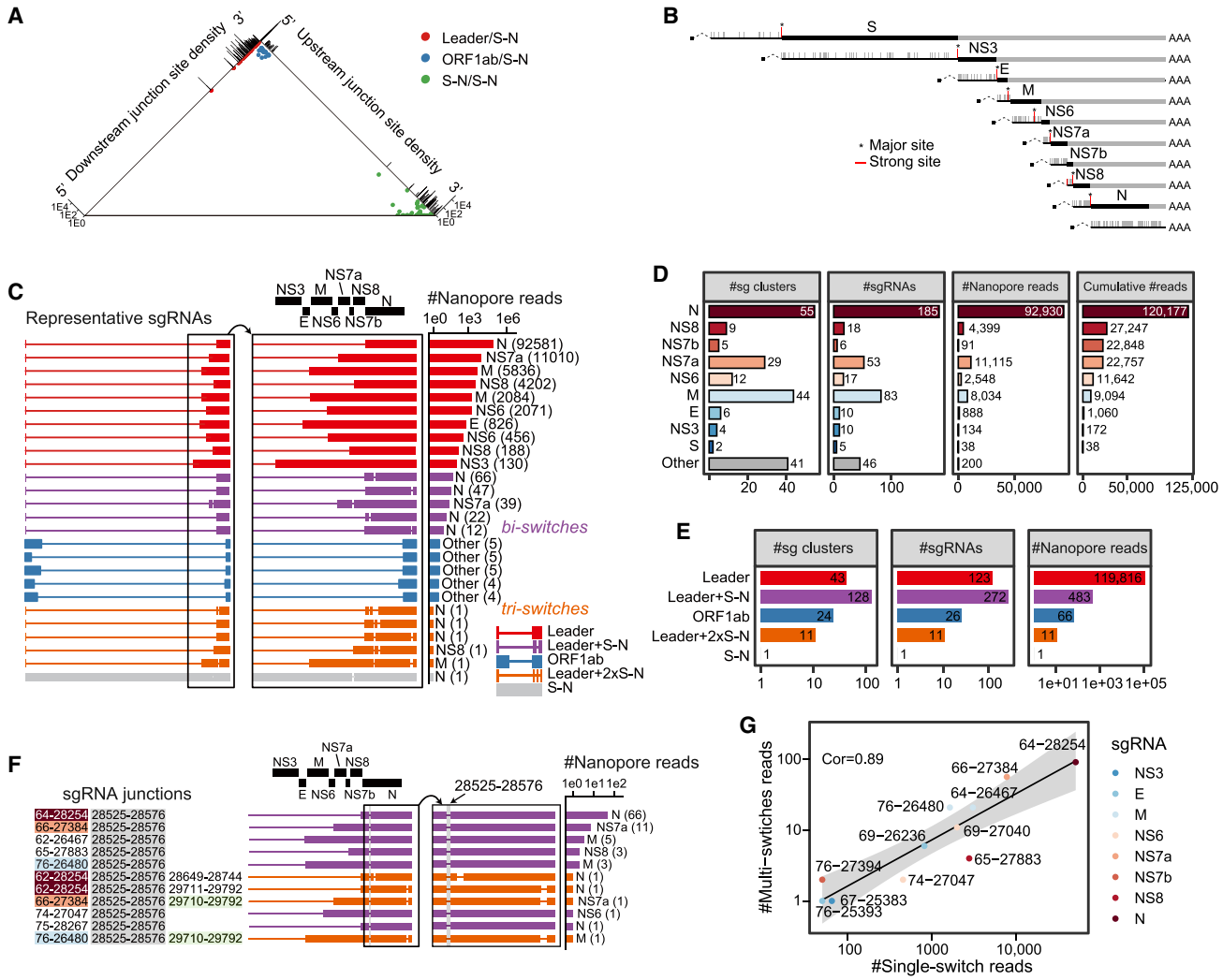


Figure 2. Global landscape of SARS-CoV-2 subgenomes

(A) Global view of consistent template switches in NGS and Nanopore data. Each template switch is represented as a point by the genomic positions of its upstream and downstream JSs 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 JSs are shown in the top right and top left bar graphs, respectively.

(B) Distribution of the downstream JSs for leader-group sgRNAs (48 h, Vero E6 cells). The sgRNA names were assigned based on the first annotated gene downstream of the junction. Strong sites (with more than 100 NGS read support) are marked as red lines, of which the major site (with the largest number) in each sgRNA group is marked with an asterisk.

(C) Subgenome clusters reconstructed from Nanopore long reads. Representative examples of five different types of subgenomes (colored legend) are shown by row in global (left) and magnified (center) views with the number of supporting reads (right). Boxes and lines represent transcribed and skipped regions, respectively, because of template switches. The top 10 leader-type and top 5 other types of subgenomes are shown. The label of the subgenome was assigned by the first ORF after the template switch.

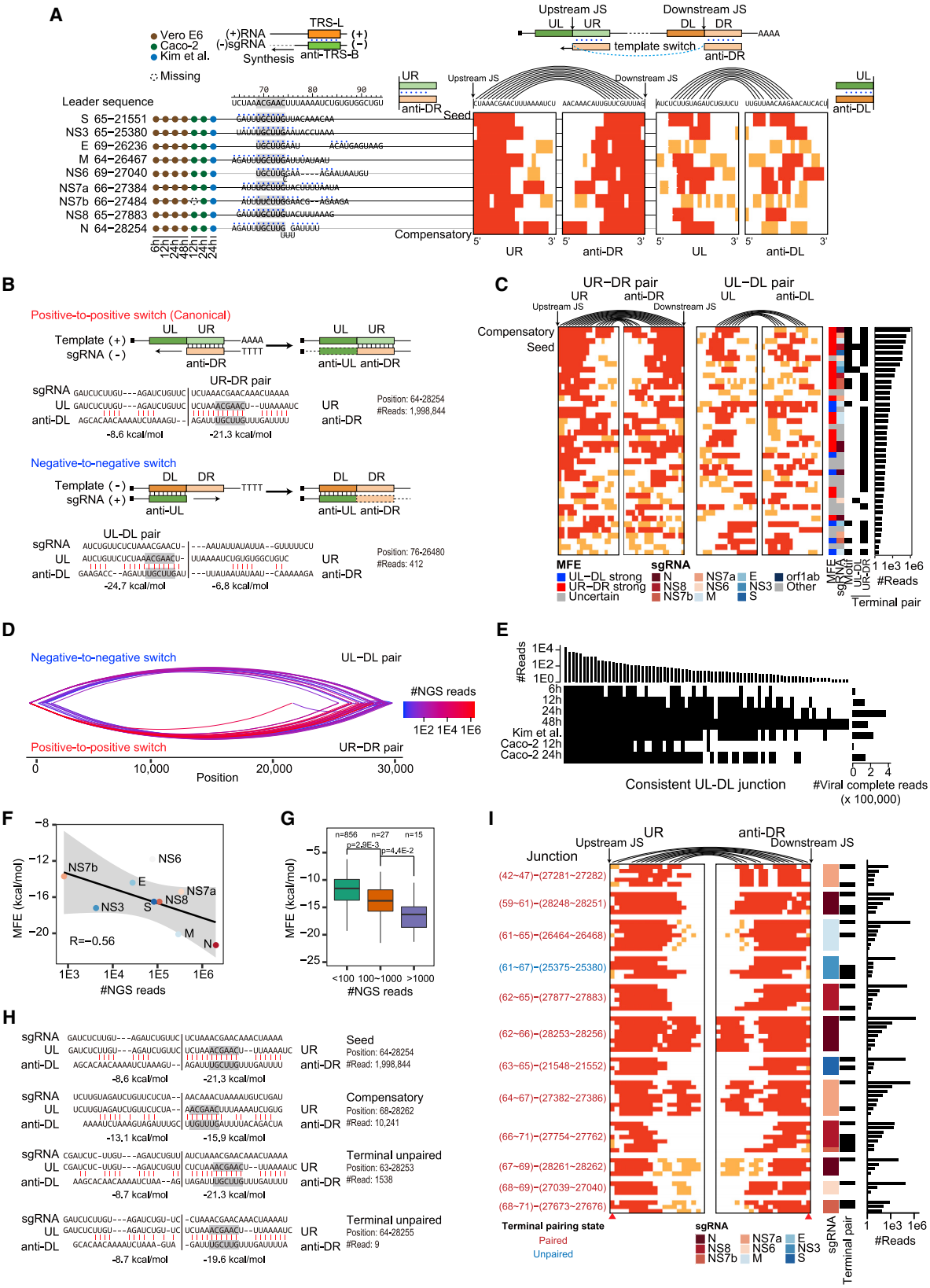
(D) Statistics for 10 subgenome types classified by the first complete ORF in the subgenome (Vero E6 cells, 48 h). For each type of sgRNA, the number of clusters, sgRNAs, Nanopore reads, and cumulative count of Nanopore reads containing the ORF are shown. Because S sgRNAs are the longest canonical sgRNAs, they might be sequenced less efficiently by Nanopore.

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

(F) Examples of multi-switch sgRNAs with common junctions (28,525–28,576). There are 7 bi-switch and 4 tri-switch sgRNAs (numbers of supporting Nanopore reads are shown on the right).

(G) Comparison of the number of multi-switch reads versus single-switch reads with a specific junction for leader-type sgRNAs. The Spearman correlation coefficient is labeled.

See also Figures S3 and S4.



(legend on next page)

7-nt flanking JSs) and compensatory mode (with additional base pairs outside of the seed region), as marked for S and N sgRNAs in Figure 3A.

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 negative-to-negative template switching also exists, occurring while generating (+)sgRNAs. 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 a UR-DR pair (UR::anti-DR for positive-to-positive mode) or a UL-DL pair (anti-UL::DL for negative-to-negative mode). We posited that the relative strength 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 MFE between the UR-DR and UL-DL pairs (method details). As shown in Figure 3B, for the top example on junction 64-28254, the MFE of UR-DR is much lower than that of 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 JSs found in all samples 48 h after infection. We found that junctions with a high expression level (top junctions) have much lower MFE in UR-DR or UL-DL mode than those with a low expression level (bottom junctions) or random junctions without a known TRS motif pair (Figure S5A). These data are further evidence that both modes exist and can be differentiated by considering the relative positioning of the MFEs for the two pairing modes.

Moreover, we investigated the detailed RNA-RNA base-pairing patterns that mediate template switching, involving the sequences UR and anti-DR or UL and anti-DL, respectively (20 nt flanking the JSs, as shown in Figures 3A and 3B). We observed that many junction reads have close JSs with small shifts, and we used a method based on a 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 the 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 of Vero E6 cells 48 h after 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 detected frequently in multiple SARS-CoV-2-infected samples, especially those with high expression (Figure 3E). The junctions resulting from the two different modes are supported in NGS RNA-seq and Nanopore sequencing (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 (Li, 2018) prefers

Figure 3. RNA-RNA pairing determinants in template switching efficacy

(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 an empty circle (NS7b sgRNA). Base pairings between the TRS-L and anti-TRS-B segments are represented by blue dots, and the TRS motifs are highlighted in gray. The heatmaps on the right represent base pairings between the UR-DR pair (UR and anti-DR) and UL-DL pair (UL and anti-DL). The red or orange squares represent paired states, whereas white squares represent an unpaired state for base pairings between two specific segments flanking the upstream and downstream JSs for template-switching events by row, as illustrated by the arcs linking the predicted base pairs for the first row of template switches (S sgRNA). Red indicates a consecutive paired state in a 6-nt segment with at least 5 nt.

(B) Illustrations and examples of 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 a gray box. The number of NGS reads in 48-h Vero E6 cell data and the MFEs between different pairing segments are shown.

(C) Heatmaps as in (A), representing RNA-RNA base pairings in two modes (UR-DR pair and UL-DL pair) for consistent and core template switch junctions from Vero E6 cell 48-h data. The junctions shown by row are detected in NGS and Nanopore reads, with the largest numbers 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 (top) and positive-to-positive (bottom) template switches for consistent junctions in NGS and Nanopore data 48 h after infection. The numbers 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) is 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 cells). The Spearman correlation coefficient is labeled.

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

(H) Representative examples showing that RRI features affect template switching efficacy. The 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 in (A) between UR and anti-DR segments flanking template switch sites. The columns indicating the pairing states of the two terminal bases are marked by red arrows. Neighboring junctions with similar pairing patterns were grouped together, and the terminal pairing states for the major junctions in each group was marked by color (red for paired and blue for unpaired). The corresponding sgRNA, terminal pairing state, and read numbers (48 h, Vero E6 cells) for each junction are shown by row on the right.

See also Figures S5 and S6.

the UR-DR setting and that 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 the pairing is a key factor because the MFEs between UR-DR or UL-DL pairs have evident negative correlations 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 in 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 with the top 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 effect 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 a paired state for the terminal pairing in 11 of 12 subgroups.

The global RNA base-pairing flanking JSs have similar patterns for SARS-CoV-2-infected Caco-2 cells (Figure S6A) and SARS- and MERS-infected Calu-3 cells (Figure S6B). The features of MFE and terminal pairing status are also observed in SARS and MERS (Figures S6C and S6D).

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

TRS motif-independent RRI-mediated template switching

Previous studies have found that the TRS is important in the biogenesis of subgenomes and that the same TRS motif in leader and body sequences can form strong base pairings during (−) sgRNA synthesis and mediate template switching (Di et al., 2017; Hussain et al., 2005; Pasternak et al., 2001; Zúñiga et al., 2004). Can the same TRS motif be found in the SARS-CoV-2 genome? We searched the canonical TRS motifs AAGAAC and ACGAAC across the positive and negative SARS-CoV-2 genome and found 30 and 12 occurrences, respectively (Figure 4A). As expected, we did find the TRS motif in 158 of 7,499 unique JSs found in two NGS replicates from Vero E6 cells 48 h after 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). TRS motifs in positive and negative strands can participate in template switching. 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 categorization into 3 subgroups of consistent junctions. We observed that the ORF1ab and S-N groups do not have the TRS motif pair in the flanking sequences of the junctions (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 the TRS motif ACGAAC depends on the pairing context in comparison with case #1 versus #5 junctions associated with the NS8 gene, where 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 effect on template switching efficacy. We further validated the NS8 #2 junction by RT-PCR and clone sequencing (Figure 4D), resulting from RNA base pairing without a TRS motif around the downstream JS.

We detected one special leader-type sgRNA without a TRS motif around upstream nor downstream JSs, although it had low expression; it was found in multiple samples based on our and published Nanopore datasets (Figure 4E). The bases in UR segment pairing with anti-DR are in the conserved loop (Rfam v.14.2 RF03117; Kalvari et al., 2018) upstream of the TRS-L motif (Figure 4F), and we hypothesize that the RNA secondary structure of this loop may lead to a low frequency of this type of sgRNA. For the leader-type sgRNAs, the major mode of template switching involves a TRS motif pair plus additional base pairs, whereas alternative modes are partially TRS- or non-TRS-mediated base pairings (Figure 4G).

Many template switches do not need the TRS motif, and RRI-mediated proximity of two long-range JSs may be the key determinants of CoV subgenome biogenesis.

Weak but extensive fusion between ORF1ab and N ORF RNA regions

In addition to the leader sequence-participating template switches, the other group of long-range template switches occurs unexpectedly between the upstream site in ORF1ab and the downstream site in 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 the ORF1ab group has a broad distribution and two peaks toward 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 (Kim et al., 2020), showing a similar trend (Figure S7A). We also observed similar patterns, including leader- and ORF1ab-type sgRNAs, in Caco-2 cells at two time points (Figure S7B). The ORF1ab-type sgRNAs were detected in another CoV infection, HCoV-229E (wild type [WT]), but not in its mutant strain infection (SL2), by analyzing the published Nanopore data (Viehweger et al., 2019), as shown in Figure S7C. The different profiles of WT and SL2 may be caused by 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 suggest that biogenesis of ORF1ab-type sgRNAs is widespread and regulated.

The ending position also has a broad distribution in the N RNA region, as exemplified by the Vero E6 cell data 48 h after infection (Figure S7E). As summarized by the number of Nanopore long reads, junctions starting in the leader sequence have 168,208 reads over 2,339 unique junctions, whereas junctions starting

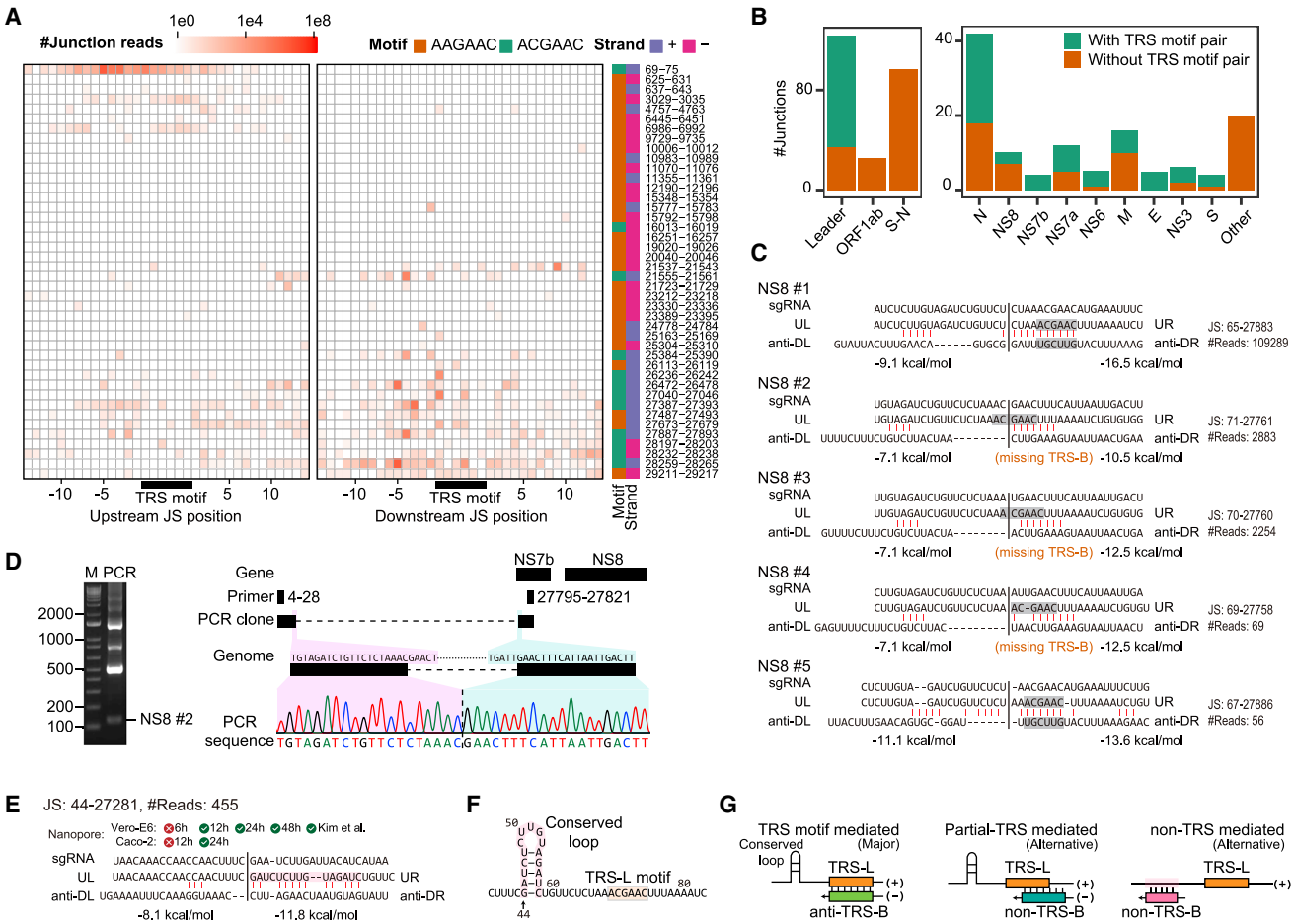


Figure 4. TRS motif-independent RRI in template switch

(A) Heatmap showing use of TRS motifs in donor and acceptor sites of template switches. The motif sequences, strands, 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 without TRS motifs. The TRS motif is marked in gray, and the number of 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 sequences, and cloned sequences are shown on the right.

(E) An example of upstream, non-TRS mediated, leader-type sgRNA. The number of NGS reads for Vero E6 48 h cells is shown. The presence of this sgRNA in Nanopore data from different samples is 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.

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 the ORF1ab type (Figure 5E).

To investigate the potential functions of these non-canonical sgRNAs, we analyzed recently published Ribo-seq (ribosome sequencing) data using ribosome footprints to infer SARS-CoV-2 coding capacity (Finkel et al., 2020). We counted the number of JSs 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). We 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 JS in two or more samples (Figures 5F, 5G, and 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 Table S3 and motivate further studies to validate their biogenesis and biological functions.

DISCUSSION

Understanding the process and mechanism of SARS-CoV-2 sgRNA biogenesis is crucial for identifying potential anti-viral

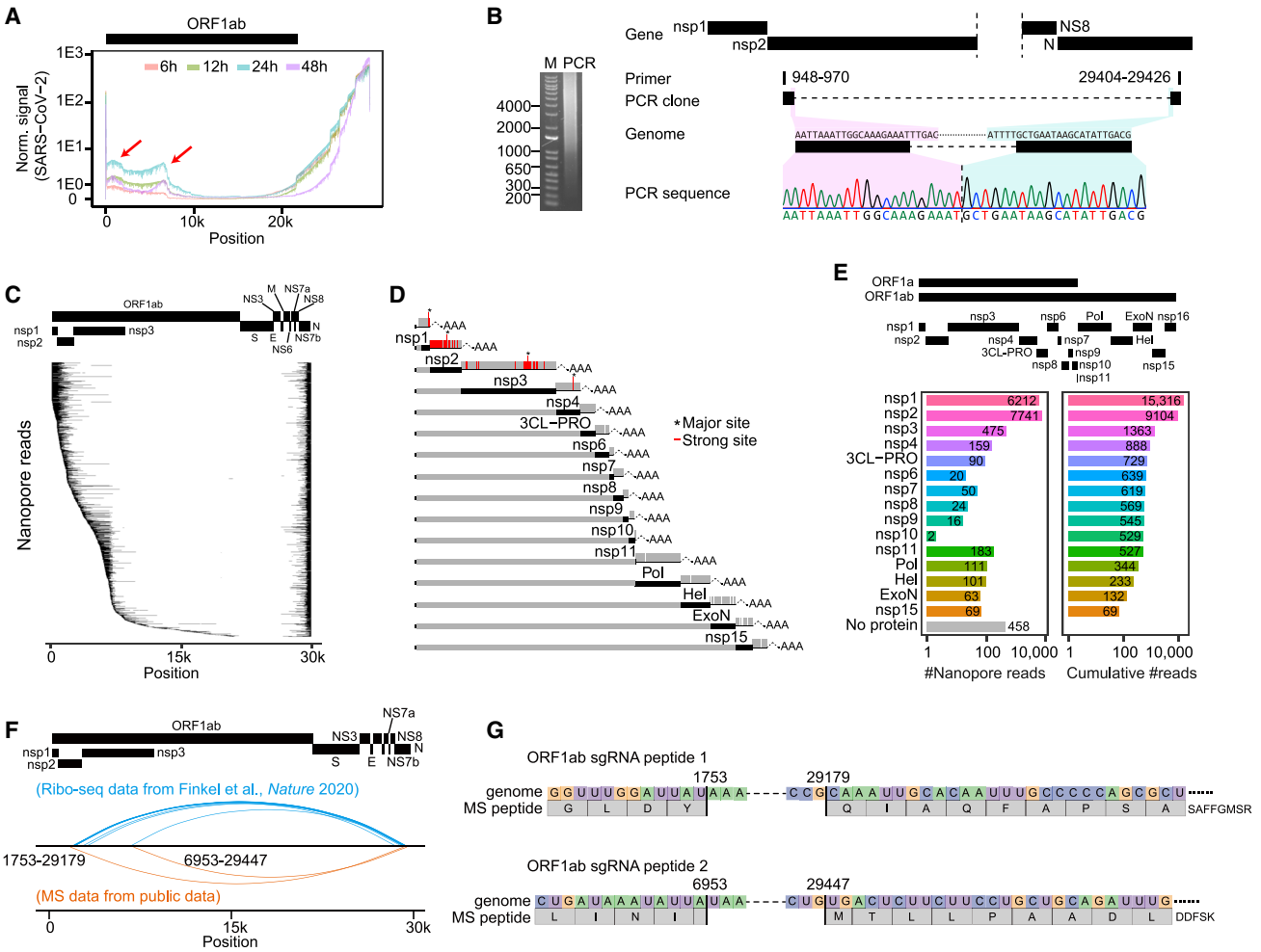


Figure 5. Extensive fusion subgenomes between ORF1ab and the N RNA region

(A) Nanopore read coverage profiles across the SARS-CoV-2 genome for different time points after infection of Vero E6 cells. The arrows in the 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 two primers. The locations of primers, genome sequences, and cloned sequences are shown on the right.

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

(D) Distribution of the upstream JS for ORF1ab sgRNAs (Vero E6 cells, 48 h). The sgRNA types were assigned according to the last protein upstream of the junction in ORF1ab. Strong sites are shown as a red line, whereas major sites are marked by asterisks.

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

(F) Illustration of 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.

drug targets. SARS-CoV genomes are known to generate several subgenomes through template switching during negative genome synthesis, mediated by interactions between leader and body TRS elements. Our results revealed diverse modes of subgenome biogenesis (Figure S7J) and several key determinants of RRIs affecting template switching efficacy (Figure S7K).

Wu and Brian (2010) have shown that transfected bovine coronavirus (BCoV) subgenomes can function as templates for negative-strand synthesis, and our data further confirm this mode by identifying many sgRNAs with multi-switch events in SARS-CoV-2-infected cells. Besides known positive-to-positive template switching, we discovered negative-to-negative template switching during (+)-strand synthesis. Wu and Brian (2007) have reported a special sgRNA ambisense chimera resulting from in *trans* positive-to-negative-strand template switching in bovine CoV, and more complex modes of sgRNA synthesis merit further investigation.

Previous studies hypothesized a three-step model of CoV transcription, including initiation pre-complex formation, base-pairing scanning by the pre-complex, and template switching (Sola et al., 2015). Our results provide detailed features of base-pairing scanning for efficient template switching. It has

also been reported that formation of local secondary structures or high-order RNA structures downstream of switching sites is important to pause continuous transcription and, thus, promote switching (Mateos-Gomez et al., 2013; Nicholson and White, 2014). Co-variational mutation analysis of multiple genomes has found conserved structural RNA elements in the terminal regions of *Alphacoronavirus* genomes (Madhugiri et al., 2014). However, at the moment, similar analyses are hindered by the low mutation rate observed in SAR-CoV-2 or would require 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 (selective 2′-hydroxyl acylation analyzed by primer extension coupled with mutational profiling; Smola et al., 2015) and PARIS (psoralen analysis of RNA interactions and structures; Lu et al., 2016) together, would be useful for decoding the mechanism from the perspective of the RNA structure and RRI network. Several RNA binding proteins (RBPs) have been proposed to participate in the biogenesis of subgenomes (Sola et al., 2011). Candidate RBPs can be identified systematically identified from specific RNA capture assays followed by MS. Functional assays, such as RBP knockout, and *in vivo* binding assays, such as cross-linking and immunoprecipitation with high-throughput sequencing (CLIP-seq; Xue et al., 2009), could be used to validate the roles and mechanisms of regulatory RBPs. The mechanisms of viral RdRP pausing and jumping are still unclear and need further investigation.

We did not capture the negative-strand intermediates of CoV gRNAs and sgRNAs because 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 non-polyadenylated RNA-seq method to sequence and quantify the negative subgenomes. Furthermore, time-course nascent RNA-seq is a promising strategy and could be used in future studies to show dynamic maps of RNAs during viral genome replication and transcription.

Our results provide a quantitative high-resolution map of subgenome structures. Which 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 sgRNAs are similar to the defective interfering RNAs (DI-RNAs) found in some RNA viruses, including HCoV-229E, as reported recently (Pathak and Nagy, 2009; Viehweger et al., 2019).

Limitations of study

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 sgRNA biogenesis in tissues under physiological conditions. The regulatory features we report to govern template switches are predicted by computational RNA-RNA base pairing analysis and could be verified by 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 because of the potential risk of creating artificial, highly pathogenic CoVs.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Cell Culture and Virus Infections
- METHOD DETAILS
  - Northern blotting and simulation
  - Reverse transcription
  - Poly(A) RNA sequencing
  - Nanopore direct RNA sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Published data collection
  - Mapping of NGS RNA-seq data
  - Processing of Nanopore data
  - Identification of significant junction sites
  - Reconstruction of subgenome
  - Characterization of pairing rules
  - Motif analysis
  - Noncanonical template switch analysis
  - RT-PCR of sgRNAs
  - Mass Spectrometry data analysis

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2021.02.036.

ACKNOWLEDGMENTS

The authors thank Dr. Zhihong Hu's lab at 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), the National Science and Technology Major Project (2018YFA0900801), and the Special Fund for COVID-19 Research of Wuhan University. We are grateful to the Beijing Taikang Yicai Foundation for support. Part of the computation of 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.L., J.F., D.G., M.S., F.L., Q.Z., and M.G. performed the experiments. D.W., Y.P., S.W., Y.Z., X.Y., S.L., and X.L. analyzed the sequencing data. D.W. and K.W. analyzed the MS data. Q.L. and J.F. designed primers and probes. X.-L.Y. and Z.-L.S. provided virus-infected cell lines. Y.Z., Y.C., K.L., D.W., and M.C. wrote the manuscript with input from all authors. All authors discussed the results and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 13, 2020
Revised: October 28, 2020
Accepted: February 24, 2021
Published: March 3, 2021

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., and de Hoon, M.J. (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., and Matthews, D.A. (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* (D1), D1100–D1106.

Di, H., Jr., Madden, J.C., Jr., 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. USA *114*, 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. *20*, 533–534.

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 *589*, 125–130.

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., Tien, P., and Guo, D. (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* (D1), D335–D342.

Kim, D., Lee, J.Y., Yang, J.S., Kim, J.W., Kim, V.N., and Chang, H. (2020). The Architecture of SARS-CoV-2 Transcriptome. Cell *181*, 914–921.e10.

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. J. *17*, 10.

Mateos-Gomez, P.A., Morales, L., Zuñiga, 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* (D1), 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., van den Born, E., Spaan, W.J., and Snijder, E.J. (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., Steffen, P., Hochsmann, M., and Giegerich, R. (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*, 59–126.

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., Zúñiga, 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, Á., Hertzig, T., Schelle, B., Bayer, S., Weißbrich, B., Snijder, E.J., Rabenau, H., Doerr, H.W., et al. (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., Zheng, X., Qin, A., Zhou, J., Guo, B., Chen, Y., Wen, X., Ye, W., Zhou, Y., and Zhu, Y. (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. USA *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.

Zúñiga, 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.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
SARS-CoV-2 virus (WIV04)	Wuhan Institute of Virology	IVCAS 6.7512
Critical commercial assays		
Direct RNA sequencing kit	Oxford Nanopore Technologies	SQK-RNA002
NEB Directional RNA Library Prep Kit for Illumina	New England BioLabs	E7760S
SuperScript III Reverse Transcriptase	ThermoFisher Scientific	18080093
KOD -Plus- Neo DNA Polymerase	TOYOBO	KOD-401
TRIzol Reagent	Invitrogen	15596026
Deposited data		
Raw sequencing data	This paper	BIG Data Center (https://bigd.big.ac.cn/): GSA:CRA002508 and GSA-human:HRA000412 under project PRJCA002477
Human reference genome NCBI build 38, GRCh38	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
Chlorocebus sabaeus Ensembl v99	Ensembl database	ftp://ftp.ensembl.org/pub/release-99/fasta/chlorocebus_sabaeus/
SARS-CoV-2 reference genome	NCBI nucleotide database	https://www.ncbi.nlm.nih.gov/nuccore/MN996528.1/
SARS reference genome	NCBI nucleotide database	https://www.ncbi.nlm.nih.gov/nuccore/NC_004718.3/
MERS reference genome	NCBI nucleotide database	https://www.ncbi.nlm.nih.gov/nuccore/NC_038294.1/
HCoV-229E reference genome	Viehweger et al., 2019	Open Science Framework: UP7B4
Public SARS-CoV-2 Nanopore and NGS data	Kim et al., 2020	Open Science Framework: 8F6N9
Public SARS and MERS NGS data	NCBI SRA database	SRP: SRP056612
Public HCoV-229E Nanopore and NGS data	Viehweger et al., 2019	Open Science Framework: UP7B4
The conserved 5′UTR secondary structures for SARS-CoV-2 and HCoV-229E viruses	Rfam database	Rfam: RF03116, RF03117
Public SARS-CoV-2 Mass Spectrometry data	ProteomeXchange database	http://www.proteomexchange.org/: PXD018241, PXD021120, PXD018594
Experimental models: cell lines		
African green monkey kidney cells (Vero E6)	ATCC	CRL-1586
Human colorectal adenocarcinoma cells (Caco-2)	ATCC	HTB-37
Oligonucleotides		
Negative probe used in Northern blot: complementary to the 3′ end (positions 29090 to 29870) of SARS-CoV-2 positive genome; Positive probe: complementary to the negative probe	This paper	N/A
Primers for RT-PCR of sgRNAs	This paper	Mendeley Data: https://doi.org/10.17632/6z78x3fds9.1

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
cutadapt	Martin, 2011	https://cutadapt.readthedocs.io/en/stable/
STAR	Dobin et al., 2013	https://github.com/alexdobin/STAR
guppy	Oxford Nanopore Technologies	https://nanoporetech.com/
NanoPlot	Oxford Nanopore Technologies	https://github.com/wdecoster/NanoPlot
nanopolish	Loman et al., 2015	https://github.com/jts/nanopolish
minimap2	Li, 2018	https://github.com/lh3/minimap2
RNAhybrid	Rehmsmeier et al., 2004	https://bibiserv.cebitec.uni-bielefeld.de/rnahybrid
MaxQuant	Tyanova et al., 2016	https://www.maxquant.org/
Other		
Source codes for the analyses	This paper	https://github.com/zhouyulab/cov2sg/

RESOURCE AVAILABILITY

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).

Materials availability

This study did not generate unique reagents.

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 numbers BIGD: CRA002508 for Vero E6 data and BIGD: HRA000412 for Caco-2 data under project PRJCA002477. The source codes for all the analysis including workflows in Snakemake (Köster and Rahmann, 2018) and scripts in Python and R are available at the https://github.com/zhouyulab/cov2sg/ in GitHub.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

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% CO_2. Cells were infected at a multiplicity of infection of 0.1 with plaque purified SARS-CoV-2 virus from Vero E6 cells (WIV04, IVCAS 6.7512) provided by Dr. Zheng-Li Shi's lab in the Wuhan Institute of Virology (Zhou et al., 2020).

METHOD DETAILS

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 negative-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: $lg(M) = -bm + k$. 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.

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.

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).

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).

QUANTIFICATION AND STATISTICAL ANALYSIS

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 Table S1. The conserved 5′UTR secondary structures for SARS-CoV-2 and HCoV-229E viruses were retrieved from the Rfam database (Kalvari et al., 2018) with accession number RF03117 (Betacoronavirus) and RF03116 (Alphacoronavirus), respectively. The information of Ribo-seq reads covering sgRNA JSs was downloaded from the Table S1 of recent study (Finkel et al., 2020). The published Mass Spectrometry datasets were downloaded from ProteomeXchange database (Deutsch et al., 2017) under identifiers PXD018241 (Davidson et al., 2020), PXD018594 (Grenga et al., 2020), and PXD021120.

Mapping of NGS RNA-seq data

The adaptors in raw reads were removed using cutadapt (v2.5) program (Martin, 2011). After filtering out potential ribosomal RNAs, the clean reads were first mapped to the host genome (Vero E6: Chlorocebus sabaeus Ensembl v99; Caco-2 and Calu-3: human hg38) using STAR (v2.7.2b) program (Dobin et al., 2013) 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.

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 (De Coster et al., 2018). Poly(A) tails were detected by nanopolish (v0.12.3) (Loman et al., 2015). 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) (Li, 2018). The reads belonging to the viral genome are used in further analysis.

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 first 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 $S_{upstream}$ and $S_{downstream}$, 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 N = #JS / sqrt($S_{upstream}$ x $S_{downstream}$). 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 dataset. 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 (Figures S2C and S2D).

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 (Figure 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 (Cock et al., 2009) 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.

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 (Rehmsmeier et al., 2004) 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 whose 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.

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.

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 (Figures 1G, 1H, S2F, and S2G). 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.

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.

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 (Tyanova et al., 2016) 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 both sides 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.