The short- and long-range RNA-RNA Interactome of SARS-CoV-2

Omer Ziv[1,4,*], Jonathan Price[1,4], Lyudmila Shalamova[2,4], Tsveta Kamenova[1], Ian Goodfellow[3], Friedemann Weber[2,*], Eric A. Miska[1,5,6,*]

[1] Wellcome Trust/Cancer Research UK Gurdon Institute and Department of Genetics, University of Cambridge, Cambridge, CB2 1QN, UK
[2] Institute for Virology, FB10-Veterinary Medicine, Justus-Liebig University, Gießen 35392, Germany
[3] Division of Virology, Department of Pathology, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 2QQ, UK
[4] These authors contributed equally
[5] Wellcome Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, UK
[6] Lead contact

* Correspondence: omer.ziv@gurdon.cam.ac.uk (O.Z.); friedemann.weber@vetmed.uni-giessen.de (F.W.), eric.miska@gurdon.cam.ac.uk (E.A.M.)

SUMMARY

The *Coronaviridae* is a family of positive-strand RNA viruses that includes SARS-CoV-2, the etiologic agent of the COVID-19 pandemic. Bearing the largest single-stranded RNA genomes in nature, coronaviruses are critically dependent on long-distance RNA-RNA interactions to regulate the viral transcription and replication pathways. Here we experimentally mapped the *in vivo* RNA-RNA interactome of the full-length SARS-CoV-2 genome and subgenomic mRNAs. We uncovered a network of RNA-RNA interactions spanning tens of thousands of nucleotides. These interactions reveal that the viral genome and subgenomes adopt alternative topologies inside cells, and engage in different interactions with host RNAs. Notably, we discovered a long-range RNA-RNA interaction - the FSE-arch - that encircles the programmed ribosomal frameshifting element. The FSE-arch is conserved in the related MERS-CoV and is under purifying selection. Our findings illuminate RNA structure based mechanisms governing replication, discontinuous transcription, and translation of coronaviruses, and will aid future efforts to develop antiviral strategies.

Keywords: Coronavirus, SARS-CoV-2, COVID-19, RNA structure, RNA-RNA interaction, host-virus, discontinuous transcription, Ribosomal Frameshifting, FSE-arch, COMRADES

INTRODUCTION

RNA viruses comprise the dominant component of the eukaryotic virome (Dolja and Koonin, 2018). Their error-prone genome replication mode allows them to rapidly evolve new variants and to jump from animals to humans (Woolhouse and Gaunt, 2007), thus presenting a high epidemic and pandemic threat. Several members of the betacoronavirus genus (family *Coronaviridae*), namely the Severe Acute Respiratory Syndrome coronavirus (SARS-CoV), the Middle East Respiratory Syndrome coronavirus (MERS-CoV), as well as the Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2) are of special concern. SARS-CoV-2, the causative agent of Coronavirus Disease 2019 (COVID-19), has spread to date to nearly every country in the world, resulting in millions of infections, over a million of deaths, and a massive global economic impact (McKibbin and Fernando, 2020). Even though worldwide efforts and resources are redirected to overcome the COVID-19 pandemic, at present, there are no approved vaccines or antiviral medicines. This illustrates the urgent need for deciphering in-depth the molecular biology of coronaviruses, especially SARS-CoV-2.

Coronaviruses have evolved the largest known single-stranded RNA genome in nature. Regulation of their mRNA transcription and translation is facilitated by cis-acting structures that interact with each other, with viral proteins, and with host machineries (Madhugiri et al., 2016). mRNA transcription in coronaviruses involves a process whereby so-called subgenomic mRNAs (sgmRNAs) are produced through discontinuous genomic RNA (gRNA) template utilization, which is in contrast to replication of the full-length genome (Sawicki et al., 2007). This discontinuous transcription is mediated by the Transcription Regulating Sequence-leader (TRS-L) at the 5′ end of the genome, and the Transcription Regulating Sequence-body (TRS-B) at the 5′ ends of each ORF. Template switching between these RNA sequence elements results in a set of 5′ and 3′ co-terminal, "nested" sgmRNAs of different sizes on which the 5′ proximal ORFs are translated into nonstructural or structural viral proteins (Moreno et al., 2008; Sola et al., 2015). The mechanisms underlying discontinuous transcription and genome replication have not been fully worked out, however long-distance RNA-RNA interactions along the viral genome have been proposed as key regulators (Mateos-Gómez et al., 2011; Mateos-Gomez et al., 2013; Moreno et al., 2008; Sola et al., 2015).

On the full-length gRNA itself, two partially overlapping open reading frames (ORF1a and ORF1b) are translated from the same start codon at the 5′ end, resulting in the polyproteins pp1a and pp1ab. Translation of the longer product pp1ab is made possible by a hairpin-type pseudoknot RNA structure known as the frameshifting element (FSE) which regulates a programmed -1 ribosomal frameshifting that overrides with about 50% efficiency the stop codon of ORF1a (Kelly et al., 2020; Namy et al., 2006). Previous studies applied RNA structure probing techniques using selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE) and DMS, as well as nuclear magnetic resonance (NMR) to effectively identify conserved cis-acting RNA structures regulating the life cycle of coronaviruses. However, when it comes to identifying long-distance base-pairing between distal nucleotides, these methods fall short. Therefore, the long-range RNA-RNA interactome of coronaviruses has never been mapped in full. Deciphering how the various structural elements along the coronavirus gRNA and sgmRNA are folded and brought together in time and space is vital for understanding, dissecting and manipulating viral replication, discontinuous transcription, and translation regulation.

We recently developed Crosslinking Of Matched RNAs And Deep Sequencing (COMRADES) for in-depth RNA conformation capture in living cells (Ziv et al., 2018). COMRADES is derived from a class of methods that combine psoralen crosslinking of base paired RNA and deep sequencing (Aw et al., 2016; Lu et al., 2016; Sharma et al., 2016). COMRADES utilises a clickable psoralen derivative to specifically crosslink paired nucleotides, and high throughput sequencing to retrieve their positions (Figure 1). Following *in vivo* crosslinking, the viral RNA is selectively captured, fragmented and subjected to a click-chemistry reaction to add a biotin tag to crosslinked fragments. Crosslinked RNA duplexes are then selectively captured using streptavidin affinity purification. Half of the resulting RNA is proximity ligated, following reversal of the crosslink to create chimeric RNA templates for high throughput sequencing. The other half is used as a control, in which reversal of the crosslink precedes the proximity ligation, and accurately represents the background level of non-specific ligation. The coupling of two biotin-streptavidin mediated enrichment steps, first of viral RNA, and second of crosslinked RNA duplexes provides high structural depth for identification of both long- and short-lived conformations. COMRADES can therefore measure *(i)* the structural diversity of alternative RNA conformations that co-exist inside cells; *(ii)* short-distance, as well as long-distance (over tens of thousands of nucleotides) base-pairing within the same RNA molecule; and *(iii)* base-pairing between different RNA molecules, such as those of host and viral origin (Kudla et al., 2020; Ziv et al., 2018).

Here we apply COMRADES to study the structural diversity of the SARS-CoV-2 gRNA and sgmRNA inside cells. We discover networks of short- and long-range RNA-RNA interactions spanning the entirety of SARS-CoV-2 gRNA and sgmRNA. We reveal site-specific interactions with the host transcriptome. Finally, we uncover a conserved long-range structure encompassing the programmed ribosomal frameshifting element. In order to make our data more accessible to the community, we have developed an interactive user-friendly interface for exploring the structures identified in this study.

RESULTS AND DISCUSSION

The SARS-CoV-2 genome and sgmRNA adopt alternative co-existing topologies that involve long-distance base-pairing

Inside the host, the gRNA of SARS-CoV-2 is transcribed into sgmRNA (Figure 2A). To compare the structure of both types of RNA, we applied the COMRADES method and set up a dual enrichment strategy to analyse the positive sense gRNA and positive sense sgmRNA separately (Figure 2B). Briefly, we selectively pulled down the full-length positive sense SARS-CoV-2 genome from *in vivo* crosslinked, SARS-CoV-2 inoculated Vero E6/TMPRSS2 cells (Matsuyama et al., 2020), using a tiling array of antisense probes for ORF1a/b, which resulted in a highly enriched gRNA fraction (Figure 2C). The full-length positive sense sgmRNA was subsequently enriched from the gRNA-depleted supernatant of the first pulldown, using a second tiling array of antisense probes to the region downstream of ORF1a/b (Figure 2B). This dual enrichment strategy resulted in a high degree of separation between the gRNA and the sgmRNA (Figure 2C). COMRADES provided >6 million non-redundant chimeric reads, which was sufficient to generate high-resolution maps for both the gRNA and sgmRNA with a high signal to noise ratio (Figure S1), and high reproducibility between independent biological replicates (r = 0.92, p value <2.2e-16, Figure 2D,E). Our structural data covered >99.99% of the coronavirus gRNA and the sgmRNA (Figure 2C), and represents the base-pairing nature of SARS-CoV-2 gRNA and sgmRNA inside cells.

Available models for the RNA structure of SARS-CoV-2 and related viruses are largely confined to short-distance base-pairing which result in local folding of important cis-acting elements (Andrews et al., 2020; Huston et al., 2020; Kelly et al., 2020; Lan et al., 2020; Manfredonia et al., 2020; Ryder, 2020; Sanders et al., 2020; Sun et al., 2020). However, long-distance base-pairing between distal RNA elements are equally essential for many RNA viruses (Huber et al., 2019), including coronaviruses (Mateos-Gómez et al., 2011; Mateos-Gomez et al., 2013; Moreno et al., 2008; Sola et al., 2015). The ability of COMRADES to capture RNA base-pairing regardless of the distance between the interacting bases enabled us to confirm *in vivo* the structure of nearly all previously characterised cis-acting elements (with one exception, discussed below) and to discover long-distance RNA-RNA interactions as they occur inside cells. Indeed, we observed a high prevalence of long-range RNA base-pairing along the SARS-CoV-2 genome, with ORF1a demonstrating more long-range connectivity than any other ORF (Figure 2F). Most of the base-pairing is confined to a single ORF, however, some interactions cross ORF boundaries. For example, ORF1a base-pairs with ORF1b, as well as with the 5′ and 3′ untranslated regions (UTRs) (Figure 2F). We additionally discovered long-distance interactions unique to the sgmRNA (Figure 2G). Previous models of the SARS-CoV-2 and related viruses mainly analysed structural population averages, i.e. assuming that all copies of the genome and sgmRNA have a single static conformation. Yet, the complex life cycle of viral RNA genomes, i.e. their engagement with multiple cellular and viral machineries such as the ones for replication, transcription, and translation, suggests a dynamic RNA structure, as we and others have reported for Zika virus (Huber et al., 2019; Li et al., 2018; Ziv et al., 2018) and for HIV-1 (Tomezsko et al., 2020). Our structural analysis of SARS-CoV-2 reveals a high level of structural dynamics whereby alternative high-order conformations, some of which involve long-distance base-pairing, co-exist *in vivo* (Figures 3 and S2A, Table S1). For example, nucleotides 5,660-5,680 in ORF1a interact with three alternative distal regions: 3.6 kb upstream, 3.4 kb downstream, and 2 kb upstream (Figure 3, arches 4, 5 and 8 respectively), and the 5′ UTR interacts with ORF1a as well as with the 3′ UTR (Figure 3, arches 2 and 3 respectively). In contrast, we find that ORF N sgmRNA is held in a single dominant conformation where the leader sequence interacts exclusively with a region 0.8 kb downstream (Figure S2B). In summary, we discover the co-existence of alternative SARS-CoV-2 gRNA and sgmRNA topologies, held by long-range base-pairing between regions tens of thousands of nucleotides apart. Each topology brings in physical proximity previously characterised and new elements involved in viral replication and discontinuous transcription, therefore offering a model for facilitating distinct patterns of template switching to produce the complete SARS-CoV-2 transcriptome.

The SARS-CoV-2 genome engages in different interactions with cellular host RNA

The infectious life cycle of coronaviruses takes place mainly in the host cell's cytoplasm, where many cellular RNAs reside (Sola et al., 2015). Host-virus RNA-RNA interactions regulate the replication of some RNA viruses, e.g. the interaction between hepatitis C virus and human microRNA miR-122 (Jopling et al., 2005), the interaction between Zika virus and human miR-21 (Ziv et al., 2018), and the priming of HIV-1 replication by human tRNAs (Mak and Kleiman, 1997). However, to the best of our knowledge, whether the SARS-CoV-2 gRNA or sgmRNA interact with cellular RNA is unknown. Our COMRADES method provides an opportunity to undertake an unbiased analysis of the host-virus RNA-RNA interactome (Ziv et al., 2018). We discovered site-specific interactions between the SARS-CoV-2 RNA and various cellular RNAs, especially small nuclear RNAs (snRNAs) (Figures 4A,B and S3A,B). Apart from their canonical role in splicing, snRNAs mature in the cytoplasm and may have additional biological roles (Matera et al., 2014). Along the viral gRNA, cellular snRNA interactions are mostly confined to ORF1a and ORF1b, and include site specific binding of U1, U2 and U4 snRNAs. The gRNA coding region for the sgmRNA ORFs and the UTRs are largely devoid of snRNA binding. In contrast, along the viral sgmRNA, both the N ORF and the 3′ UTR show high occupancy of U1 and U2 snRNA binding. In order to explore the conservation of these snRNA interactions in a related coronavirus, we performed COMRADES on MERS-CoV-inoculated Huh-7 cells. Similarly to SARS-CoV-2, we identified a site specific interaction of U2 snRNA within the MERS-CoV ORF1a (Figures 4C and S3C), illustrating the evolutionary conservation of the U2 snRNA base-pairing with ORF1a of betacoronaviruses. In addition to cellular small RNAs we also detected long cellular RNAs interacting with SARS-CoV-2 RNA, although to a lesser extent. Of specific interest, the ribonuclease (RNase) MRP RNA was found to base-pair with an extended 3′ region of the sgmRNA, but not the gRNA, of SARS-CoV-2 (Figure S3D). The RNase MRP RNA has a conserved secondary structure similar to that of the RNA component of the bacterial RNase P ribonucleoprotein (RNP) complex (Dávila López et al., 2009; Welting et al., 2006). The RNase MRP RNA has a role in human pre-ribosomal RNA processing (Goldfarb and Cech, 2017), when mutated leads to a spectrum of human disease (Ridanpää et al., 2001), and has been implicated in viral RNA degradation (Jaag et al., 2011). Targeting host-virus RNA-RNA interactions provides an attractive platform for developing new antiviral therapies, as resistance would require the virus to acquire considerable mutational changes to become independent of the host RNA. However, whereas multiple tools and efforts are dedicated to identifying host-virus protein-protein interactions, the crosstalk between host and virus RNA remains largely unexplored. Coupled with the recent advancement in techniques to target RNA *in vivo*, COMRADES's capacity to map the host-virus RNA-RNA interactome opens up new opportunities to control emerging RNA viruses. The data we present here could be valuable for the development of new targets for antiviral drugs.

The UTRs of SARS-CoV-2 interact with distal genomic regions and with each other

The 5′ UTR of coronaviruses contain five evolutionary conserved stem-loop structures (denoted SL1-SL5) that are essential for genome replication and discontinuous transcription (Madhugiri et al., 2016). The 3′ UTR contains 3 structural elements important for replication: an evolutionary conserved bulged stem-loop (BSL) (Hsue and Masters, 1997), a partially overlapping hairpin-type pseudoknot (Goebel et al., 2004; Williams et al., 1999), and a 3′ terminal multiple stem-loop structure containing a hyper-variable region (HVR), which folds back to create a triple helix junction (Liu et al., 2013). Our analysis identified seven of these eight cis-acting elements within the UTRs (Figures 5A and S4A). However, our data did not support the folding of the stem-loop pseudoknot at the 3′ UTR. Of note, two recent studies using SHAPE methods to map the structure of SARS-CoV-2 inside cells similarly failed to identify this pseudoknot (Huston et al., 2020; Sun et al., 2020), and a previous study demonstrated the instability of this pseudoknot in the related mouse hepatitis virus (MHV) (Stammler et al., 2011).

In addition to the canonical UTR structures, we provide here a direct *in vivo* evidence for genome cyclization in SARS-CoV-2, mediated by long-range base-pairing between the 5′ and 3′ UTRs (Figures 5B and S4B). This base-pairing spans a distance of 29.7 kilobases and is among the longest distance RNA-RNA interactions ever reported. Genome cyclization was previously hypothesised from mutational analyses of murine coronavirus (MHV) and was suggested to facilitate discontinuous transcription (Li et al., 2008). However, while MHV genome cyclization involves the 5′ SL1 structure, we find that in SARS-CoV-2, this process is mediated by the 5′ SL3 instead, and results in complete opening of SL3 and disruption of the triple helix junction in the 3′ UTR (Figure 5B). In agreement with this observation, SL3 of related betacoronaviruses was suggested to be weakly folded or unfolded (Chen and Olsthoorn, 2010; Li et al., 2008). Genome cyclization plays an essential role in the replication of a number of RNA viruses, including flaviviruses (Hahn et al., 1987; Ziv et al., 2018). The evolutionary selection of such a mechanism might stem from in-cell competition between intact and defective viral genomes, as it ensures that only genomes bearing two intact UTRs engage with the replication machinery. The SARS-CoV-2 genome cyclization we report here results in a complete opening of the 5′ SL3 where the Transcription Regulating Sequence-Leader (TRS-L) resides, raising the possibility that genome cyclization regulates SARS-CoV-2 discontinuous transcription, as was previously suggested for MHV (Li et al., 2008). It remains to be seen whether this base-pairing can be targeted to inhibit viral replication *in vivo*.

In addition to genome cyclization, we identified two alternative conformations involving long-distance RNA-RNA interactions between each UTR and ORF1a. These long-distance conformations result in unfolding of SL2 and SL3 in the 5′ UTR (Figures 5C and S4C), and unfolding of the terminal stem-loop in the 3′ UTR (Figures 5D and S4D). Of note, unlike the gRNA, the leader sequence within the 5′ UTR of ORF N sgmRNA is held in a single long-range conformation through base-pairing with a region 0.8 kb downstream (Figures 5E, S2B and S4E). All of the long-range interactions described above are strongly supported through chimeric reads (Figures 5F and S2A). Overall, our data demonstrate the existence of alternative, mutually-exclusive UTR conformations inside cells, involving interactions between functional UTR elements and distal regions within the ORFs. We further show that the N ORF sgmRNA folds differently than the viral genome. The long-distance RNA structure map for SARS-CoV-2 provides a practical starting point to dissect the regulation of discontinuous transcription, as it identifies cis-acting elements that interact with each other to create genome topologies that favour the synthesis of the ensemble of sgmRNAs.

A longe-range structural arch encircling the SARS-CoV-2 frameshifting element (FSE-arch) is under purifying selection

RNA viruses evolve sophisticated mechanisms to enhance the functional capacity of their size-restricted genomes and to regulate the expression levels of their replicase components. In coronaviruses, one such mechanism is programmed -1 ribosomal frameshifting to facilitate translation of ORF1b which contains the viral RdRp activity, and to set a defined ratio of ORF1a and ORF1b products (Plant et al., 2010). This is mediated by a ~120 nucleotide long cis-acting frameshifting element (FSE) composed of a stem-loop attenuator, and a slippery sequence followed by a single-stranded spacer and an RNA pseudoknot (Kelly et al., 2020). It has been suggested that pausing the progression of the ribosome upstream of the pseudoknot facilitates a tandem-slippage of the peptidyl-tRNA and aminoacyl-tRNA to the −1 reading frame, thus allowing continuous translation through the stop codon at the end of ORF1a (Brierley et al., 1989). Altering the frameshifting mechanism had a deleterious effect on SARS-CoV replication (Plant et al., 2013), making the FSE an attractive target for antiviral therapy. Understanding the surrounding RNA structure and function is therefore of great importance as it might aid the design of drugs targeting the FSE. Unexpectedly, we find that the FSE of SARS-CoV-2 is embedded within a much larger, ~1.5 kb long higher-order structure that bridges the 3′ end of ORF1a with the 5′ region of ORF1b, which we termed the FSE-arch (Figure 6A,B). To the best of our knowledge, this is the first time such a long-range structural bridge has been reported for any coronavirus, and importantly this structure is supported by the largest number of chimeric reads in our data (more than tens of thousands of non-redundant chimeric reads) (Figures S6A,B), reflecting its high folding stability *in vivo*. The FSE-arch results in a stem-loop structure encompassing 1,475 nucleotides, and bearing the FSE within it (Figures 6B and S5C). We hypothesized that if an RNA-RNA interaction is functionally important, there should be purifying selection and hence reduced nucleotide evolution rate in this region. Therefore we used a recent dataset (Firth, 2020) to explore the nucleotide conservation of the FSE-arch (Figure 6C). Strikingly, the FSE-arch is under a strong purifying selection and is among the most conserved regions within the SARS-CoV-2 genome. Consistent with this, analysing the phylogeny of the SARS-related coronavirus subgenus (taxid: 2509511) revealed two positions of covariance that support the conservation of the FSE-arch (Figure 6B). To further explore this structure experimentally, we analysed its existence in MERS-CoV. MERS-CoV shares only ~50% sequence identity with SARS-CoV-2 (Chen et al., 2020; Lu et al., 2020), yet even so, performing COMRADES on MERS-CoV-inoculated Huh-7 cells revealed a strong evidence for an homologous FSE-arch surrounding the MERS-CoV FSE, bridging ORF1a with ORF1b (Figure 6D,E). While the mechanism governing the FSE-arch formation will require further investigation, similar long-distance interactions around the frameshifting elements of several plant RNA viruses were previously demonstrated to regulate frameshifting, possibly by assisting in back-stepping of ribosomes at the slippery sequence, and by stabilising the FSE, allowing it to refold after the passage of each ribosome (Barry and Miller, 2002; Cimino et al., 2011; Gao and Simon, 2016; Tajima et al., 2011).

In addition to their coding capacity, nucleic acids have evolved structural capabilities to sense metabolites (Mandal and Breaker, 2004), catalyse reactions (Pyle, 1993), and interact with other cellular components. When brought in physical proximity, different combinations of cis-acting sequences can lead to new biological activities. For example, interactions between promoters and enhancers dictate the rate of transcription along the eukaryotic genome (Rowley and Corces, 2018). Great effort is being made to reveal the structural landscapes of the SARS-CoV-2 genome (Andrews et al., 2020; Huston et al., 2020; Kelly et al., 2020; Lan et al., 2020; Manfredonia et al., 2020; Ryder, 2020; Sanders et al., 2020; Sun et al., 2020). However, without deciphering the long-range connectivity, our understanding is far from being complete. Here we reveal how cis-acting elements along the coronavirus genome are folded and alternate between different topologies to create spatial combinations of functional RNA elements. The combinatorial nature of the coronavirus genome inside the host cell as discussed here, provides molecular insights into the replication, discontinuous transcription and ribosomal frameshifting machineries of coronaviruses and will facilitate the discovery of new functional cis-acting elements and the design of RNA-based antiviral therapies for SARS-CoV-2. To accelerate the research of the SARS-CoV-2 RNA-RNA interactome by the scientific community, we developed a freely downloadable interactive web interface for visualisation and exploration of our structural data, as well as for its integration with related datasets (https://github.com/JLP-BioInf/SARS-CoV2-COMRADES-APP).

Limitations

(I) Intramolecular base-pairing within the same viral genome is more likely to occur than intermolecular base-pairing between two different viral genomes. However, we cannot rule out the possibility that some of the long-range interactions reported here might stem from intermolecular interactions between different viral genomes. *(ii)* snRNAs are known to be heavily modified. Since certain chemical modifications may affect the base-pairing capacity of RNA, modeling the exact structure of snRNA - viral interactions (Figure S3B) is particularly challenging, and will benefit from followup experimental work.

ACKNOWLEDGMENTS

The authors thank Christian Drosten and John Ziebuhrfor for providing the SARS-CoV-2 and MERS-CoV strains used in this study. We thank B. Luisi, G. Evan, and the Department of Biochemistry, University of Cambridge for facilitating the study; A. Yerra, Wellcome Trust/Cancer Research UK Gurdon Institute for critical reading and editing of the manuscript; D. Jordan for advice on statistics; I. Brierley, Department of Pathology, University of Cambridge for advising on coronavirus FSE; A.E. Firth, Department of Pathology, University of Cambridge for advising on structure conservation; S. Moss, Wellcome Trust/Cancer Research UK Gurdon Institute for assisting with risk assessments; P. Coupland and K. Kania, Genomics Core Facility, Cancer Research UK Cambridge Institute for processing the sequencing libraries; and C. Flandoli and I Kaminski for illustrations. This work was supported by Cancer Research UK grants (C13474/A18583, C6946/A14492) to E.A.M.; Wellcome grants (104640/Z/14/Z, 092096/Z/10/Z) to E.A.M.; Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 197785619 – SFB 1021 grant to F.W.; the RAPID consortium of the Bundesministerium für Bildung und Forschung (BMBF) grant (01KI1723E) to F.W.; the MAD-CoV-2 consortium that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 101005026. to F.W.; and Wellcome grant (REF207498/Z/17/Z) to I.G.

AUTHOR CONTRIBUTIONS

O.Z., L.S., and F.W. designed the study; O.Z., F.W., I.G., and E.A.M. supervised the study; L.S. carried viral infections and RNA crosslinking; O.Z. performed the COMRADES method and supervised the analysis; J.L.P. designed the analysis pipeline and analysed the data with input from O.Z. and T.K.; T.K. carried covariation analysis; O.Z., J.L.P. and E.A.M. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests. E.A.M. is a founder and director of STORM Therapeutics. STORM Therapeutics had no role in the design, performance, analysis, interpretation, and writing of the study. O.Z is a consultant in Evotec Int. Evotec Int. had no role in the design, performance, analysis, interpretation, and writing of the study.

FIGURE LEGENDS

Figure 1. The COMRADES method.
Virus inoculated cells are crosslinked using clickable psoralen. Viral RNA is pulled down from the cell lysate using an array of biotinylated DNA probes, following digestion of the DNA probes and fragmentation of the RNA. Biotin is attached to crosslinked RNA duplexes via click chemistry, enabling pulling down crosslinked RNA using StreptAvidin beads. Half of the RNA duplexes are proximity ligated, following reversal of the crosslinking to enable sequencing. The other half serves as a control, in which crosslink reversal proceeds the proximity ligation. See Figure S1 for numbers and percentages of chimeric reads.

Figure 2. The *in vivo* base-pairing of the SARS-CoV-2 gRNA and sgmRNA
(A) The organisation of the SARS-CoV-2 gRNA and sgmRNA.
(B) Dual-enrichment strategy for separation of SARS-CoV-2 gRNA and sgmRNA from inside cells.
(C) Coverage of structural data (chimeric reads) for the gRNA and sgmRNA samples aligned to the gRNA coordinates. Each chimeric-read originated from *in vivo* crosslinking of base-paired RNA. Chimeric reads aligned to the Leader in the sgmRNA samples are not shown due to the use of gRNA coordinates.
(D) Heat map of RNA-RNA interactions along the SARS-CoV-2 gRNA. Signal represents base-pairing between the genomic coordinates on the *x* and *y* axes. Top left and bottom right represent two independent biological replicates. Colour code corresponds to the number of non-redundant chimeric reads supporting each interaction.
(E) Left panel: Heatmap of RNA-RNA interactions along the ORF S sgmRNA. Right panel: Zoom-in of the gRNA region from (D) corresponding to ORF S sgmRNA coordinates. Colour code as in (D). Top left and bottom right represent two independent biological replicates.
(F) Arch plot representation of long-range RNA-RNA interactions along the SARS-CoV-2 gRNA. Interactions that span at least 500 nt are shown. Colours represent the number of non-redundant chimeric reads supporting each arch.
(G) Arch plots representation of long-range RNA-RNA interactions along ORF S sgmRNA (bottom) and the gRNA region corresponding to the S sgmRNA coordinates (top). Interactions that span at least 500 nt are shown. Colours as in (F).

Figure 3. Long-range RNA-RNA interactions along the SARS-CoV-2 gRNA
RNA-RNA interactions between regions that are separated by at least 2,000 nucleotides are shown. Top panel illustrates the different patterns assigned to different parts of the genome. Coloured rectangles below the top panel and near each arch number represent the number of non-redundant chimeras supporting each conformation. The sequence of part of the base-pairing is shown above each conformation. Numbers within the loops represent the loop size. See Figure S2 and Table 1S for numbers of chimeric reads and significance of the arches.

Figure 4. Interactions between cellular snRNAs and viral RNA
(A) snRNAs binding positions along the SARS-CoV-2 gRNA (bottom), and sgmRNA (top right). Arrows mark the binding positions of individual snRNAs.
(B) Base-pairing model for the viral gRNA - U1 snRNA interaction. **Ψ** denotes pseudoUridine.
(C) U2 snRNA binding positions along the SARS-CoV-2 gRNA (top), and MERS-CoV (bottom).
T test p values: ** <0.05; *** <0.001. See Figure S3 for base-pairing models and for COMRADES controls.

Figure 5. The UTRs of SARS-CoV-2 adopt alternative conformations inside cells
(A) The canonical SARS-CoV-2 UTRs structure identified in this study. Colours represent the number of non-redundant chimeric reads supporting each base-pair.

(B) RNA structure corresponding to genome cyclization of SARS-CoV-2 inside cells. Colour code as in (A).
(C,D) Long-distance interactions between the 5′ UTR (C) or the 3′ UTR (D) and ORF1a. Colour code as in (A).
(E). Interaction between the 5′ Leader sequence and a downstream region in ORF N sgmRNA. Colour code as in (A).
(F) Representation of the left- and right-side of the chimeric reads supporting the long-range interactions shown in (B-E).
Numbers within loops in (B-E) represent the loops sizes. Grey arches adjacent to nucleotide sequences in (B-E) mark unpaired bases. Full sequences are available in Figure S4.

Figure 6. The structure of the ribosomal frameshifting element arch (FSE-arch) inside cells
(A) Heatmap of RNA-RNA interactions around the SARS-CoV-2 FSE. Signal represents base-pairing between the genomic coordinates on the *x* and *y* axes. Top left and bottom right represent two independent biological replicates. Colour code corresponds to the number of non-redundant chimeric reads supporting each interaction. Arrows indicate the FSE-arch signal.
(B) Arch plot representation of long-range RNA-RNA interactions around SARS-CoV-2 FSE. Conservation significance is shown below the arches. Black rectangle indicates the FSE position. Top Colour code corresponds to the number of non-redundant chimeric reads supporting each arch. Bottom colour code corresponds to the conservation significance (Synplot p values).
(C) The structure around the SARS-CoV-2 FSE. Colours represent the number of chimeric reads supporting each base-pair. Red circles (positions 13,944 and 14,256) indicate covariation positions.
(D) Arch plot representation of long-range RNA-RNA interactions around the MERS-CoV FSE. Colour code corresponds to the number of non-redundant chimeric reads supporting each arch.
(E) Heatmap of RNA-RNA interactions around the MERS-CoV FSE. Signal represents base-pairing between the genomic coordinates on the *x* and *y* axes. Top left and bottom right represent two independent biological replicates. Colour code corresponds to the number of non-redundant chimeric reads supporting each interaction. Arrows indicate the MERS-CoV FSE-arch signal.
See Figure 5S for statistical significance and the full sequence of the FSE-arch.

STAR★Methods

Key Resources Table

The key resources table is supplied as a supplemental file.

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Eric A. Miska (eric.miska@gurdon.cam.ac.uk).

Materials Availability

No unique materials were generated in this study.

Data and Code Availability

All sequencing data sets have been deposited in GEO under accession number GSE154662. Computer code has been deposited on GitHub: https://github.com/JLP-BioInf/SARS-Cov2-COMRADES. Base-pairing prediction, structure prediction and clustering data are available for exploration as a web interface: https://github.com/JLP-BioInf/SARS-CoV2-COMRADES-APP. Additional Supplemental Items are available from Mendeley Data at http://dx.doi.org/10.17632/ghh6w67vmx.1. Additional data supporting the findings of this study are available from the corresponding authors upon request.

Experimental Model and Subject Details

Chlorocebus sabaeus (Green monkey) VeroE6 (female, RRID:CVCL_YQ49) were purchased from American Type Culture Collection (ATCC, id: ATCC CRL-1586). VeroE6/TMPRSS2 cells (female), (Matsuyama et al., 2020) were kindly provided by Prof. Dr. Stefan Pöhlmann (German Primate Center, Göttingen). Vero E6 and VeroE6/TMPRSS2 cells were cultured in Dulbecco's modified Eagles medium (DMEM) supplemented with 10% fetal bovine serum at 37 °C in a humidified CO2 incubator. HuH7 (Homo sapiens, male, adult hepatocellular carcinoma, RRID:CVCL_0336) cells were purchased from the Japanese Collection of Research Bioresources (JCRB) Cell Bank (JCRB No. JCRB0403) and cultured in Dulbecco's modified Eagles medium (DMEM) supplemented with 10% fetal bovine serum at 37 °C in a humidified CO2 incubator. All cell lines were regularly examined to exclude mycoplasma contamination.

Method Details

Viral Infection and psoralen crosslinking. Infection experiments were performed under biosafety level 3 conditions. Independent biological replicates were performed using 90-120 million cells each. Titration of virus stocks was conducted using Vero E6 cells. For SARS-CoV-2 infection, VeroE6/TMPRSS2 cells (PMID: 32165541) were inoculated with SARS-CoV-2 strain München-1.2/2020/984 (Rothe et al., 2020) at MOI=2 pfu/cell for 20 hours. For MERS infection, HuH7 cells were inoculated with MERS-CoV strain EMC/2012 (GenBank: JX869059.2; PMID: 23170002) (van Boheemen et al., 2012) at MOI=2 pfu/cell for 20 hours. Following inoculation, cells were washed 3 times with PBS and were incubated for 20 minutes with 0.4 mg/ml Psoralen-triethylene glycol azide (psoralen-TEG azide, Berry & Associates) diluted in PBS and supplemented with OptiMEM I with no phenol-red (Gibco). Cells were subsequently irradiated on ice with 50 KJ/m2 365 nm UVA using a CL-1000 crosslinker (UVP). Cell lysis was performed by RNeasy lysis buffer (Qiagen) supplemented with DTT. Proteins were degraded using proteinase K (NEB) and RNA was purified using RNeasy maxi kit (Qiagen).

Viral RNA enrichment. Total cellular RNA was mixed with a tiling array of 50 biotinylated DNA probes, 20 nucleotides-long each (IDT), antisense to ORF1a and ORF1b of the viral genomic RNA (Supplemental Data 1), and was maintained at 37 °C for 12 hours rotating in 500 mM NaCl, 0.7% SDS, 33 mM Tris-Cl pH 7, 0.7 mM EDTA, 10% Formamide. Dynabeads MyOne Streptavidin C1 (Invitrogen) were added during the final incubation hour. Beads containing the gRNA were captured on a magnet, while gRNA depleted supernatants were used for isolating the viral sgmRNA using a second tiling array of 50 biotinylated probes antisense to the sgmRNA ORFs as described in the text. Beads were washed 5 times with 2x SSC buffer supplemented with 0.5% SDS. RNA was released from the beads using 0.1 units/µl Turbo DNase (Invitrogen) at 37 °C for 30 minutes and was cleaned using RNA Clean & Concentrator (Zymo Research).

Cross-linked RNA enrichment. Viral enriched gRNA and sgmRNA fractions were fragmented by incubating at 37 °C for 20 minutes with 0.1 units/µl RNase III (Ambion).

Reactions were terminated by cleaning with RNA Clean & Concentrator (Zymo Research). Biotin was attached to cross-linked RNA duplexes by incubating with 150 µM Click-IT Biotin sDIBO Alkyne (Life technologies) under constant agitation at 37 °C for 1.5 hours . Residual Biotin sDIBO Alkyne was removed by RNA Clean & Concentrator (Zymo Research). Biotinylated RNA duplexes were enriched using Dynabeads MyOne Streptavidin C1 (Invitrogen) at the following reaction conditions: 100 mM Tris-Cl pH 7.5, 10 mM EDTA, 1 M NaCl, 0.1% Tween-20, 0.5 unit/µl Superase-In (Invitrogen). Beads were washed 5 times on a magnet with 100 mM Tris-HCl pH 7.5, 10 mM EDTA, 3.5 M NaCl, 0.1% Tween-20. RNA was eluted by adding 95% Formamide, 10 mM EDTA solution preheated to 65oC and purified using RNA Clean & Concentrator (Zymo Research).

Proximity ligation and crosslink reversal. Each RNA sample was divided in two: one half was used for proximity ligation and then crosslink reversal (i.e. COMRADES sample), while in the other half, crosslink reversal was done before proximity ligation (i.e. control sample). Prior to proximity ligation, RNA was denatured by briefly heating to 95 °C. Proximity ligation was done under the following conditions: 1 unit/µl RNA ligase 1 (New England Biolabs), 1x RNA ligase buffer, 50 µM ATP, 1 unit/µl Superase-in (Invitrogen), final volume: 200 µl. Reactions were incubated for 16 hours at 16 °C and were terminated by cleaning with RNA Clean & Concentrator (Zymo Research). Crosslink reversal was done by irradiating the RNA on ice with 2.5 KJ/m2 254 nm UVC using a CL-1000 crosslinker (UVP).

Sequencing library preparation. Library preparation was done as described in (Ziv et al., 2018), using 6N unique molecular identifiers to eliminate PCR biases. Pre-adenylated adapters were used and all ligation reactions were carried without ATP to reduce ligation artefacts. All libraries and controls went through 13 PCR cycles using KAPA HiFi HotStart Ready Mix (KAPA Biosystems). PCR products were size-selected on a 1.8% agarose gel before loading on a Novaseq (Illumina) for paired-end 150 bp runs. Total of ~1.6 billion sequences were achieved for this study.

Quantification and Statistical Analysis

Data Pre-processing. Data preprocessing was performed according to (Ziv et al., 2018). In brief, raw paired-end reads were trimmed for adaptors and checked for quality using cutadapt (Martin, 2011). Trimmed paired-end reads were assembled into single reads using the program pear (https://cme.h-its.org/exelixis/web/software/pear/doc.html). PCR duplicates were removed using unique molecular identifiers via collapse.py (https://gitlab.com/tdido/tstk). Chimeric reads were identified and annotated to the respective genome using hyb (Travis et al., 2014). SARS-CoV2 samples were processed using the Chlorocebus sabaeus reference genome (ChlSab1.1) with the addition of the SARS-CoV-2 sequence (NC_045512.2). MERS samples were processed using the Homo sapiens reference genoem (GRCh38) with the addition of MERS (NC_019843.3).

Clustering of chimeras into chimeric groups. Due to crosslinking and fragmentation, the COMRADES datas can provide redundant structural information whereby the same *in vivo* structure produces sequencing reads differing by a few nucleotides. This results in increased computation load of folding each chimeric read separately. To overcome this issue, and to gain better structure predictions, the reads were clustered into chimeric groups. Each chimeric read is composed of a left side (*L*) and right side (*R*), each originated from a different position along the gRNA or sgmRNA. Each chimeric read can therefore be described as (*g*): the genomic distance between *L* and *R*, and chimeric reads that originated from the same structure will have a similar *g* and can be clustered based on their *g* values. Clustering of chimeric reads that originated from the same structure was performed using a network based approach whereby an adjacency matrix is created for all chimeric reads based on the nucleotide difference between their *g* values (*Deltagap*).

$$Deltagap(gi, gj) = max(width(gi), width(gj)) - widthOfIntersection(gi, gj)$$

This results in *Deltagap=0 for* identically overlapping gaps, and increasing *Deltagap values for* chimeric reads that share less overlapping sites. The clustering was performed twice per sample, once for the chimeric reads that represent short stem structures (*g* <= 10 nt) and once for chimeric reads that represent long distance interactions (g > 10 nt). Short range interactions weights were calculated as:

$$E(gi, gj) = 10 - Deltagap(gi, gj)$$

This allows exactly overlapping gaps to have the highest weight, and gaps with no overlaps to have a weight of 0. For long range interactions weights were calculated as:

$$E(gi, gj) = 15 - Deltagap(gi, gj)$$

Long range interactions with weights lower than 0 were set to 0, meaning that gaps that differ by more than 15 nucleotides could not be considered as part of the same chimeric group. The weighted graphs created for long and short range interactions consisted of *g* as vertices and weights as edges: G = (V,E) using iGraph (http://www.interjournal.org/manuscript_abstract.php?361100992). To identify densely connected subgraphs (communities) with chimeric groups containing chimeric reads that originated from the same structure, we clustered the graph using random walks with the cluster_waltrap function (steps = 2) from the iGraph package. Chimeric groups containing less than 10 chimeric reads were discarded. Chimeric groups often contained a small amount of longer *L* or *R* sequences due to the random fragmentation in the COMRADES protocol. To avoid introducing biases in the folding results, clusters were trimmed to the region from *L* to *R* for which evidence in the cluster is higher than the mean evidence - 2 standard deviations. The trimmed clusters can be found in Supplemental Data 2.

Folding. Folding of the chimeric groups was performed using the vienna package (Lorenz et al., 2011). For short range chimeric groups RNAFold was used (with default parameters) and for long range chimeric groups RNADuplex was used (with default parameters).

Heatmaps, viewpoints and downstream analysis. Heatmaps, viewpoints and all other downstream analysis was performed using custom R scripts, all scripts to reproduce the analysis as well as functions to analyse other COMRADES data sets are available on GitHub (https://github.com/JLP-BioInf/SARS-Cov2-COMRADES)

Covariation analysis. A step-by-step guide R script with code used for the following analysis can be found at: https://github.com/JLP-BioInf/SARS-Cov2-COMRADES/tree/master/covariation. Complete Sarbecovirus sequences were taken from the NCBI Virus Database website (taxid: 2509511, 17 June 2020 version). Sequences containing unidentified nucleotides were discarded. Four sequence sets for MSA (multiple sequence analysis) were generated:

(I) MSA-1-SARSrel-3515seq: includes all complete non-redundant Sarbecovirus sequences. MUSCLE ((Edgar, 2004)), (default parameters).

(II) MSA-2-SARSrel-3515seq: Includes all complete non-redundant Sarbecovirus sequences containing only the four canonical nucleotide identifiers. MUSCLE (parameters, -maxiters 2).

(III) MSA-3-SARSrel-137seq: pairwise distances between the sequences in *(I)* were calculated using the mbed function in the kmer package (Blackshields et al., 2010). The “seeds” attribute was used to extract the sequence indices of all seed sequences. These seed sequences were included in this multiple sequence alignment. This resulted in a smaller sequence set but with a representative variation as (I). MUSCLE (default parameters).

(IV) MSA-4-SARSrel-559seq: The unaligned sequences used to generate *(II)* were divided into seven smaller sequence sets (six 500-sequences sets and one 515-sequence set). The seed sequences ("seeds" attribute of the mbed function in the kmer package (https://cran.r-project.org/package=kmer)) for all seven sets were combined in a new sequence set to be aligned to make up multiple sequence alignment *(IV)*. This resulted in a sequence set with less sequences than (I), but more than (III) and representative variation as (I). MUSCLE (default parameters).

In all cases the NCBI reference genome for SARS-CoV-2 (NC_045512.2) was used as reference. For each of the four sequence alignments, the following steps were taken:
(I) SARS-CoV-2 COMRADES cluster coordinates were taken from Supplemental Data 2. For long-range clusters (defined above) the segments defined by the coordinates of the left and right side of the respective cluster were extracted from the MSA, and fused together to form a smaller MSA containing only the aligned left and right side sequences. For short-range clusters (defined as above) the whole region defined by the start position of the left side and the end position of the right side was extracted. Those segments of the full MSA will be referred to as "cluster alignments". In both cases, any sequence starting with more than 10 empty positions was removed from that cluster alignment.
(II) The cluster alignments were analyzed with R-Scape (Rivas et al., 2017) using default parameters.
(iii) The R-Scape output for each candidate co-varying pair includes an E-value statistical score (probability of a false positive result for the respective position pair in the cluster alignment). The default significance level of 0.05 was kept, so only position pairs with E-values smaller than 0.05 were considered in the subsequent analysis.
(iv) Results tables of R-scape output combined with the corresponding coordinates in the non-aligned SARS-CoV-2 reference sequence, as well as nucleotide combination frequencies at the two positions across the alignment. We defined secondary base-pairing frequency as the percentage of sequences in which the pair of nucleotides differed from the most common base-pair type at the position but could still form a base-pair. For example, an imaginary pair of nucleotides has the composition:

A--T (94%); G--C(3%), G--T(2%); C--T(1%),

And its secondary base-pairing frequency is 83.3%. For further steps of the analysis (folding predictions), we consider only candidate pairs with ≥90% secondary base-pairing frequency. A list of the candidate base-pairs is found in Supplemental Data 3.

Covariation analysis of sgmRNA chimera clusters. For the covariation analysis of sgmRNA chimera clusters, cluster coordinates were taken from Supplemental Data 2. The alignments described above were shortened to include the leader sequence fused to the full-length of mRNA-S, and were subsequently used here. A modified version of the code used from full genome chimera clusters was used (available at https://github.com/JLP-BioInf/SARS-Cov2-COMRADES/tree/master/covariation). The identified base pairs were included in Supplemental Data 3.

Sequence conservation analysis of the extended FSE structure. Genome conservation data analyzed with synplot2 (Firth, 2014) was taken from (Firth, 2020). These data were aligned with our structural data and displayed in Figure 6C.

Additional Resources

Computer code has been deposited on GitHub: https://github.com/JLP-BioInf/SARS-Cov2-COMRADES. Base-pairing prediction, structure prediction and clustering data are available for exploration as a web interface: https://github.com/JLP-BioInf/SARS-CoV2-COMRADES-APP. Additional data supporting the findings of this study are available from the corresponding authors upon request.

Supplemental Information

Three additional data files are available with this manuscript (http://dx.doi.org/10.17632/ghh6w67vmx.1).

Supplemental Data 1, related to STAR★Methods. Antisense probes for viral RNA pulldowns.
Supplemental Data 2, related to STAR★Methods. Clustering of chimeric reads.
Supplemental Data 3, related to STAR★Methods. Covarying bases in SARS-CoV-2.

REFERENCES

Andrews, R.J., Peterson, J.M., Haniff, H.S., Chen, J., Williams, C., Grefe, M., Disney, M.D., and Moss, W.N. (2020). An in silico map of the SARS-CoV-2 RNA Structurome. bioRxiv, https://doi.org/10.1101/2020.04.17.045161.

Aw, J.G.A., Shen, Y., Wilm, A., Sun, M., Lim, X.N., Boon, K.-L., Tapsin, S., Chan, Y.-S., Tan, C.-P., Sim, A.Y.L., et al. (2016). In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. Mol. Cell *62*, 603–617.

Barry, J.K., and Miller, W.A. (2002). A -1 ribosomal frameshift element that requires base pairing across four kilobases suggests a mechanism of regulating ribosome and replicase traffic on a viral RNA. Proc. Natl. Acad. Sci. U. S. A. *99*, 11133–11138.

Blackshields, G., Sievers, F., Shi, W., Wilm, A., and Higgins, D.G. (2010). Sequence embedding for fast construction of guide trees for multiple sequence alignment. Algorithms Mol. Biol. *5*, 1–11.

van Boheemen, S., de Graaf, M., Lauber, C., Bestebroer, T.M., Raj, V.S., Zaki, A.M., Osterhaus, A.D.M.E., Haagmans, B.L., Gorbalenya, A.E., Snijder, E.J., et al. (2012). Genomic Characterization of a Newly Discovered Coronavirus Associated with Acute Respiratory Distress Syndrome in Humans. MBio *3*, 806.

Brierley, I., Digard, P., and Inglis, S.C. (1989). Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot. Cell *57*, 537–547.

Chen, S.-C., and Olsthoorn, R.C.L. (2010). Group-specific structural features of the 5'-proximal sequences of coronavirus genomic RNAs. Virology *401*, 29–41.

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

Cimino, P.A., Nicholson, B.L., Wu, B., Xu, W., and White, K.A. (2011). Multifaceted regulation of translational readthrough by RNA replication elements in a tombusvirus. PLoS Pathog. *7*, e1002423.

Dávila López, M., Rosenblad, M.A., and Samuelsson, T. (2009). Conserved and variable domains of RNase MRP RNA. RNA Biol. *6*, 208–220.

Dolja, V.V., and Koonin, E.V. (2018). Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res. *244*, 36–52.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Firth, A.E. (2014). Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. Nucleic Acids Res. *42*, 12425.

Firth, A.E. (2020). A putative new SARS-CoV protein, 3c, encoded in an ORF overlapping ORF3a. Journal of General Virology, https://doi.org/10.1099/jgv.0.001469.

Gao, F., and Simon, A.E. (2016). Multiple Cis-acting elements modulate programmed -1 ribosomal frameshifting in Pea enation mosaic virus. Nucleic Acids Res. *44*, 878–895.

Goebel, S.J., Hsue, B., Dombrowski, T.F., and Masters, P.S. (2004). Characterization of the RNA Components of a Putative Molecular Switch in the 3′ Untranslated Region of the Murine Coronavirus Genome. Journal of Virology *78*, 669–682.

Goldfarb, K.C., and Cech, T.R. (2017). Targeted CRISPR disruption reveals a role for RNase MRP RNA in human preribosomal RNA processing. Genes Dev. *31*, 59–71.

Hahn, C.S., Hahn, Y.S., Rice, C.M., Lee, E., Dalgarno, L., Strauss, E.G., and Strauss, J.H. (1987). Conserved elements in the 3' untranslated region of flavivirus RNAs and potential cyclization sequences. J. Mol. Biol. *198*, 33–41.

Hsue, B., and Masters, P.S. (1997). A bulged stem-loop structure in the 3' untranslated region of the genome of the coronavirus mouse hepatitis virus is essential for replication. Journal of Virology *71*, 7567–7578.

Huber, R.G., Lim, X.N., Ng, W.C., Sim, A.Y.L., Poh, H.X., Shen, Y., Lim, S.Y., Sundstrom, K.B., Sun, X., Aw, J.G., et al. (2019). Structure mapping of dengue and Zika viruses reveals functional long-range interactions. Nat. Commun. *10*, 1408.

Huston, N.C., Wan, H., de Cesaris Araujo Tavares, R., Wilen, C., and Pyle, A.M. (2020). Comprehensive in-vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. bioRxiv, https://doi.org/10.1101/2020.07.10.197079.

Jaag, H.M., Lu, Q., Schmitt, M.E., and Nagy, P.D. (2011). Role of RNase MRP in viral RNA degradation and RNA recombination. J. Virol. *85*, 243–253.

Jopling, C.L., Yi, M., Lancaster, A.M., Lemon, S.M., and Sarnow, P. (2005). Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. Science *309*, 1577–1581.

Kelly, J.A., Olson, A.N., Neupane, K., Munshi, S., Emeterio, J.S., Pollack, L., Woodside, M.T., and Dinman, J.D. (2020). Structural and functional conservation of the programmed −1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). Journal of Biological Chemistry *295*, 10741-10748.

Kudla, G., Wan, Y., and Helwak, A. (2020). RNA Conformation Capture by Proximity Ligation. Annu. Rev. Genomics Hum. Genet. *21*, 81-100.

Lan, T.C.T., Allan, M.F., Malsick, L.E., Khandwala, S., Nyeo, S.S.Y., Bathe, M., Griffiths, A., and Rouskin, S. (2020). Structure of the full SARS-CoV-2 RNA genome in infected cells. bioRxiv, https://doi.org/10.1101/2020.06.29.178343.

Li, L., Kang, H., Liu, P., Makkinje, N., Williamson, S.T., Leibowitz, J.L., and Giedroc, D.P. (2008). Structural Lability in Stem–Loop 1 Drives a 5′ UTR–3′ UTR Interaction in Coronavirus Replication. Journal of Molecular Biology *377*, 790–803.

Li, P., Wei, Y., Mei, M., Tang, L., Sun, L., Huang, W., Zhou, J., Zou, C., Zhang, S., Qin, C.-F., et al. (2018). Integrative Analysis of Zika Virus Genome RNA Structure Reveals Critical Determinants of Viral Infectivity. Cell Host Microbe *24*, 875–886.e5.

Liu, P., Yang, D., Carter, K., Masud, F., and Leibowitz, J.L. (2013). Functional analysis of the stem loop S3 and S4 structures in the coronavirus 3′UTR. Virology *443*, 40–47.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., et al. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet *395*, 565–574.

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. (2016). Coronavirus cis-Acting RNA Elements. Adv. Virus Res. *96*, 127–163.

Mak, J., and Kleiman, L. (1997). Primer tRNAs for reverse transcription. J. Virol. *71*, 8087–8095.

Mandal, M., and Breaker, R.R. (2004). Gene regulation by riboswitches. Nat. Rev. Mol. Cell Biol. *5*, 451–463.

Manfredonia, I., Nithin, C., Ponce-Salvatierra, A., Ghosh, P., Wirecki, T.K., Marinus, T., Ogando, N.S., Snider, E.J., van Hemert, M.J., Bujnicki, J.M., et al. (2020). Genome-wide mapping of therapeutically-relevant SARS-CoV-2 RNA structures. bioRxiv, https://doi.org/10.1101/2020.06.15.151647.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

Mateos-Gómez, P.A., Zuñiga, S., Palacio, L., Enjuanes, L., and Sola, I. (2011). Gene N proximal and distal RNA motifs regulate coronavirus nucleocapsid mRNA transcription. J. Virol. *85*, 8968–8980.

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.

Matera, A.G., Gregory Matera, A., and Wang, Z. (2014). A day in the life of the spliceosome. Nature Reviews Molecular Cell Biology *15*, 108–121.

Matsuyama, S., Nao, N., Shirato, K., Kawase, M., Saito, S., Takayama, I., Nagata, N., Sekizuka, T., Katoh, H., Kato, F., et al. (2020). Enhanced isolation of SARS-CoV-2 by TMPRSS2-expressing cells. Proc. Natl. Acad. Sci. U. S. A. *117*, 7001–7003.

McKibbin, W.J., and Fernando, R. (2020). The Global Macroeconomic Impacts of COVID-19: Seven Scenarios. CAMA Working Paper No. 19/2020, http://dx.doi.org/10.2139/ssrn.3547729.

Moreno, J.L., Zúñiga, S., Enjuanes, L., and Sola, I. (2008). Identification of a Coronavirus Transcription Enhancer. Journal of Virology *82*, 3882–3893.

Namy, O., Moran, S.J., Stuart, D.I., Gilbert, R.J.C., and Brierley, I. (2006). A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. Nature *441*, 244–247.

Plant, E.P., Rakauskaite, R., Taylor, D.R., and Dinman, J.D. (2010). Achieving a golden mean: mechanisms by which coronaviruses ensure synthesis of the correct stoichiometric ratios of viral proteins. J. Virol. *84*, 4330–4340.

Plant, E.P., Sims, A.C., Baric, R.S., Dinman, J.D., and Taylor, D.R. (2013). Altering SARS coronavirus frameshift efficiency affects genomic and subgenomic RNA production. Viruses *5*, 279–294.

Pyle, A.M. (1993). Ribozymes: a distinct class of metalloenzymes. Science *261*, 709–714.

Ridanpää, M., van Eenennaam, H., Pelin, K., Chadwick, R., Johnson, C., Yuan, B., vanVenrooij, W., Pruijn, G., Salmela, R., Rockas, S., et al. (2001). Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia. Cell *104*, 195–203.

Rivas, E., Clements, J., and Eddy, S.R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nat. Methods *14*, 45–48.

Rothe, C., Schunk, M., Sothmann, P., Bretzel, G., Froeschl, G., Wallrauch, C., Zimmer, T., Thiel, V., Janke, C., Guggemos, W., et al. (2020). Transmission of 2019-nCoV Infection from an Asymptomatic Contact in Germany. N. Engl. J. Med. *382*, 970–971.

Rowley, M.J., and Corces, V.G. (2018). Organizational principles of 3D genome architecture. Nat. Rev. Genet. *19*, 789–800.

Ryder, S.P. (2020). Analysis of Rapidly Emerging Variants in Structured Regions of the SARS-CoV-2 Genome. bioRxiv, https://doi.org/10.1101/2020.05.27.120105.

Sanders, W., Fritch, E.J., Madden, E.A., Graham, R.L., Vincent, H.A., Heise, M.T., Baric, R.S., and Moorman, N.J. (2020). Comparative analysis of coronavirus genomic RNA structure reveals conservation in SARS-like coronaviruses. bioRxiv, https://doi.org/10.1101/2020.06.15.153197.

Sawicki, S.G., Sawicki, D.L., and Siddell, S.G. (2007). A contemporary view of coronavirus transcription. J. Virol. *81*, 20–29.

Sharma, E., Sterne-Weiler, T., O'Hanlon, D., and Blencowe, B.J. (2016). Global Mapping of Human RNA-RNA Interactions. Mol. Cell *62*, 618–626.

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

Stammler, S.N., Cao, S., Chen, S.-J., and Giedroc, D.P. (2011). A conserved RNA pseudoknot in a putative molecular switch domain of the 3'-untranslated region of coronaviruses is only marginally stable. RNA *17*, 1747–1759.

Sun, L., Li, P., Ju, X., Rao, J., Huang, W., Zhang, S., Xiong, T., Xu, K., Zhou, X., Ren, L., et al. (2020). In vivo structural characterization of the whole SARS-CoV-2 RNA genome identifies host cell target proteins vulnerable to re-purposed drugs. bioRxiv, https://doi.org/10.1101/2020.07.07.192732.

Tajima, Y., Iwakawa, H.-O., Kaido, M., Mise, K., and Okuno, T. (2011). A long-distance RNA-RNA interaction plays an important role in programmed -1 ribosomal frameshifting in the translation of p88 replicase protein of Red clover necrotic mosaic virus. Virology *417*, 169–178.

Tomezsko, P.J., Corbin, V.D.A., Gupta, P., Swaminathan, H., Glasgow, M., Persad, S., Edwards, M.D., Mcintosh, L., Papenfuss, A.T., Emery, A., et al. (2020). Determination of RNA structural diversity and its role in HIV-1 RNA splicing. Nature *582*, 438–442.

Travis, A.J., Moody, J., Helwak, A., Tollervey, D., and Kudla, G. (2014). Hyb: A bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data. Methods *65*, 263.

Welting, T.J.M., Kikkert, B.J., van Venrooij, W.J., and Pruijn, G.J.M. (2006). Differential association of protein subunits with the human RNase MRP and RNase P complexes. RNA *12*, 1373–1382.

Williams, G.D., Chang, R.-Y., and Brian, D.A. (1999). A Phylogenetically Conserved Hairpin-Type 3′ Untranslated Region Pseudoknot Functions in Coronavirus RNA Replication. Journal of Virology *73*, 8349–8355.

Woolhouse, M., and Gaunt, E. (2007). Ecological origins of novel human pathogens. Crit. Rev. Microbiol. *33*, 231–242.

Ziv, O., Gabryelska, M.M., Lun, A.T.L., Gebert, L.F.R., Sheu-Gruttadauria, J., Meredith, L.W., Liu, Z.-Y., Kwok, C.K., Qin, C.-F., MacRae, I.J., et al. (2018). COMRADES determines in vivo RNA structures and interactions. Nat. Methods *15*, 785–788.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
SARS-CoV-2 strain München-1.2/2020/984	Laboratory of Prof. Dr. Christian Drosten (Charité – Universitätsmedizin, Berlin)	doi: 10.1056/NEJMc2001468
MERS-CoV strain EMC/2012	Laboratory of Prof. Dr. Christian Drosten (Charité – Universitätsmedizin, Berlin)	doi: 10.1128/mBio.00473-12
Chemicals, Peptides, and Recombinant Proteins		
Dulbecco's modified Eagles medium (DMEM)	Gibco™	A4192101
FBS	Gibco™	16000044
Psoralen-triethylene glycol azide (Psoralen-TEG azide)	Berry & Associates	PS 5030
Opti-MEM™ I Reduced Serum Medium, no phenol red	Gibco™	11058021
Proteinase K	New England Biolabs	P8107S
Dynabeads™ MyOne™ Streptavidin C1	Invitrogen™	65001
TURBO™ DNase	Invitrogen™	AM2238
Phosphate buffered saline	Sigma-Aldrich	P5493
Ambion™ RNase III	Invitrogen™	AM2290
Click-IT Biotin DIBO Alkyne	Invitrogen™	C20023
SUPERase In RNase Inhibitor	Invitrogen™	AM2696
Formamide - deionised	Sigma-Aldrich	F9037
Novex TBE-Urea Gels, 10%	Invitrogen™	EC6875BOX
SYBR Gold Nucleic Acid Gel Stain	Invitrogen™	S11494

T4 RNA Ligase 1, High concentration	New England Biolabs	M0437
KAPA HiFi HotStart ReadyMix	Roche Sequencing Store	KK2601
Critical Commercial Assays		
RNeasy Maxi Kit	Qiagen	75162
RNeasy Mini Kit	Qiagen	74104
RNA Clean & Concentrator-5	Zymo	R1015
Deposited Data		
Deposited sequencing datasets (all)	This manuscript	GEO under accession number: GSE154662
Integrated base-pairing prediction, structure prediction and clustering data	This manuscript	https://github.com/JLP-BioInf/SARS-CoV2-COMRADES-APP
Sarbecovirus Sequences	NCBI Virus Database	taxid: 2509511
Merbecovirus Sequences	NCBI Virus Database	taxid: 2509494
Experimental Models: Cell Lines		
Vero E6	ATCC	ATCC CRL-1586
VeroE6/TMPRSS2 cells (PMID: 32165541)	Laboratory of Prof. Dr. Stefan Pöhlmann (German Primate Center, Göttingen)	doi:10.1073/pnas.2002589117
HuH7 cells	JCRB	JCRB0403
Oligonucleotides		
Oligonucleotides for SARS-CoV-2 pulldown	See Supplemental data 1	N/A
Oligonucleotides for MERS-CoV pulldown	See Supplemental data 1	N/A
Software and Algorithms		
R 4.0.1	The R foundation	https://www.r-project.org/
R-scape	Rivas et al., 2017	http://eddylab.org/R-scape/
MUSCLE	Edgar, 2004	https://www.drive5.com/muscle/
kmer 1.1.2	Wilkinson SP (2018)	https://cran.r-project.org/web/packages/kmer/index.html
Bios2Cor 2.1	Taddese et al., 2020	https://cran.r-project.org/web/packages/Bios2cor/index.html
Seqinr	Charif and Lobry (2007)	https://cran.r-project.org/web/packages/seqinr/index.html

MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms	Kumar et al., 2018	https://www.megasoftware.net/
JalView	JavA Bioinformatics Analysis Web Services/// Waterhouse et al., 2009	https://www.jalview.org/
VARNA	Darty et al., 2009	http://varna.lri.fr/index.php?lang=en&page=home&css=varna
Adobe Illustrator (figure preparation)	Adobe	https://www.adobe.com/uk/
Affinity Designer (figure preparation)	2020 Serif (Europe) Ltd.	https://affinity.serif.com/en-gb/designer/

Figure 1

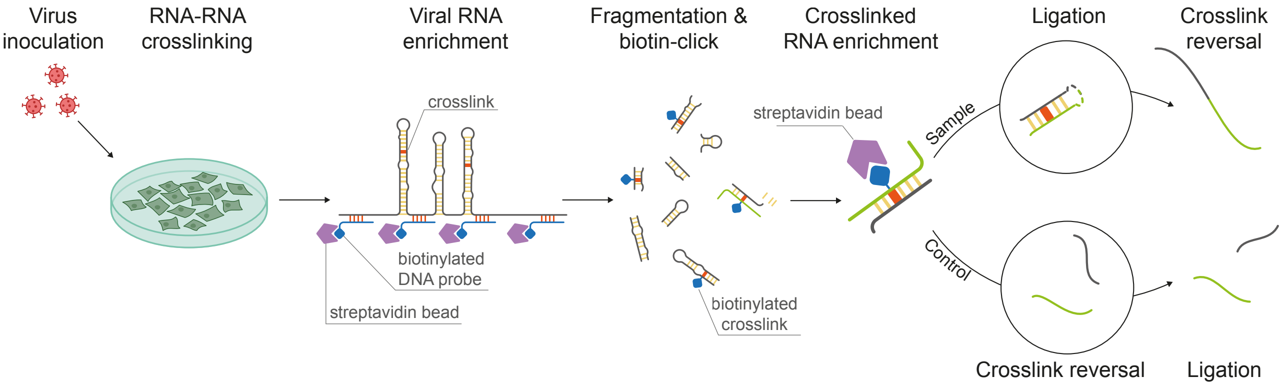


Figure 1
Ziv *et al*. 2020

Figure 2

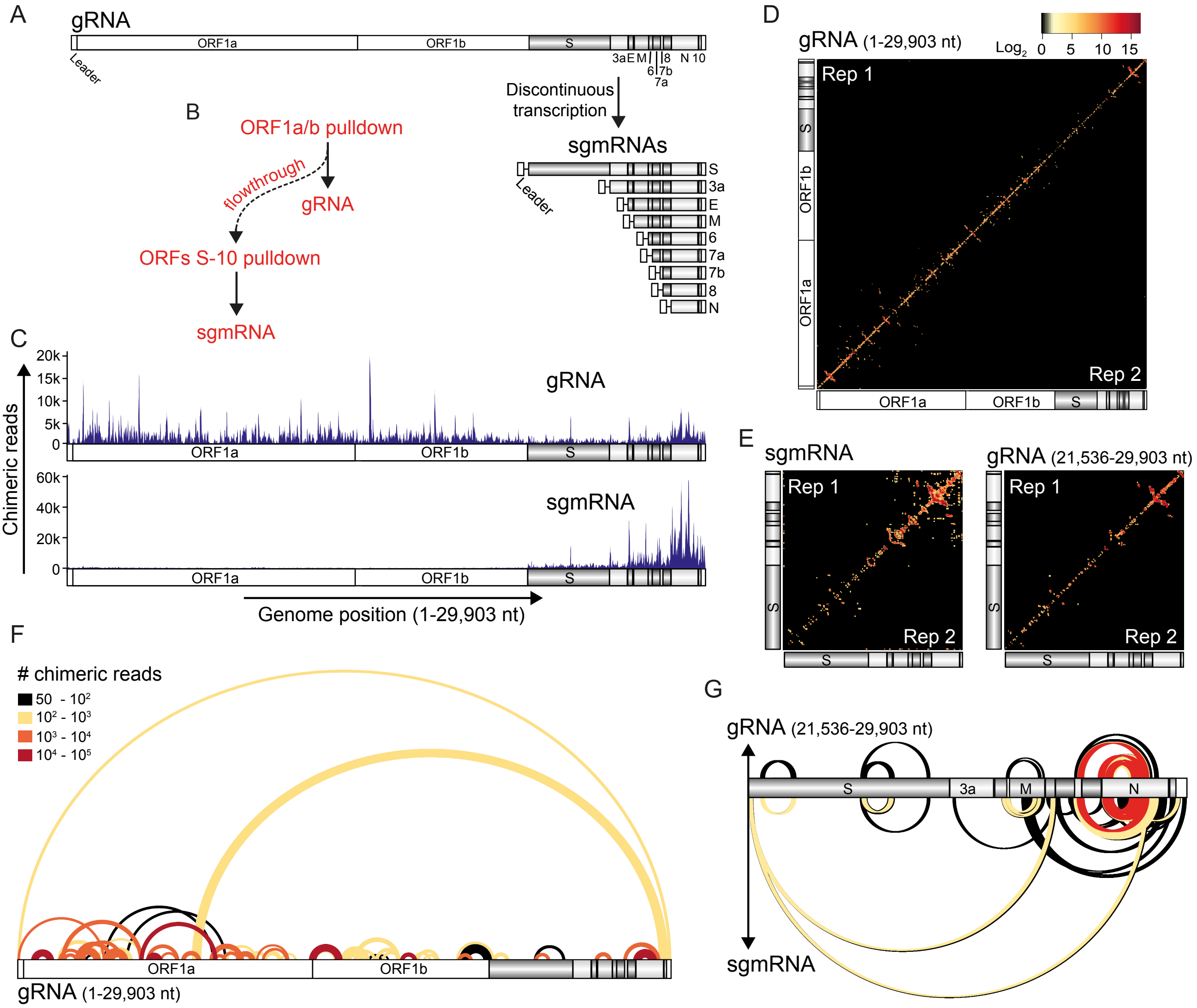


Figure 2
Ziv *et al*. 2020

Figure 3

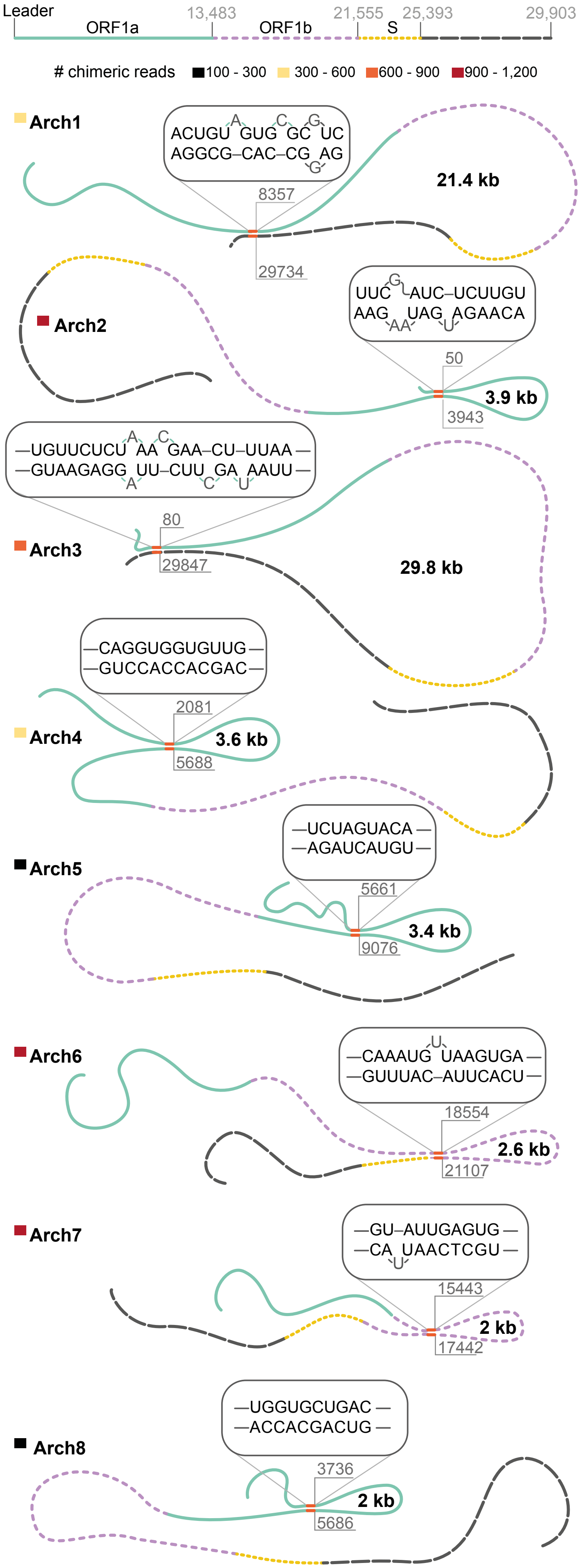


Figure 3
Ziv *et al.* 2020

Figure 4

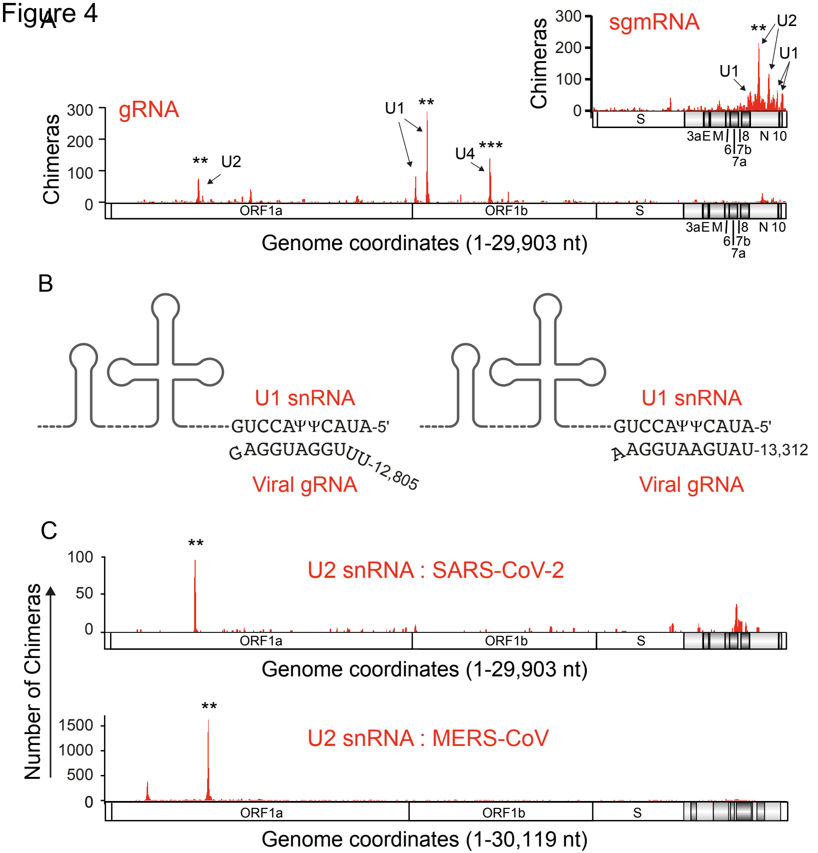


Figure 4

Ziv *et al.* 2020

Figure 5

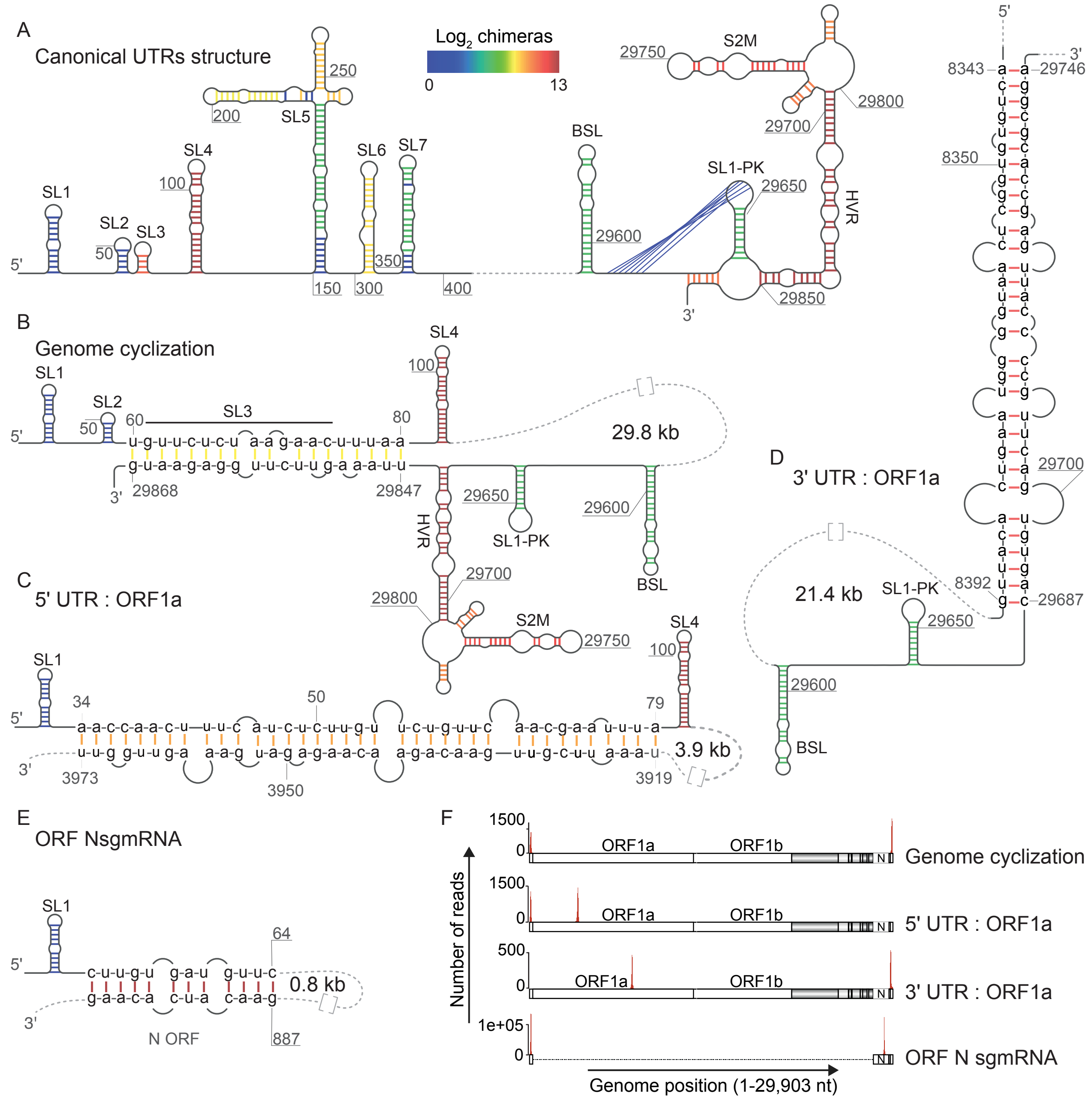


Figure 5
Ziv *et al.* 2020

Figure 6

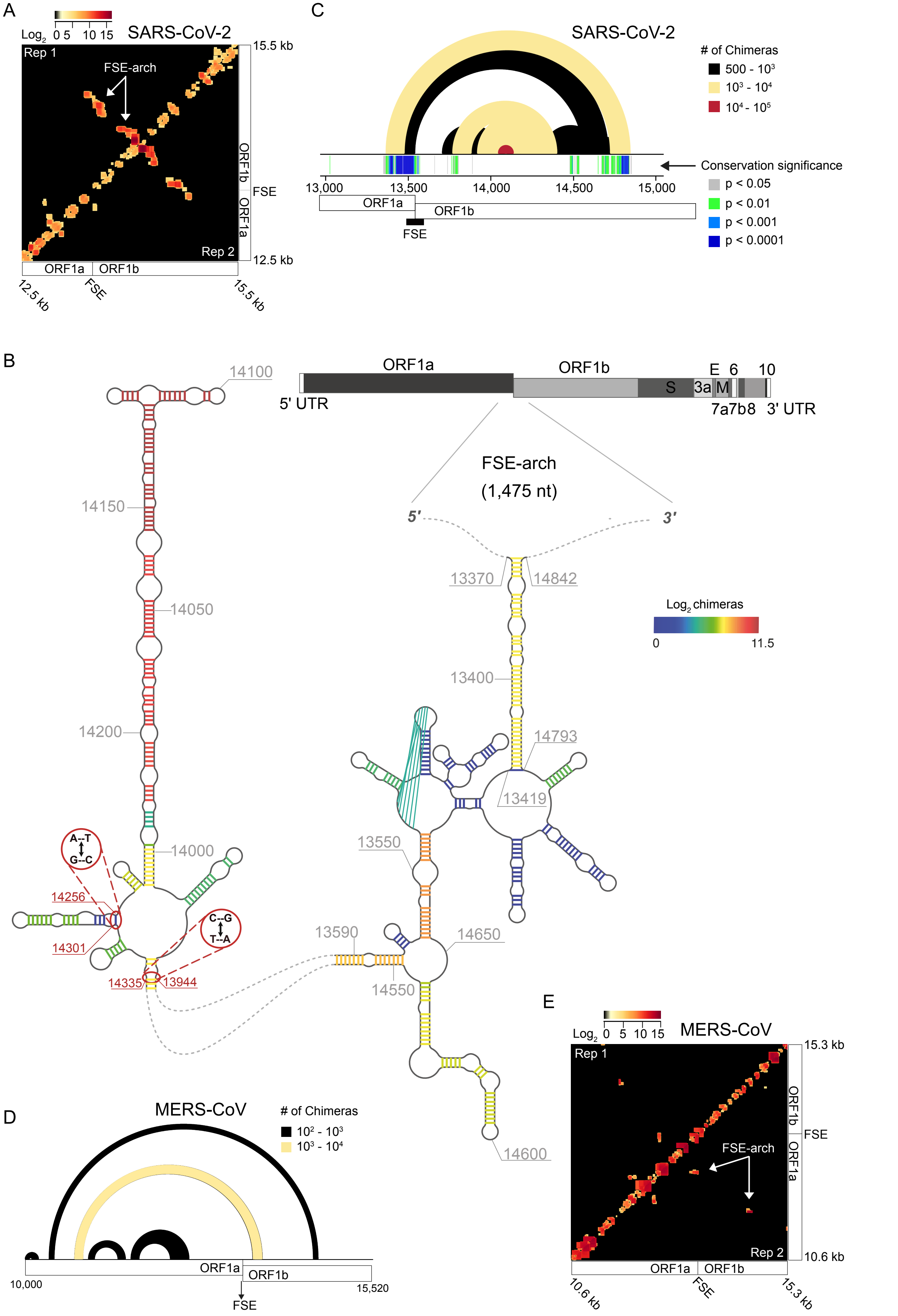
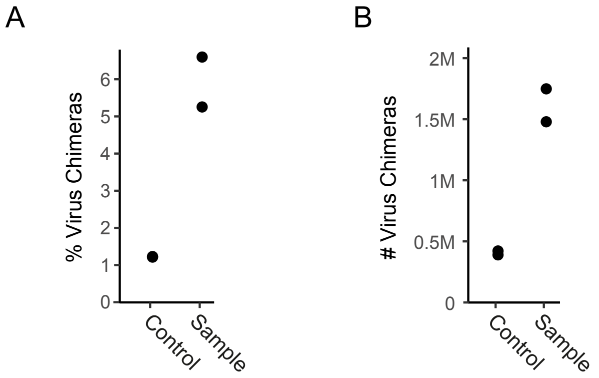
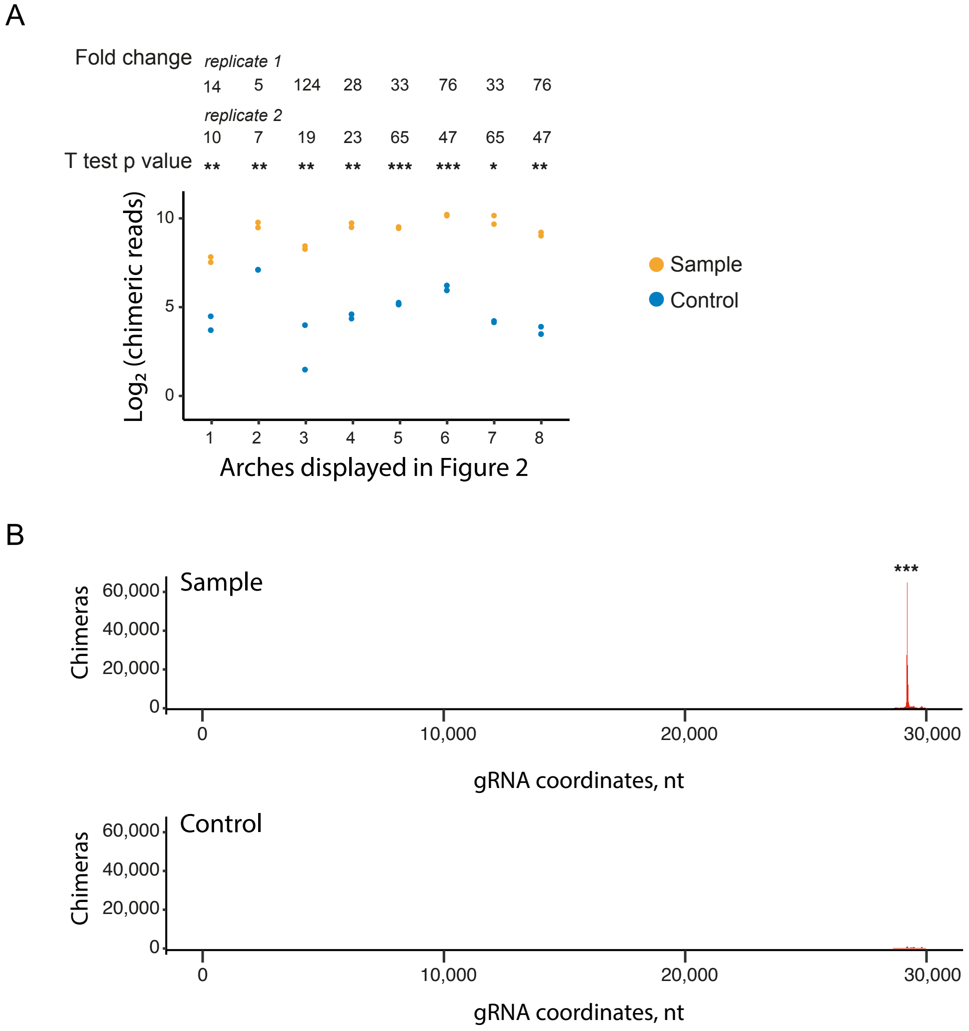


Figure 6
Ziv *et al.* 2020

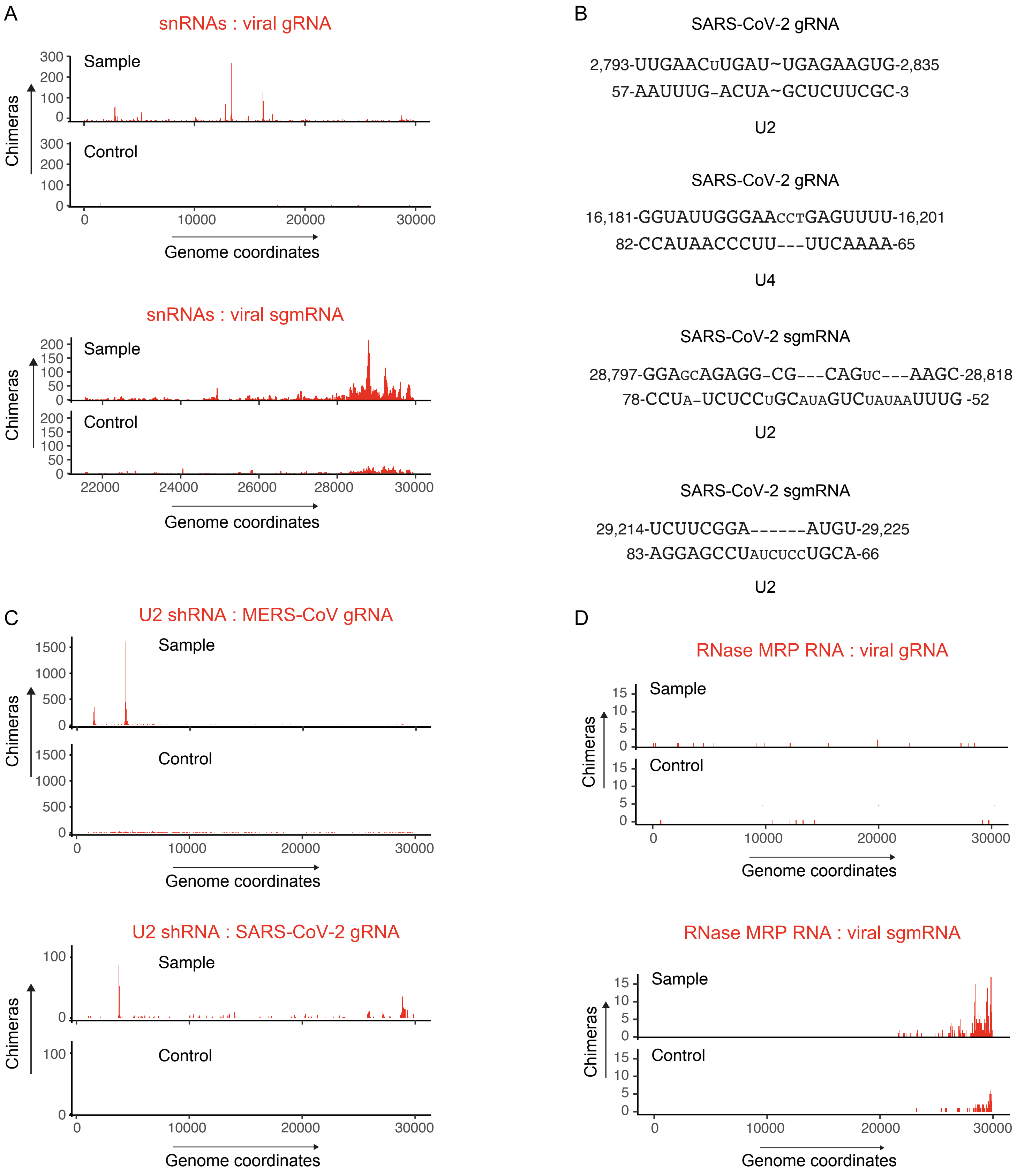
Supplemental Text and Figures



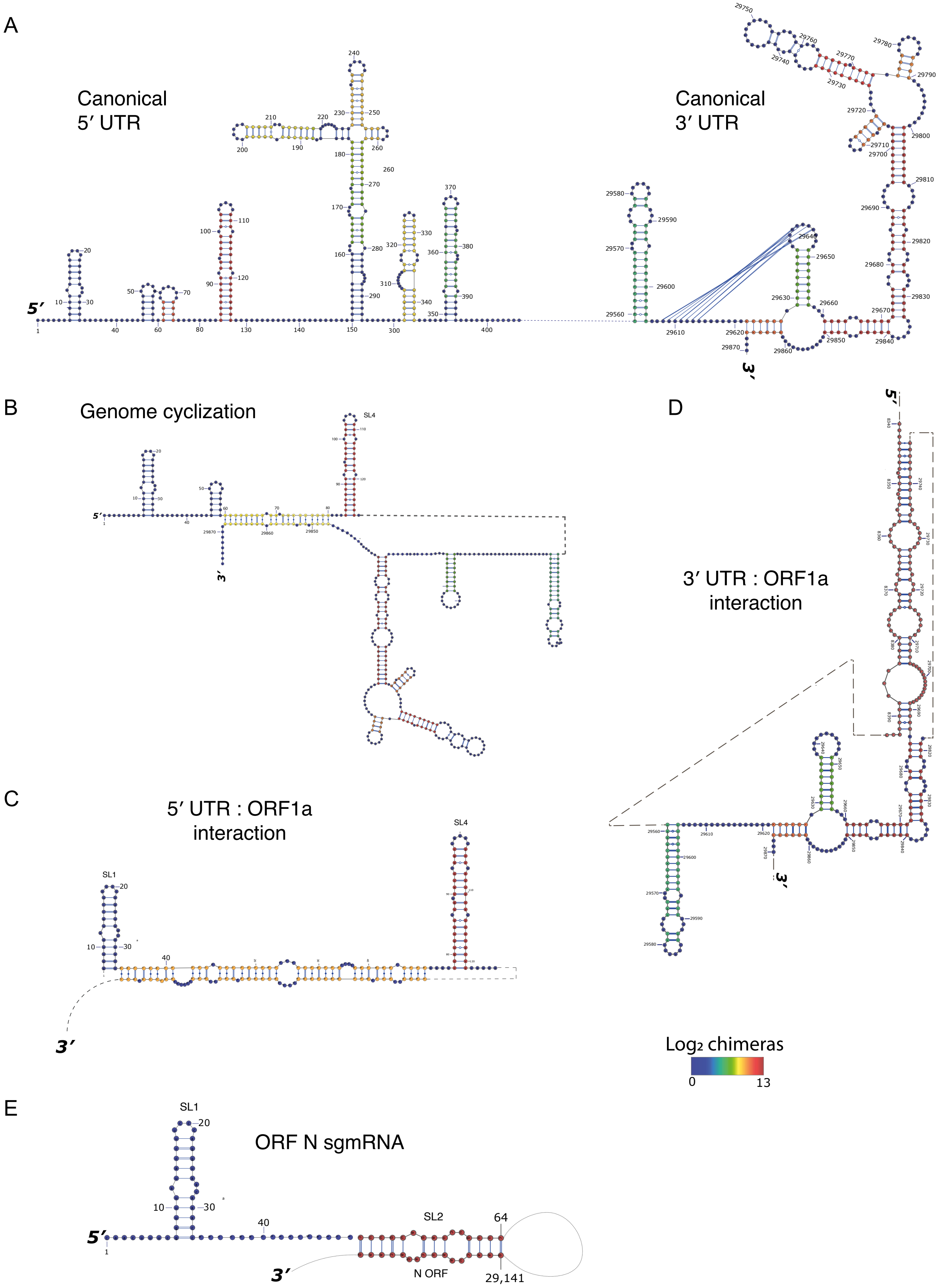
Supplementary Figure 1, related to Figure 1. The COMRADES method.
(A) Percentage of chimeric reads in the samples and control experiments.
(B) Number of chimeric reads in the samples and control experiments.



Supplementary Figure 2, related to Figure 3. Long-range RNA-RNA interactions along the SARS-CoV-2 gRNA and N ORF sgmRNA
(A) Number of chimeric reads supporting the arches shown in Figure 2 in COMRADES samples and controls libraries. Log_2 values are shown. Fold changes represent the ratios of chimeras in sample / chimeras in control for each biological replicate. T test p values indicate the significant of each arch (* < 0.1, ** < 0.05, *** < 0.01).
(B) Interactions of the leader sequence with a downstream position in ORF N sgmRNA. Number of chimeric reads supporting this base-pairing is shown. *** denotes T test p value < 0.01

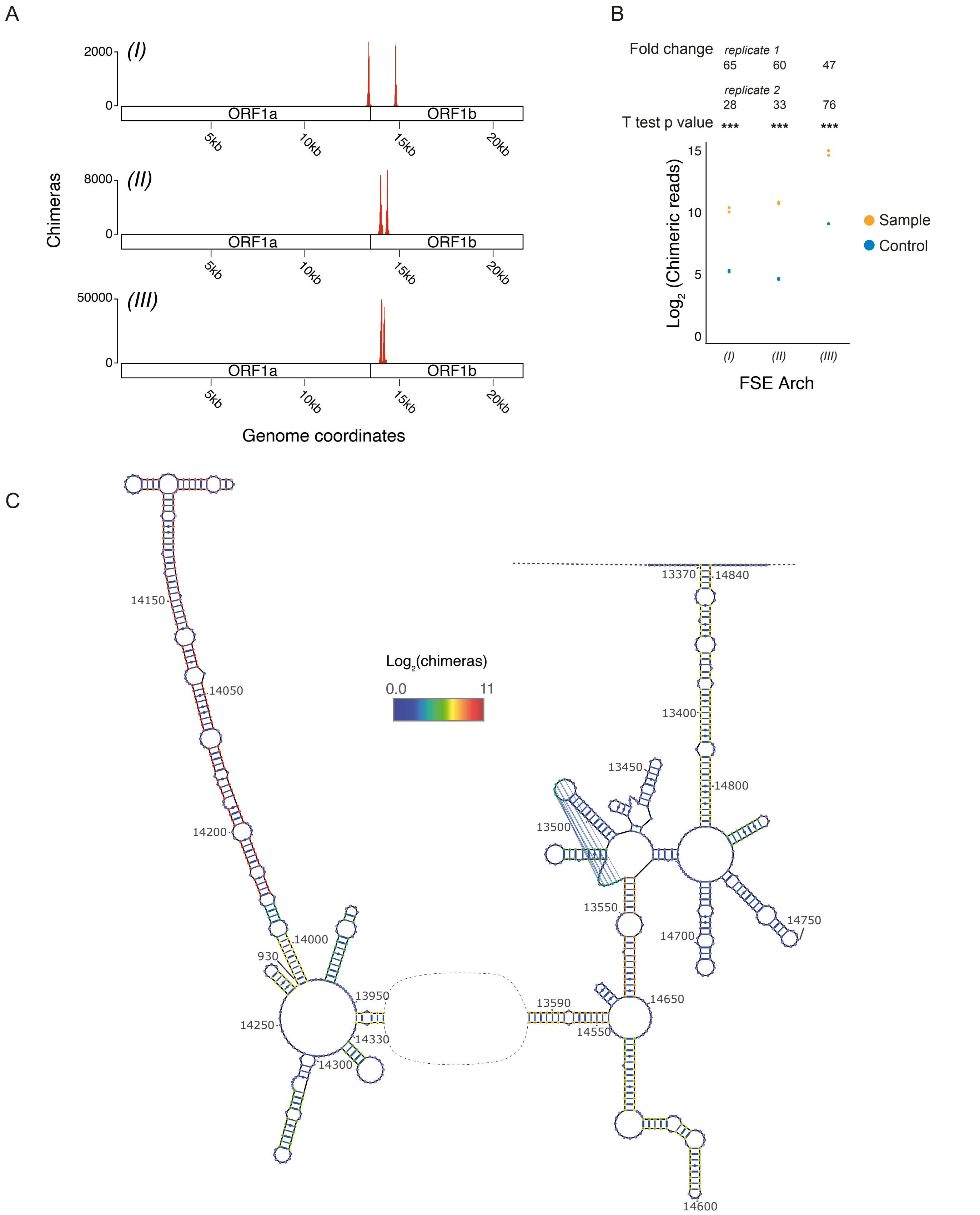


Supplementary Figure 3, related to Figure 4. Interactions between cellular and viral RNA
(A) snRNAs binding positions along the SARS-CoV-2 gRNA (top) and sgmRNA (bottom) and their COMRADES-controls.
(B) Base-pairing models for the interactions between viral RNA (top strand) and host snRNAs (bottom strand)
(C) U2 snRNA binding position along MERS-CoV gRNA (top) and SARS-CoV-2 gRNA (bottom) and their COMRADES controls
(D) RNase MRP binding positions along the SARS-CoV-2 gRNA (top) and sgmRNA (bottom) and their COMRADES controls.



Supplementary Figure 4, related to Figure 5. The UTRs of SARS-CoV-2 adopt alternative conformations inside cells

(A-E) Detailed representation of the canonical UTRs structure (A); genome cyclization (B); UTRs binding to ORF1a (C,D); and ORF N sgmRNA conformation (E). Colour code represents the number of non-redundant chimeric reads supporting each base-pair.



Supplementary Figure 5, related to Figure 6. The structure of the SARS-CoV-2 ribosomal frameshifting element arch (FSE-arch) inside cells

(A) Representation of the left- and right-side of the chimeric-reads supporting the FSE-arch. *(I)*-*(III)* correspond to the gRNA positions marked in (C).

(B) Number of chimeric reads supporting the FSE-arch in samples and controls. Log$_2$ values are shown. Fold changes represent the ratios of chimeras in sample / chimeras in control for each biological replicate. T test p values indicate the significant of each position (*** < 0.01). *(I)*-*(III)* correspond to the gRNA positions marked in (C).

(C) Detailed representation of the FSE-arch structure. Colour code represents the number of non-redundant chimeric reads supporting each base-pair.

Supplementary Table 1. Prevalence of chimeras supporting the long-range interactions shown in Figure 3.

Arch name	Long-range chimeras[a]	Short-range chimeras[b]	Ratio of Long-range / Short-range chimeras
1	442	14,506	3.0%
2	1,873	5,980	31.3%
3	704	5,980	11.8%
4	1,690	13,546	12.5%
5	1,529	9,050	16.9%
6	2,493	7,447	33.5%
7	2,102	16,531	12.7%
8	1,193	11,766	10.1%

[a] Long-range chimeras supporting each arch

[b] Chimeras supporting alternative, short-range interactions within the arch regions

Supplemental Data 1

Supplemental Data 2

Click here to access/download

Supplemental Videos and Spreadsheets

Supplemental Data 2.xlsx

Supplemental Data 3