The architecture of the SARS-CoV-2 RNA genome

Changchang Cao
Chinese Academy of Sciences
Zhaokui Cai
Chinese Academy of Sciences
Xia Xiao
Chinese Academy of Medical Sciences & Peking Union Medical College
Jian Rao
Chinese Academy of Medical Sciences & Peking Union Medical College
Naijing Hu
Chinese Academy of Sciences
Minnan Yang
Chinese Academy of Sciences
Xiaorui Xing
Chinese Academy of Sciences
Bing Zhou
Chinese Academy of Sciences
Xiangxi Wang
Chinese Academy of Sciences
Jianwei Wang
Chinese Academy of Medical Sciences and Peking Union Medical College
Yuanchao Xue (ycxue@ibp.ac.cn)
Chinese Academy of Sciences

Research Article

Keywords: SARS-CoV-2, vRIC-seq, RNA structure, D614G mutation, siRNA

Posted Date: December 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-132578/v1

Abstract

SARS-CoV-2 carries the largest single-stranded RNA genome and is the causal pathogen of the ongoing COVID-19 pandemic. How the SARS-CoV-2 RNA genome is folded in the virion remains unknown. To fill the knowledge gap and facilitate structure-based drug development, we developed a virion RNA in situ conformation sequencing technology, named vRIC-seq, for probing viral RNA genome structure unbiasedly. Using vRIC-seq data, we reconstructed the tertiary structure of the SARS-CoV-2 genome and revealed a surprisingly "unentangled globule" conformation. We uncovered many long-range duplexes and higher-order junctions, both of them were under purifying selections and contributed to the sequential package of the SARS-CoV-2 genome. Unexpectedly, the D614G and two accompanying mutations might remodel duplexes into more stable forms. Lastly, the structure-guided design of potent small interfering RNAs could obliterate the SARS-CoV-2 in Vero cells. Overall, our work provides a framework for studying the genome structure, function, and dynamics of emerging deadly RNA viruses.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causal pathogen of Coronavirus Disease 2019 (COVID-19) [1-3]. As a single-stranded and positive-sense RNA virus, SARS-CoV-2, together with SARS-CoV and Middle East respiratory syndrome coronavirus (MERS-CoV), all belong to the *Coronaviridae* family [4]. SARS-CoV-2 carries one of the largest RNA genomes (~30 kilobases, kb) among all RNA virus families and encodes about 29 proteins [1,3,5-7]. Since its outbreak in late December 2019, SARS-CoV-2 has infected tens of millions of peoples and caused over one million deaths worldwide (https://covid19.who.int/). Although global efforts and resources have been redirected to fighting against SARS-CoV-2, there are no effective antiviral medicines or approved vaccines available yet. Considering the RNA nature of SARS-CoV-2, RNA-based therapeutics such as small interference RNAs (siRNAs) or antisense oligos (ASOs) are emerging as potent agents to cleave the viral RNA genome in infected host cells. Because RNA structure can significantly influence the efficacy of siRNAs and ASOs [8,9], deciphering the 3D structure of SARS-CoV-2 becomes an urgent need prior to RNA-based drug development.

RNA structures are widely recognized as critical modulators in regulating transcription, translation, and replications of coronavirus and other RNA viruses [10-17]. At this frontier, many efforts have been devoted to study the structure of SARS-CoV-2. Even though Cryo-electron microscopy and 3D electron tomography are powerful in delineating the global architectures of SARS-CoV-2, the entire RNA genome inside virions remains unrevealed [18-20]. Besides these physical approaches, several chemical-based high-throughput sequencing methods, such as icSHAPE and DMS-MaPseq, have been recently applied to probe single-stranded regions of SARS-CoV-2 in infected cells or in vitro [21-24]. The mapped single-stranded information could be further used as restraints to predict the base-paired regions within 500 nt [25]. Yet the current secondary structural model of SARS-CoV-2 might be incomplete since it missed information of long-range duplexes, which are prevalent and vital for completing the life cycles of positive-strand RNA viruses [26,27]. As a significant advance, a psoralen-based method called COMRADES recently identified many

long-range RNA duplexes of SARS-CoV-2 inside cells, further highlighting the importance of RNA duplexes in maintaining SARS-CoV-2 fitness [28,29]. Significantly lagging behind those in-cell structural studies, how the 30 kb genome of SARS-CoV-2 is organized and arranged in virions remains unclear.

We recently developed an RNA *in situ* conformation sequencing technology, named RIC-seq, for unbiased mapping of RNA-RNA spatial interactions in living cells [30]. RIC-seq utilizes a pCp-biotin to label proximally interacting chimeric RNAs and high-throughput sequencing to retrieve their spatial proximity information. We demonstrated that RIC-seq could successfully detect short- and long-range duplexes, multiple-way junctions, and loop-loop contacts without base pairing potentials. These merits make RIC-seq more competent to decipher the 3D structure of the SARS-CoV-2 RNA genome. But SARS-CoV-2 virions are typically 80 nanometers in diameter and can't be pelleted down as human cells by standard centrifugation [1]. This feature makes virions not compliant with our current RIC-seq protocol that includes extensive washing and standard centrifugation at every enzymatic step [30]. To overcome the major challenges, we design a way to hold SARS-CoV-2 virions on Concanavalin A (ConA) beads that bind specifically to glycoproteins present on the surface of virions. By optimizing virion capture, crosslinking, and enzymatic conditions, we further developed a virion RNA in situ conformation sequencing technology, named vRIC-seq, for global mapping of viral RNA genome structures in intact virions.

In this study, we successfully applied the vRIC-seq technology to probe the SARS-CoV-2 RNA genome structure in intact virions. We reconstructed a model of the surprisingly compact yet unentangled tertiary structure of the SARS-CoV-2 RNA genome. At the secondary structure level, we discovered that the in-virion structure of SARS-CoV-2 keeps mostly unchanged in the cell except for many long-range RNA duplexes. We noted that many long-range RNA-RNA interactions are under purifying selection and might contribute to the 3D package of the SARS-CoV-2 RNA genome. Finally, we uncovered several highly accessible single-stranded regions in SARS-CoV-2 for efficient viral RNA cleavage in Vero cells by using siRNAs.

Results

Overview of vRIC-seq technology

Like other coronaviruses, the envelope of SARS-CoV-2 contains two major glycoproteins spike (S) and membrane (M) [31,32]. According to this feature, we used magnetic beads coated with ConA, a kind of plant lectin that can specifically bind glycoproteins, to capture the SARS-CoV-2 virions prepared from the supernatants of infected Vero cells (see Methods). After capturing the virions on beads, formaldehyde was further applied to crosslink the nucleocapsid (N) protein-mediated proximal RNA-RNA interactions, as well as ConA and the surface glycoproteins (Fig. 1a). Next, the virions were permeabilized and treated with micrococcal nuclease (MNase) to digest away free RNAs that were not close to each other. Subsequently, all the proximal RNAs were 3' end-labeled with pCp-biotin and ligated together by T4 RNA ligase. Lastly, the resulting chimeric RNA marked with pCp at the juncture were enriched and converted into libraries for paired-end sequencing (about 260 bp, Fig. 1a and Supplementary Fig. 1a).

We obtained approximately 24.6 million unique reads for each replicate and ~3.4 million chimeric reads resulting from different RNA fragments. Using the RIC-seq data analysis workflow established earlier [30], we found that 97.3% of the chimeric reads could be aligned to the SARS-CoV-2 genome, and 2.3% were mapped to host RNA (Fig. 1b). The trace amounts of host RNA reads might be derived from detached Vero cells at the virion collection stage. Notably, approximately 0.4% of the chimeric reads were virus-to-host RNA interactions (Fig. 1b), which may reflect the random ligation rates between virions and Vero cells, therefore, further highlighting the specificity of vRIC-seq technology. Besides, the virus-to-host interactions also raised the possibility that some host RNAs might be packaged into the virion, just like tRNA in the HIV virions for its replication [33]. However, such a possibility was excluded because we observed a random virus-to-host RNA interaction pattern across all the green monkey chromosomes (Supplementary Fig. 1b).

The pCp-biotin labelling and selection were successful because approximately 85% of the additional nucleotides at chimeric junctions were cytosine (Supplementary Fig. 1c). vRIC-seq is highly reproducible on chimeric reads coverage (R = 0.999) and interaction strength (R = 0.995) of pairwise sites (Fig. 1c and Supplementary Fig. 1d). Moreover, we noted that 99.7% of the SARS-CoV-2 genome was covered at least 2500× by chimeric reads (Supplementary Fig. 1e), and the remaining 0.3% covered by less than 2500× was mainly located at the stem-loop 1 of 5’ UTR and the poly (A) region at the 3’ terminal. Next, we aligned the pairwise interacting RNA fragments to the SARS-CoV-2 genome, and such analysis revealed many regions that are preferably cut by MNase and subsequently labelled with pCp-biotin (Fig. 1d). We also noticed that the first 3,500 nucleotides (nt) of ORF1b had a relative lower vRIC-seq coverage, and nucleotide content didn’t contribute to this difference (Fig. 1d).

Recapitulate known structures of the SARS-CoV-2 genome

We first determined to validate the viral RNA spatial interactions revealed by vRIC-seq. For this purpose, we deduced several conserved RNA duplexes in coronavirus using vRIC-seq data and compared it with recently proposed SARS-CoV-2 secondary structure models [13,21-24,34-36]. As expected, vRIC-seq faithfully recapitulated all of the stem-loops in the 5’ UTR (1-395 nt) of SARS-CoV-2 except stem-loop 1 (SL1), which is approximately 40 nt in length and can’t be purified by AMPure XP beads (Fig. 1e). Moreover, the 3’ UTR of SARS-CoV-2 contains several conserved structural elements known to functionally impact viral RNA synthesis and translation, including the stem-loop II-like motif (s2m), hyper-variable region (HVR), and mutually exclusive bulged stem-loop (BSL) or pseudoknot (PK) [36]. We found that vRIC-seq successfully captured the canonical s2m and HVR structures (Fig. 1f). However, in contrast to the previous theoretical model [36], our data preferentially supported a double hairpin conformation for BSL and P2 stem, rather than the pseudoknot conformation (Fig. 1f). These results demonstrate that vRIC-seq can faithfully probe RNA spatial interactions in the virion, and support that this proximity information can be used for structure modelling.

Topological organization of the SARS-CoV-2 genome

After validating the vRIC-seq data, we next investigated the features of SARS-CoV-2 genome organization in the virion. To this end, we first calculated the spanning distance of pairwise interacting RNA fragments. Approximately 90.6% of the pairwise interactions happened within 100 nt, whereas approximately 6.3% of the interactions spanned more than 600 nt, and three sharp peaks with 810, 1360, and 2090 nt separately, were observed (Fig. 2a). Of note, those long-distance interactions were not caused by the discontinuous transcription of SARS-CoV-2 during negative-strand synthesis [6], because we observed a clear enrichment of pCp at the chimeric junctions (see red lines, Supplementary Fig. 2a,b).

Following our previous approach, we divided the SARS-CoV-2 genome into 10-nt windows and constructed an RNA interaction map (Fig. 2b). Using the map, we identified 254 clustered interactions positioned perpendicular to the diagonal, suggesting the widespread occurrence of local duplexes in the SARS-CoV-2 genome (Fig. 2b). Surprisingly, we also noticed 77 long-range interactions that were sequentially distributed and covered almost the entire genome (Fig. 2b). Similar to the organization of human primary transcripts [30], we observed 49 RNA topological domains with a median length of 630 nt in the SARS-CoV-2 genome (Fig. 2b, Supplementary Fig. 2c, and Supplementary Table 1). The largest domain is located before the 3' UTR and contains sequences encoding ORF7a, ORF7b, ORF8, and the N protein.

Theoretically, vRIC-seq can capture base-paired RNA duplexes as well as protein-mediated indirect RNA contacts. To examine the base-pairing probabilities for the observed long-range interactions within 2.5 kb, we calculated the minimum free energy (MFE) for pairwise interacting fragments that spanned over 400 nt. We observed significantly lower MFE values than randomly shuffled sequences with the same nucleotide content (P = 1.62e-10, Supplementary Fig. 2d), indicating that those long-range RNA-RNA interactions may be directly base-paired. Notably, we also uncovered 62 pairwise interacting RNA fragments that could span over 2.5 kb (Supplementary Fig. 2e). However, those individual RNA fragments showed 12-fold stronger interactions with its local partners rather than distal partners (Supplementary Fig. 2f), as exampled for long-range interactions between 1,180-1,190 nt and 29,343-29,354 nt, which both showed stronger local interactions (Supplementary Fig. 2g-i). These data suggest that some alternative topology of the SARS-CoV-2 genome might be present in the virion.

3D globule configuration of the SARS-CoV-2 genome in virions

Previous microscopy studies revealed a spherical shape with a mean diameter of ~80 nm for CoV and SARS-CoV-2 viral particles [1,37,38]. To explore how the 30 kb RNA genome of SARS-CoV-2 is folded in the tiny virion, we utilized the contact frequencies data of different RNA fragments as constraints to model the global conformation of the SARS-CoV-2 genome. Pursuing this, we adopted a widely used miniMDS (multidimensional scaling) approach to infer the 3D structure of SARS-CoV-2 [39]. Our modeling revealed a 3D globule conformation of the RNA genome (Fig. 2c and Supplementary Video 1). Notably, the thread-like genome was unentangled, appearing to have a mildly helical conformation. Moreover, different segments of the viral genome seemed to occupy separate territories (see different colors, Fig. 2c), forming an organization apparently similar to known structures of mammalian genomes [40]. Of note, this

“globule” and knot-free configuration might reflect the arrangement of the N protein, which has been shown to bind the CoV genome and interacted with M protein via its C-terminal domain in the interior of the lipid membrane of virions [18].

Secondary structure model of the SARS-CoV-2 genome

Based on the 3D RNA interaction map, we further developed an adaptive strategy to model the secondary structure of the entire SARS-CoV-2 genome. Briefly, we first predicted local duplexes, which were positioned in our SARS-CoV-2 RNA interaction map as perpendicular signals to the diagonal, and then used these duplexes as constraints to predict long-range duplexes (Supplementary Fig. 3a). To evaluate the performance of this strategy, we first predicted the secondary structure of 28S rRNA using our previously published RIC-seq data in HeLa cells [30]. The structural model achieved a sensitivity of 83.0% and a positive predictive value (PPV) of 78.3%, and both criteria were significantly higher than structures merely based on the minimal free energy values provided by several computational tools (Supplementary Fig. 3b). Importantly, our algorithm showed higher PPV and sensitivity if RIC-seq detected duplexes spanning over 600 nt were counted (Supplementary Fig. 3c). Together, these data demonstrate the accuracy of our adaptive strategy in deducing RNA structures.

Having validated the prediction algorithm, we next applied it to reconstruct the secondary structure of the whole SARS-CoV-2 genome (Fig. 3). Our structural model is highly favoured by the vRIC-seq data, as base-paired regions showed stronger interaction strength than size-matched unpaired control sequences (P < 2.2e-16, Supplementary Fig. 3d). In our structural model, the median and mean distances between two paired regions in SARS-CoV-2 were 28 and 111 nt, respectively. The maximal spanning distance (2,145 nt) was observed for a duplex formed between two fragments: 27,357-27,396 nt and 29,465-29,502 nt (Green dash line boxed region, Fig. 3). We found that about 63.8% (19,064 nt) of the SARS-CoV-2 genome were base-paired, and our model precisely recapitulated the stem-loop structures (SL1-SL7) in the 5’ UTR, as well as 3’ UTR structures including the s2m and the HVR (Fig. 3). Notably, our vRIC-seq data strongly supported an extended duplex SL8, rather than the two previously proposed separate duplexes which were theoretically predicted using the coding sequence of nsp1 (410-470 nt) [36] (see blue dash line marked region, Fig. 3). Besides identifying many duplexes, our structural model unexpectedly revealed 167 multi-way junctions, including 57 three-way junctions and three 12-way junctions; these junctions seemed to organize the SARS-CoV-2 RNA genome into many petaloid structures (Fig. 3 and Supplementary Fig. 3e).

Next, we examined whether there are any correlations between structural elements and viral RNA abundance. During negative-strand RNA synthesis, SARS-CoV-2 produces nine subgenomic RNAs (sgRNAs) via the template-switching activity of RNA-dependent RNA polymerase (RdRP) between the 5’ leader sequence and the transcription-regulatory sequence in the body (TRS-B) [6]. Interestingly, the TRS-B of ORF7a was located in a long-range duplex spanning over 2,000 nt, while the other eight TRS-B were located in local stem-loops (cyan boxed regions, Fig. 3 and Supplementary Fig. 3f). Moreover, we found that the number of single-stranded nucleotides within TRS-B was negatively correlated to the abundance

of the corresponding sgRNA (R = - 0.52, Supplementary Fig. 3g). By contrast, the number of single-stranded nucleotides in regions adjacent to the TRS-B showed positive correlations with sgRNA levels (Supplementary Fig. 3h). It will be interesting in the future to examine how these structural features determine viral RNA transcriptions.

The in-cell and in-virion structural dynamics

SARS-CoV-2 can hijack host cells by engaging its spike proteins with the host receptor angiotensin-converting enzyme 2 (ACE2) [7]. After entering into the host cell, the SARS-CoV-2 genome is released from the densely coated N proteins for translating nonstructural proteins to replicate its genome, as well as to produce structural proteins to assemble new virions (Fig. 4a). Whether there is any genome structure remodeling before and/or after the infection is still unclear. By analyzing the in vivo SHAPE-MaP data [24], we found that the single-stranded regions revealed by vRIC-seq tend to have stronger SHAPE signals than base-paired regions (P < 2.2e-16, Fig. 4b). Unexpectedly, 76% of the duplexes were observed in both the host cell and the virion, indicating that the SARS-CoV-2 genome structure was largely unchanged before and after infection (Fig. 4c). Additionally, the duplexes spanning longer distances tended to be changed in cells (Fig. 4d).

Moreover, the duplexes in the 5’ UTR regions barely changed during the life cycle of SARS-CoV-2 (Fig. 4e). One of the most dynamic regions was 12.6 kb - 13.6 kb, which covered the frame-shifting element (FSE) and resided at the boundary of ORF1a and ORF1b. The in-cell structure reported a pseudoknot conformation, but the in-virion structure suggested that the Stem 1 of pseudoknot (PK) in FSE preferably formed an alternative duplex with the upstream fragment (Fig. 4f and Supplementary Fig. 4a-c). Consistent with our observation, another study indirectly predicted this alternative duplex conformation in infected cells by using DMS-MaPseq revealed single-stranded information [23]. Although both conformations might be dynamically present in virions, we indeed observed 2.9-fold more vRIC-seq signals to support the alternative duplex than the PK (79.95 vs. 27.91, Supplementary Fig. 4b,c). It will be interesting to explore the functionality of these structural changes in the future. We also examined another well-known PK in the 3’ UTR of betacoronavirus [13]. Our vRIC-seq data barely support the presence of the 3’ UTR PK (Supplementary Fig. 4d-f), and both the in-virion and in-cell structures predict an extended BSL rather than a pseudoknot (Fig. 3, Fig. 4g, and Supplementary Fig. 4d-f). Together, these results demonstrate the accuracy of our secondary structure model of the SARS-CoV-2 genome in virions and highlight the potentially impactful structural dynamics during infection.

Co-variant in duplexed regions of SARS-CoV-2

RNA duplexes usually are under the selection pressure to maintain the viral genome’s conformation to facilitate its replication [17]. To uncover duplexes under purifying selection, we first conducted an analysis to examine the co-variant base pairs in 429 non-redundant coronavirus genomes. This comparison revealed eight co-variant base pairs for SL5-6 in the 5’ UTR and 15 co-variant base pairs for the HVR in the 3’ UTR (Supplementary Fig. 5a,b and Supplementary Table 2). Importantly, we detected 13 duplexes

that spanned more than 600 nt and showed focused co-variations among multiple strains, highlighting the potential functionality of these long-range duplexes revealed by vRIC-seq (Supplementary Fig. 5c-e).

Next, we investigated the co-variant base pairs among 50,300 SARS-CoV-2 strains. This analysis identified 74 co-variant events across the whole SARS-CoV-2 genome (see red arc lines, Fig. 5a and Supplementary Table 3), of which six co-variants were located in SL1 and four in the s2m (see red shadowed regions in 5' UTR and 3' UTR, respectively. Fig. 5b,c). Moreover, we found ten co-variant events in a three-way junction located in ORF7a (Supplementary Fig. 5f), indicating that this junction might be important, and maintaining such a junction structure seemed critical to viral infection. Future investigations that experimentally determine the function of each three sequences should help elucidate the practical impact(s) of this structurally fascinating junction.

D614G and accompanying mutations on structure remodeling

SARS-CoV-2 genome is frequently mutated during evolution due to lacking proofreading activity of RdRP. Some mutations are beneficial to SARS-CoV-2 and emerge as dominant strains in the global pandemic, such as D614G and three accompanying mutations [41]. We found that these four mutations were located in local duplexed regions, and no interactions could be observed between them (Fig. 6a). The most prevalent D614G mutant caused by an A-to-G nucleotide transition at position 23,403, was located in the single-nucleotide bulge of a stem-loop (Fig. 6b). Interestingly, we found that the A23403G mutation fine-tuned the two local bulge structures into a thermodynamically more favorable six-nucleotide bulge structure (-10.3 kcal/mol vs. -13.7 kcal/mol) (Fig. 6b). Agreeing well with our model, the three uracil (U) in the bulge also showed strong SHAPE-MaP signals in the SARS-CoV-2 infected cells [24].

Additionally, the D614G accompanying mutations at 241 nt, 3,037 nt, and 14,408 nt all reside in the single-stranded internal or apical loops (Fig. 6c-e). These loops could also be partially supported by the SHAPE-MaP signals revealed in the host cells [24]. However, the C14408U mutation in RdRP (P323L) might introduce a novel structure with a smaller apical loop (14,380-14,441 nt) (Fig. 6e). It might be worthwhile to study further how these mutations are introduced during infection and how they enable viral fitness advantage. In addition, we systematically analyzed all the observed mutations in different SARS-CoV-2 strains stains. We found that the base-paired nucleotides are less mutated and have slightly lower entropy scores (Supplementary Fig. 6), further highlighting the functional importance of maintaining duplexes during evolution.

Structure-guided cleavage of SARS-CoV-2 RNA

RNA cleavage mediated by small interfering RNA (siRNA), antisense oligonucleotides, and RNA-targeting CRISPR, has emerged as a therapeutic modality to restrain viral gene expression and inhibit viruses in host cells [42]. However, it bears emphasis that target RNA cleavage efficiency is strongly affected by duplex structures [8,9]. In order to screen for cleavage-vulnerable regions of the SARS-CoV-2 RNA genome to develop anti-viral drugs, we systematically selected 130 single-stranded regions that i) have at least 10

consecutive nucleotides and ii) contain strong conservation among the different SARS-CoV-2 strains we examined. Next, we synthesized six siRNAs specifically targeting single-stranded regions and three siRNAs targeting duplex regions to cleave the SARS-CoV-2 RNA genome, and conducted infectivity assays using Vero cells (Fig. 7a and Supplementary Table 4).

qPCR analysis of cells at 24 hours post viral exposure showed that, compared to the non-targeting siRNA control (denoted as NC), four out of the six tested siRNAs targeting single-stranded regions could completely prevent SARS-CoV-2 amplification in Vero cells (see si-2, si-3, si-4, and si-5; Fig. 7b); accordingly, we observed no viral infection on cells treated with these four siRNAs (Fig. 7c). Agreeing well with the qPCR results, we found that the Vero cells treated with these four siRNAs looked healthier than those with the non-targeting siRNA control (Supplementary Fig. 7a). In sharp contrast, none of the siRNAs targeting duplexes could completely prevent infection or viral amplification (Fig. 7b,c). As each of the four effective siRNAs target sequences common to more than 90% of the 50,300 known SARS-CoV-2 strains, combining two or more of these siRNAs is expected to result in cleavage for nearly all known SARS-CoV-2 strains.

Importantly, we found that the structure of siRNA target regions revealed in virion is largely consistent with the structures uncovered in host cells (Fig. 7d). Furthermore, the siRNA targeting region designed by vRIC-seq showed more single-stranded base numbers than those from targeting regions selected by in silico methods (Supplementary Fig. 7b). This data highlight the SARS-CoV-2 RNA structure's values in guiding the design of potent siRNAs. Together, these results validated the accuracy of vRIC-seq deduced structures and showcased its practical utility for supporting drug development to combat SARS-CoV-2 during the ongoing COVID-19 pandemic.

Discussion

The emerging deadly RNA viruses have caused severe global epidemics and pandemics. RNA viruses usually form intricate and dynamic structures to orchestrate their translation, replication, and packaging [13,43]. Seeking to understand the in situ structure of viral RNA genome in virions, we developed vRIC-seq technology to map proximal contacts between different RNA fragments of the SARS-CoV-2. This global proximity information enabled us to construct a high-confidence RNA interaction map and build a 3D structure model of the SARS-CoV-2 genome, which is organized into an unentangled, globule, and seemingly spiral overall architecture. This knot-free globule conformation may allow for rapid unfolding after SARS-CoV-2 enters into the host cells and support maximally dense packaging of its RNA genome inside the viral particles. We also uncovered many short- and long-range duplexes, pseudoknots, and multiway junctions in SARS-CoV-2 RNA. Unexpectedly, these long-range duplexes could further isolate the SARS-CoV-2 genome into RNA topological domains. However, they are generally disrupted in the infected cells, indicating an active unwinding process that might be regulated by some RNA helicases.

We also developed an adaptive algorithm to reconstruct a complete SARS-CoV-2 secondary structure model by integrating the 3D pairwise interactions of vRIC-seq data. This approach is significantly

different from other widely used strategies that indirectly predict viral RNA duplex structures based on the nucleotide flexibility information [14,15,17,44,45]. Considering the duplex length restraint during prediction, the previously modelled structures of SARS-CoV-2 in the host cells might be incomplete because many functionally relevant long-range duplexes spanning over 600 nt were omitted [21-24].

The formation and maintenance of RNA duplexes are known to strongly influence viral fitness [46]. Thus, understanding single- and double-stranded regions in the viral genome is informative for the efficient development of RNA-targeted drugs to fight SARS-CoV-2 infections. The structural model we proposed here lays the foundation for developing and deploying highly potent siRNAs, single guide RNAs, and antisense oligos. Moreover, our model provides novel insights on how some pathogenic SARS-CoV-2 variants, such as D614G, enable viral fitness advantage at the structure level. However, it should be noted that the current structural model did not include binding information for the nucleocapsid (N) protein, which densely packages the SARS-CoV-2 genome into the ribonucleoprotein core in the interior of the virion [47]. Therefore, identification of the N protein's binding sites along the genome may further narrow down the targeting regions to develop agents for efficient RNA cleavage.

In summary, we developed a high-throughput approach for probing the in situ genome structure of any RNA viruses theoretically. We further used SARS-CoV-2 as a model to illustrate the power of vRIC-seq in delineating its general architecture and sophisticated long-range RNA-RNA interactions, such as duplexes and multi-way junctions. The vRIC-seq approach is also easily adaptable to probe viral RNA structurome and interactome in host cells in the future. More importantly, we could also apply the vRIC-seq technology in profiling RNA spatial interactions with animal cells and plant cells. The inclusion of the ConA beads capture step may significantly reduce the cell numbers used in our original RIC-seq protocol, thus enable RNA spatial interaction mapping started with a small number of cells.

References

1 Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine* (2020).

2 Ren, L. L. *et al.* Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. *Chinese medical journal* **133**, 1015-1024, doi:10.1097/CM9.0000000000000722 (2020).

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

4 de Wit, E., van Doremalen, N., Falzarano, D. & Munster, V. J. SARS and MERS: recent insights into emerging coronaviruses. *Nature reviews. Microbiology* **14**, 523-534, doi:10.1038/nrmicro.2016.81 (2016).

5 Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459-468, doi:10.1038/s41586-020-2286-9 (2020).

6 Kim, D. *et al.* The Architecture of SARS-CoV-2 Transcriptome. *Cell* **181**, 914-921.e910, doi:https://doi.org/10.1016/j.cell.2020.04.011 (2020).

7 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273, doi:10.1038/s41586-020-2012-7 (2020).

8 Patzel, V. *et al.* Design of siRNAs producing unstructured guide-RNAs results in improved RNA interference efficiency. *Nature biotechnology* **23**, 1440-1444, doi:10.1038/nbt1151 (2005).

9 Vickers, T. A., Wyatt, J. R. & Freier, S. M. Effects of RNA secondary structure on cellular antisense activity. *Nucleic acids research* **28**, 1340-1347, doi:10.1093/nar/28.6.1340 (2000).

10 Hsue, B. & Masters, P. S. 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, doi:10.1128/JVI.71.10.7567-7578.1997 (1997).

11 Madhugiri, R. *et al.* Structural and functional conservation of cis-acting RNA elements in coronavirus 5'-terminal genome regions. *Virology* **517**, 44-55, doi:10.1016/j.virol.2017.11.025 (2018).

12 Plant, E. P. *et al.* A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. *PLoS biology* **3**, e172, doi:10.1371/journal.pbio.0030172 (2005).

13 Yang, D. & Leibowitz, J. L. The structure and functions of coronavirus genomic 3' and 5' ends. *Virus Res* **206**, 120-133, doi:10.1016/j.virusres.2015.02.025 (2015).

14 Watts, J. M. *et al.* Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711-716 (2009).

15 Li, P. *et al.* Integrative Analysis of Zika Virus Genome RNA Structure Reveals Critical Determinants of Viral Infectivity. *Cell host & microbe* **24**, 875-886 e875, doi:10.1016/j.chom.2018.10.011 (2018).

16 Huber, R. G. *et al.* Structure mapping of dengue and Zika viruses reveals functional long-range interactions. *Nature communications* **10**, 1408, doi:10.1038/s41467-019-09391-8 (2019).

17 Dethoff, E. A. *et al.* Pervasive tertiary structure in the dengue virus RNA genome. *Proceedings of the National Academy of Sciences* **115**, 11513-11518, doi:10.1073/pnas.1716689115 (2018).

18 Barcena, M. *et al.* Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 582-587, doi:10.1073/pnas.0805270106 (2009).

19 Ke, Z. *et al.* Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature*, doi:10.1038/s41586-020-2665-2 (2020).

20 Yao, H. *et al.* Molecular Architecture of the SARS-CoV-2 Virus. *Cell* **183**, 730-738 e713, doi:10.1016/j.cell.2020.09.018 (2020).

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

22 Manfredonia I, N. C., Ponce-Salvatierra A, Ghosh P, Wirecki TK, Marinus T, Ogando, NS, Snider EJ, Hemert MJ, Bujnicki JM, Incarnato D. Genome-wide mapping of therapeutically-relevant SARS-CoV-2 RNA structures. *bioRxiv*, doi:https://doi.org/10.1101/2020.06.15.151647 (2020).

23 Lan TCT, A. M., Malsick LE, Khandwala S, Nyeo SSY, Bathe M, Griffiths A, Rouskin S. Structure of the full SARS-CoV-2 RNA genome in infected cells. *bioRxiv*, doi:doi:10.1101/2020.06.29.178343 (2020).

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

25 Kwok, C. K., Tang, Y., Assmann, S. M. & Bevilacqua, P. C. The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends in biochemical sciences* **40**, 221-232, doi:10.1016/j.tibs.2015.02.005 (2015).

26 Nicholson, B. L. & White, K. A. Functional long-range RNA–RNA interactions in positive-strand RNA viruses. *Nature Reviews Microbiology* **12**, 493-504, doi:10.1038/nrmicro3288 (2014).

27 Lange, S. J. *et al.* Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic acids research* **40**, 5215-5226, doi:10.1093/nar/gks181 (2012).

28 Ziv Omer, P. J., Shalamova Lyudmila, Kamenova Tsveta, Goodfellow Ian, Weber Friedemann, Miska Eric A. . The short- and long-range RNA-RNA Interactome of SARS-CoV-2. *Molecular cell*, doi:https://doi.org/10.1016/j.molcel.2020.11.004 (2020).

29 Ziv, O. *et al.* COMRADES determines in vivo RNA structures and interactions. *Nature methods* **15**, 785-788, doi:10.1038/s41592-018-0121-0 (2018).

30 Cai, Z. *et al.* RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature*, 1-6 (2020).

31 Watanabe, Y., Allen, J. D., Wrapp, D., McLellan, J. S. & Crispin, M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* **369**, 330-333, doi:10.1126/science.abb9983 (2020).

32 Oostra, M., de Haan, C. A., de Groot, R. J. & Rottier, P. J. Glycosylation of the severe acute respiratory syndrome coronavirus triple-spanning membrane proteins 3a and M. *Journal of virology* **80**, 2326-2336, doi:10.1128/JVI.80.5.2326-2336.2006 (2006).

33 Wakefield, J. K., Wolf, A. G. & Morrow, C. D. Human immunodeficiency virus type 1 can use different tRNAs as primers for reverse transcription but selectively maintains a primer binding site complementary to tRNA(3Lys). *Journal of virology* **69**, 6021-6029, doi:10.1128/JVI.69.10.6021-6029.1995 (1995).

34 Sanders W, F. E., Madden EA, Graham RL, Vincent HA, Heise MT, Baric RS, Moorman NJ. Comparative analysis of coronavirus genomic RNA structure reveals conservation in SARS-like coronaviruses. *bioRxiv*, doi:10.1101/2020.06.15.153197 (2020).

35 De Cesaris Araujo Tavares R, M. G., Pyle AM. The global and local distribution of RNA structure throughout the SARS-CoV-2 genome. *bioRxiv*, doi:https://doi.org/10.1101/2020.07.06.190660 (2020).

36 Rangan, R. *et al.* RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. *Rna* **26**, 937-959, doi:10.1261/rna.076141.120 (2020).

37 Goldsmith, C. S. *et al.* Ultrastructural characterization of SARS coronavirus. *Emerging infectious diseases* **10**, 320 (2004).

38 Gui, M. *et al.* Electron microscopy studies of the coronavirus ribonucleoprotein complex. *Protein & cell* **8**, 219-224 (2017).

39 Rieber, L. & Mahony, S. miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics* **33**, i261-i266, doi:10.1093/bioinformatics/btx271 (2017).

40 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).

41 Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827 e819, doi:10.1016/j.cell.2020.06.043 (2020).

42 Qureshi, A., Tantray, V. G., Kirmani, A. R. & Ahangar, A. G. A review on current status of antiviral siRNA. *Reviews in Medical Virology* **28**, e1976 (2018).

43 Masters, P. S. Coronavirus genomic RNA packaging. *Virology* **537**, 198-207, doi:10.1016/j.virol.2019.08.031 (2019).

44 Simon, L. M. *et al.* In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. *Nucleic Acids Research* **47**, 7003-7017, doi:10.1093/nar/gkz318 (2019).

45 Tomezsko, P. J. *et al.* Determination of RNA structural diversity and its role in HIV-1 RNA splicing. *Nature* **582**, 438-442, doi:10.1038/s41586-020-2253-5 (2020).

46 Boerneke, M. A., Ehrhardt, J. E. & Weeks, K. M. Physical and Functional Analysis of Viral RNA Genomes by SHAPE. *Annual review of virology* **6**, 93-117, doi:10.1146/annurev-virology-092917-043315 (2019).

47 Kang, S. *et al.* Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm Sin B* **10**, 1228-1238, doi:10.1016/j.apsb.2020.04.009 (2020).

48 Gao, Q. *et al.* Development of an inactivated vaccine candidate for SARS-CoV-2. *Science* (2020).

49 Cameron, V. & Uhlenbeck, O. C. 3'-Phosphatase activity in T4 polynucleotide kinase. *Biochemistry* **16**, 5120-5126, doi:10.1021/bi00642a027 (1977).

50 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

51 Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639-1645 (2009).

52 Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95-98, doi:https://doi.org/10.1016/j.cels.2016.07.002 (2016).

53 Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis* **33**, 1029-1047 (2013).

54 Reuter, J. S. & Mathews, D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics* **11**, 1-9 (2010).

55 Popenda, M. *et al.* Automated 3D structure composition for large RNAs. *Nucleic acids research* **40**, e112, doi:10.1093/nar/gks339 (2012).

56 Cleveland, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* **74**, 829-836, doi:10.1080/01621459.1979.10481038 (1979).

57 Darty, K., Denise, A. & Ponty, Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974 (2009).

58 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26 (2011).

59 Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research* **31**, 3406-3415 (2003).

60 Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for molecular biology* **6**, 26 (2011).

61 Huang, L. *et al.* LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics* **35**, i295-i304, doi:10.1093/bioinformatics/btz375 (2019).

62 Consortium, T. R. RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Research* **45**, D128-D134, doi:10.1093/nar/gkw1008 (2016).

63 Pickett, B. E. *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic acids research* **40**, D593-D598 (2012).

64 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* **30**, 772-780 (2013).

65 Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935 (2013).

66 Rivas, E., Clements, J. & Eddy, S. R. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature methods* **14**, 45-48 (2017).

67 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1**, 33-46 (2017).

68 Lai, D., Proctor, J. R., Zhu, J. Y. A. & Meyer, I. M. R- chie : a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Research* **40**, e95-e95, doi:10.1093/nar/gks241 (2012).

69 Zhao, W.-M. *et al.* The 2019 novel coronavirus resource. *Yi chuan = Hereditas* **42**, 212-221 (2020).

70 Lagana, A., Shasha, D. & Croce, C. M. Synthetic RNAs for Gene Regulation: Design Principles and Computational Tools. *Front Bioeng Biotechnol* **2**, 65, doi:10.3389/fbioe.2014.00065 (2014).

Methods

Experimental model and subject details

Vero cells (CCL-81) were cultured in DMEM (Thermo Fisher, C11965500BT) containing 1% penicillin/streptomycin (Life Technologies, 15140) and 10% fetal bovine serum. All the live SARS-CoV-2 viruses related experiments were carried out in the enhanced biosafety level 3 (P3+) facilities, which are authorized by the National Health Commission of the People's Republic of China. The SARS-CoV-2 strain used in this study, IPBCAMS-YL01/2020, was isolated from a clinical sample by the team of Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College. The virus was passaged three times in Vero cells (ATCC, CCL-81) before use. Infectious titers of SARS-CoV-2 were determined by plaque assay in Vero cells. SARS-CoV-2 virions were prepared as previously described and inactivated by b-propiolactone before bringing to the general laboratory for vRIC-seq analysis [48].

Virion capture and crosslinking

We used BioMag®Plus Concanavalin A beads (Polysciences Inc, 86057) to capture the SARS-CoV-2 virions via spike glycoprotein. To this end, 150 µl of Concanavalin A beads were first washed twice with binding buffer (20 mM HEPS-KOH pH 7.9, 10 mM KCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$). The supernatant was discarded and the beads were resuspended in 400 µl of binding buffer. Subsequently, the 100 µl of virions

were added to the tube and incubated for 10 min at room temperature (RT). After washing three times with binding buffer and once with PBST buffer (1 × PBS, 0.1% Tween 20), the beads were resuspended in 1 ml of PBST buffer. To fix nucleocapsid (N) protein-mediated RNA-RNA interactions, 27 µl of 37% formaldehyde was applied and incubated for 10 min at RT with rotation at 20 rpm. The crosslinking was stopped by adding glycine to a final concentration of 0.125 M and incubated at RT for 5 min.

Permeabilization and MNase treatment

The Concanavalin A beads and captured virions were resuspended in 1 ml of permeabilization buffer (10 mM Tris-HCl pH 7.5, 10 mM NaCl, 0.5% NP-40, 0.3% Triton X-100, 0.1% Tween 20, 1 × protease inhibitors (Sigma Aldrich, P8340), 2 U/ml SUPERase In RNase inhibitor (Thermo Fisher Scientific, AM2694)), and rotated at 4 °C for 15 min at 20 rpm. After washing three times with 1× PNK buffer (50 mM Tris-HCl pH 7.4, 10 mM $MgCl_2$, 0.1% Tween 20), the beads were further treated with 200 µl of 1 × MN mixture (50 mM Tris-HCl pH 8.0, 5 mM CaCl2, 0.03 U/µl micrococcal nuclease (Thermo Fisher, EN0181)) for 10 min at 37 °C. The MNase treatment was stopped by washing twice with 1× PNK + EGTA buffer (50 mM Tris-HCl pH 7.4, 20 mM EGTA, 0.1% Tween 20) and twice with 1× PNK buffer.

pCp-biotin labelling and proximity ligation

The beads were gently resuspended in 100 µl of 1× FastAP buffer containing 10 U of FastAP alkaline phosphatase (Thermo Fisher Scientific, EF0651) and incubated at 37 °C for 15 min. To stop the reaction, the beads were sequentially washed twice in 1× PNK + EGTA buffer, twice in 1× high-salt wash buffer (5× PBS, 0.1% Tween 20) (Pipette 6-8 times, this step should be brief), and three times in 1× PNK buffer. To deposit pCp-biotin, beads were resuspended in 100 µl of ligation mixture containing 10 µl of 10× RNA ligase reaction buffer, 6 µl of 40 U/µl RNase inhibitor, 4 µl of 1 mM pCp-biotin (Thermo Fisher Scientific, 20160), 100 U of T4 RNA ligase, 20 µl of nuclease-free water and 50 µl of 30% PEG 20000, and incubated overnight at 16 °C. After washing three times with 1× PNK buffer, the beads were resuspended in a PNK mixture (10 µl of 10× Imidazole buffer (500 mM imidazole-HCl pH 6.4, 100 mM $MgCl_2$), 6 µl of 10 mM ATP, 4 µl of T4 polynucleotide kinase (Thermo Fisher Scientific, EK0032), 10 µl of 0.1 M DTT, 70 µl of nuclease-free water) and incubated at 37 °C for 45 min. Of note, T4 PNK possesses maximal kinase and 3’ phosphatase activity simultaneously in Imidazole buffer with pH 6.4 [49]. Subsequently, the beads were washed twice with 1× PNK + EGTA buffer and twice with 1× PNK buffer. For in situ proximity ligation, the ligation mixture containing 20 µl of 10× RNA ligase reaction buffer, 8 µl of 40 U/µl RNA inhibitor (Thermo Fisher Scientific, EO0381), 10 µl of 10 U/µl T4 RNA ligase (Thermo Fisher Scientific, EL0021), 20 µl of 1 mg/ml BSA, and 142 µl of nuclease-free water was used to gently resuspend the beads and incubated overnight at 16 °C.

RNA purification and library construction

At the next day, the ligation was stopped by washing three times with 1× PNK buffer. To extract viral RNAs, the beads were resuspended and incubated with 200 µl of proteinase K buffer (10 mM Tris-HCl pH 7.5, 10 mM EDTA, 0.5% SDS) and 50 µl of proteinase K (Takara, 9034) at 37 °C for 60 min and 56 °C for

15 min. The Eppendorf tube was placed on a magnet stand for 1 min, and the supernatant was transferred to a new tube. The viral RNA was extracted from the supernatant with 750 µl of TRIzol LS (Thermo Fisher, 10296028) and 220 µl of chloroform according to the manufacturer's instruction. After centrifugation at 13,000 rpm for 15 min at 4 °C, the supernatant was transferred to a 1.5 ml Eppendorf tube and mixed with 500 µl of isopropanol and 1 µl of glycoblue (15 mg/ml, Thermo Fisher, AM9515). The RNA was precipitated overnight at -20 °C and pelleted at 13,000 rpm for 20 min at 4 °C. After washing twice with 75% ethanol, the RNA pellet was resuspended in 15 µl of nuclease-free water. The subsequent steps of vRIC-seq were performed exactly as we previously described [30], including RNA fragmentation, pCp-biotin selection, and strand-specific library preparation.

siRNA transfection and RT-qPCR

We synthesized the siRNAs in GenePharma and transfected 10 pmol of each siRNAs into 8×10^4 Vero cells with Lipofectamine RNAiMAX by following the manufacturer's instructions. The 24-well plates containing the transfected cells were brought to the biosafety level 3 (P3+) facilities for SARS-CoV-2 infection after 24 h. The cells were infected with SARS-CoV-2 (MOI = 0.05) for 1 h at 37°C in 500 µl Opti-MEM medium. After removing the incubation medium, the infected Vero cells were washed once with Opti-MEM medium and cultured for an additional 24 h in maintenance medium (OPTI-MEM medium containing 1% BSA and 1% penicillin/streptomycin).

For measuring the SARS-CoV-2 levels in the supernatant, 100 µl of virus-containing medium was collected. The viral RNA was extracted with TRIzol LS Reagent (Invitrogen, 10296028) and purified by using Direct-zol™ RNA MiniPrep (ZYMO RESEARCH, R2050) according to the manufacturer's instructions. The TaqMan RT-PCR assays were performed using TaqMan Fast Virus 1-Step Master Mix (Thermo Fisher Scientific, 4444432). The primers targeting the nucleocapsid (N) gene of SARS-CoV-2 and probe were listed in Supplementary Table 4.

To quantify viral RNA levels inside the cell, we first extracted total RNAs from the infected Vero cells using TRIzol Reagent (Invitrogen, 15596026). For RT-qPCR, 1 µg of total RNA was treated with RQ1 RNase-free DNase (Promega, M6101) and converted into cDNA using MMLV reverse transcriptase (Promega, M1701) with oligo dT (20) primer. qPCR was performed with Hieff qPCR SYBR Green Master Mix (YEASEN, 11203ES08) on a StepOnePlus real-time PCR machine (Applied Biosystems). The primers targeting the RdRp gene of SARS-CoV-2 were used for qPCR.

Processing of vRIC-seq data

After removing adapters, PCR duplicates, and low-quality sequences, we first aligned the unique vRIC-seq reads to the human 45S pre-rRNA (Refseq accession number NR_046235.3). For all the unmapped reads, we further mapped them to a pan-genome consisting of the *Chlorocebus sabaeus* reference genome (genome assembly version: chlSab2) and the SARS-CoV-2 reference genome (Refseq accession number NC_045512.2) using a STAR software (v020201) [50]. Chimeric reads were identified using the previously

described RIC-seq analysis pipeline [30]. Chimeric reads coverage along the genome was visualized by the Circos suite (v0.69-5) [51].

The RNA interaction map of SARS-CoV-2

The chimeric reads mapped to the SARS-CoV-2 reference genome were collected and used for generating an RNA interaction map. Of note, we removed spliced reads containing gaps resulted from discontinuous transcriptions [6]. As described previously [30], we identified the pairwise junction sites in the chimeric reads and used them for building a two-dimensional RNA interaction matrix. This matrix could be converted into *.hic* format and visualized by the Juicebox tool (v1.11.08) [52]. The Knight-Ruiz algorithm [53] was used to balance this interaction matrix to eliminate sequencing bias along the virus genome.

Predict local structures of SARS-CoV-2 for validating vRIC-seq data

We used the extended 5' UTR (1-480 nt) and 3' UTR (29,546-29,870 nt) sequences to generate candidate secondary structures by the Fold program from RNAstructure software suite (v6.2) [54]. As a general approach, the maximum distance between two paired nucleotides was allowed within 250 nt. The structure that agreed well with the maximized RIC-seq signals between pairwise interacting RNA fragments was selected and visualized by the StructureEditor program (v6.0). The RNAComposer software was applied to deduce the three-way junction's 3D structure [55], and the configuration agreed with vRIC-seq data was chosen.

RNA topological domain in SARS-CoV-2 genome

The SARS-CoV-2 genome was separated into isolated topological domains using our previously published iterative algorithm with minor revisions [30]. Briefly, we iteratively chose the optimal boundary that minimized the inter-domain's vRIC-seq density as a new candidate domain boundary. To avoid tiny domains resulting from over division, we stopped the iteration once more than 35% of the pairwise 10-nt windows (connection score greater than 0.01) were classified as inter-domain. Lastly, adjacent domains were merged if both did not contain interactions between their 5' and 3' boundary.

MFE analysis of the duplexes revealed by vRIC-seq

To explore whether the pixels with a strong vRIC-seq signal in the RNA interaction map could form long-range duplexes, we first divided the SARS-CoV-2 genome into non-overlapping 10-nt windows. The pairwise 10-nt windows with a connection score greater than 0.01 were used for downstream analyses [30]. Next, we clustered pairwise 10-nt windows adjoining or overlapping at both ends as one interaction. The lowest hybrid free energy was then computed for the possible hybrids formed between these pairwise RNA stretches using the bifold function in the RNAstructure suite (v6.2) with default parameters [54]. Lastly, artificial sequences with the same nucleotide content as real interactions were generated ten times. The lowest hybrid free energy for those shuffled sequences was also calculated.

3D structural simulation of SARS-CoV-2 genome

Based on the RNA interaction map, we used the miniMDS program to model the spatial conformation of the SARS-CoV-2 genome using the following parameters [39]: minimds.py -l 10 -m 0.01 -p 0.01. The spatial coordinates reported by miniMDS were smoothed with the LOWESS (locally weighted scatterplot smoothing) algorithm and then visualized [56].

SARS-CoV-2 RNA secondary structure modeling

The secondary structure of SARS-CoV-2 genomic RNA was constructed in silico based solely on the vRIC-seq data by an adaptively optimized algorithm we developed in this study. We first split the SARS-CoV-2 genome into shorter segmental domains by maximizing the ratio between intra-domain and inter-domain's vRIC-seq signals. Notably, the domains smaller than 4 kb will not be further split to avoid the potential loss of long-range duplexes over the domains' boundaries. Like a previously described approach [15], we determined the secondary structure for each domain independently. To this end, we systematically screened pairwise 5-nt windows with connection scores higher than 0.03, and the windows adjoining or overlapping at both ends were further clustered as high-confidence interactions. For each interaction spanned region within a domain, we used the Fold program in the RNAstructure software suite (v6.2) to perform structure prediction [54]. The maximum distance between any two paired positions was allowed within 2500 nt. From the structural candidates reported by the Fold program, we selected the one that matched best with vRIC-seq data and forced it as a constraint in the subsequent prediction. Of note, we generated duplexes for short local interactions first and then used them as restrains to perform prediction for long-range interactions spanned regions. Moreover, interactions having stronger vRIC-seq signals were processed with priority. Finally, by restraining duplexes generated in the former stage, we folded each domain's entire sequence, including regions not covered by the high-confidence interactions. The structure agreed best with vRIC-seq signals were selected. The final secondary structure model of the viral genome RNA in SARS-CoV-2 was visualized by the VARNA program (v3-93) [57] and the Integrative Genomics Viewer (IGV) visualization tool (v.2.3.92) [58].

Evaluate the accuracy of the algorithm

To evaluate the adaptive algorithm's accuracy, we first used our previously published rRNA+ RIC-seq data in HeLa cells to predict the secondary structure of 28S rRNA [30]. We also used four widely used computational algorithms to predict the secondary structure based on the minimum free energy, including RNAstructure (v6.2) [54], Mfold (v3.6) [59], RNAfold [60], and LinearFold [61]. All parameters were set as default except for the maximum pairing distance within 1600 nt was allowed.

We used two criteria to evaluate a predicted secondary structure's accuracy: sensitivity and PPV (positive predictive value). Sensitivity was defined as the fraction of base pairs that were correctly predicted. PPV was the fraction of base pairs in the predicted structure that were correct. We used a relaxed structure comparison mode. A base pair *i*/*j* was considered correctly predicted if any of the following pairs exist in

the reference structure: $i\pm5/j\pm5$ and vice versa. The annotated structural model of 28S rRNA was downloaded from the RNAcentral database (https://rnacentral.org/rna/ URS000075EC78/9606) [62].

Identification of co-variant base pairs

A total of 429 non-redundant coronavirus genomes were downloaded from the ViPR database [63] and aligned by the MAFFT program (v7.471) [64] with default parameters to identify co-variant base pairs in SARS-CoV-2. Based on the multiple sequence alignment results and the SARS-CoV-2 secondary structure, we employed the cmbuild and cmcalibrate program in the Infernal package (v1.1.3) [65] to build the covariance model and used the cmserach in the same package to search homologs. Homologs with E-value higher than 0.01 were removed, and the alignment of left sequences was subjected to the R-scape program (v1.5.4) [66] to explore covariance in the SARS-CoV-2. Co-variant base pairs among different SARS-CoV-2 strains were inferred from multiple sequence alignment of 50,300 SARS-CoV-2 strains, which were downloaded from the Global Initiative for Sharing All Influenza Data (GISAID) database [67] (date: June 29, 2020). A base pair was classified as a co-variant event when both nucleotides were different from the reference SARS-CoV-2 sequence but still base-paired. Co-variant base pairs were plotted using the R-chie program [68].

SNP density and entropy score calculation

SNP annotation for SARS-CoV-2 was downloaded from China National Center for Bioinformation (https://bigd.big.ac.cn/ncov, date: July 13, 2020) [69]. Each nucleotide's entropy score was calculated according to the multiple sequence alignment of 50,300 SARS-CoV-2 genome sequences. The following formula as previously described was used: $S = -100\times\mathrm{Sum}\ (P_i\times\log_2 P_i)$, in which P_i is the frequency of the i^{th} allele [63]. Gaps '-' at either end of a sequence in the alignment were removed, and ambiguous residues 'N' were excluded from entropy calculation.

siRNA design

Based on the single-stranded regions revealed by vRIC-seq and the siRNA designing principles described earlier [70], we randomly selected six siRNAs for experimental validation of SARS-CoV-2 silencing in Vero cells. To explore whether the in-virion SARS-CoV-2 structure could guide siRNA design, we collected 28 siRNAs designed by in silico modeling [36], 30 siRNAs by the ViennaRNA web server (http://rna.tbi.univie.ac.at/cgi-bin/RNAxs/RNAxs.cgi), as well as all the 331 siRNAs designed from the linear sequence of SARS-CoV-2. Using these collections, we compared the number of single-stranded nucleotides at siRNA-targeted regions in infected cells. We found that the six potent siRNAs target regions also tend to be single-stranded in the host cells.

Comparison with the in-cell structure

The normalized SHAPE-MaP reactivities on each nucleotide of the SARS-CoV-2 RNA in the infected Vero E6 cells were downloaded from Anna Pyle's laboratory with the link of https://github.com/pylelab/SARS-

CoV-2_SHAPE_MaP_structure [24]. To evaluate the structural dynamics along the SARS-CoV-2 genome, we counted the percentage of shared base pairs within 1 kb sliding windows for every ten nt along the entire RNA genome. The center of sliding windows moved from the first nucleotide to the 29901st nucleotide. When the center of a sliding window is smaller than 501 or larger than 29401, the 5' or 3' end of the sliding window exceeds the boundary of the viral genome. Therefore, we truncated these sliding windows at the endpoint of the viral genome and counted the percentage of shared base pairs within the truncated windows.

Declarations

Acknowledgements

This work was supported by National Key R&D Program (2017YFA0504400), the NSFC (91740201, 91940306, and 81921003), and the Strategic Priority Program of CAS (XDB37000000) to Y.X., by the NSFC (31900465) to C.C., by the National Key R&D Program (2020YFA0707500, 2018YFA0900801) and the Strategic Priority Research Program (XDB29010000) to X.W., by National Key R&D Program (2020YFA0707600), National Major Sciences & Technology Project for Control and Prevention of Major Infectious Diseases in China (2018ZX10301401), and Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (2016-I2M-1-014) to J.W.

Author contributions

Y.X. and J.W. initiated and planned the study; C.C. performed the bioinformatics analysis with the help of N.H. and B.Z.; Z.C. developed vRIC-seq technology, created the deep-sequencing library, and performed most experiments; X.X. and J.R. performed SARS-CoV-2 infection and qPCR quantification experiments under the guidance of J.W.; M.Y. and X.X. prepared virions under the guidance of X.W.; Y.X., C.C. and J.W. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

vRIC-seq data have been deposited in the GEO under accession number GSE155733.

Code availability

Homemade scripts for reconstructing the secondary structure of SARS-CoV-2 genome and for simulating global configuration could be found at https://github.com/caochch /RIC2Structure.

Figures

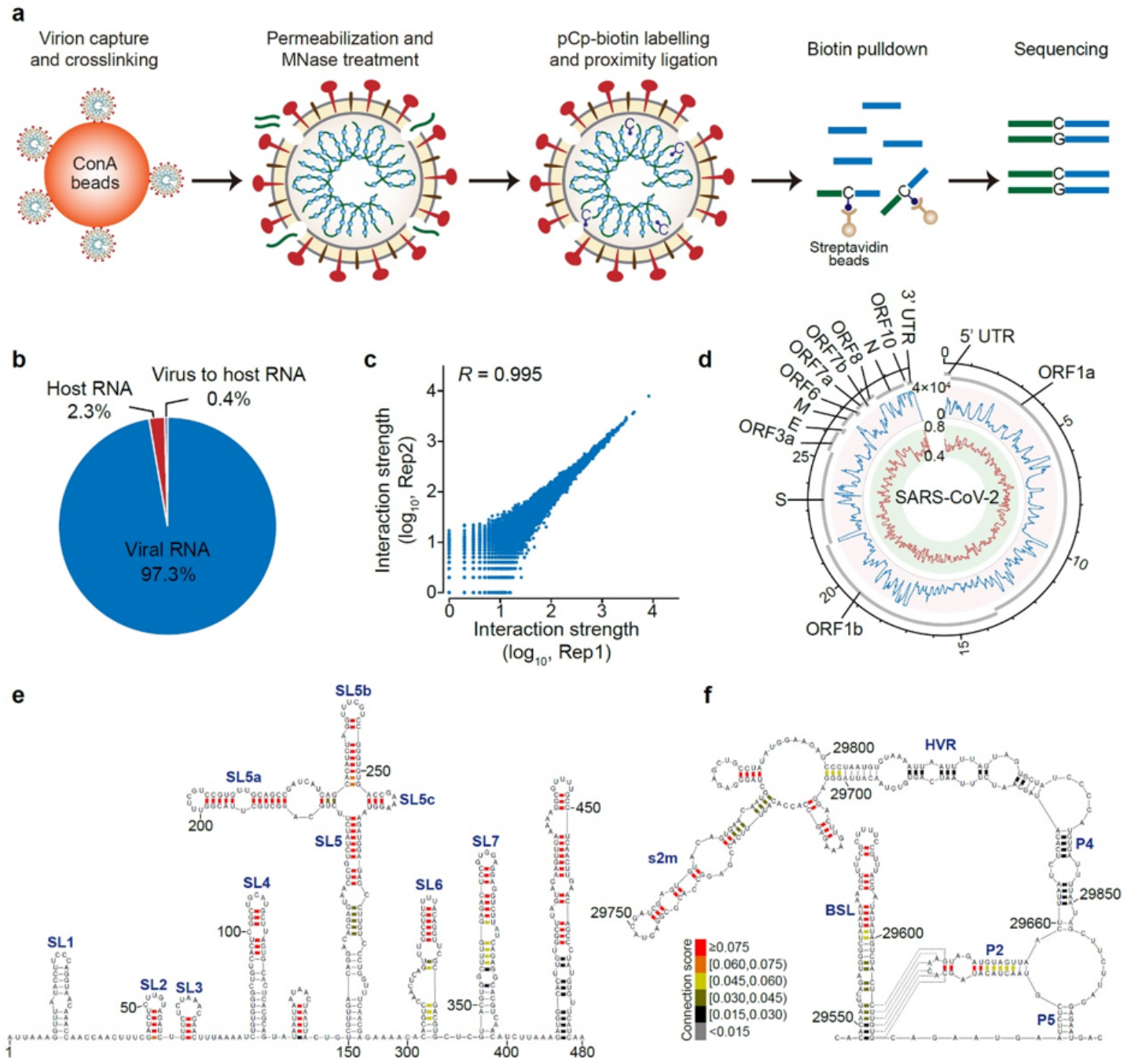

Figure 1

Overview and evaluation of vRIC-seq technology. a, Scheme of vRIC-seq technology. Concanavalin A (ConA) beads were used to capture the virion for diverse enzyme treatments in subsequent steps. b, The proportions of chimeric reads mapped to SARS-CoV-2. c, Scatter plots showing the correlation between two biological replicates for the number of chimeric reads (interaction strength). R, Pearson correlation coefficient. d, Circos plot showing the distribution of chimeric reads along the SARS-CoV-2 genome. The inner red circle stands for the fractions of adenine or uracil within 100 nt windows, and the outer blue circle shows the coverage of chimeric reads. e, f, vRIC-seq confirmed known coronavirus RNA structures in the 5' UTR (1-480 nt, e) and 3' UTR (29,546-29,870 nt, f) of the SARS-CoV-2 RNA genome. Connection

scores shown in different colors were used for assessing the base-pairing probability. The dashed lines illustrated the pseudoknot.

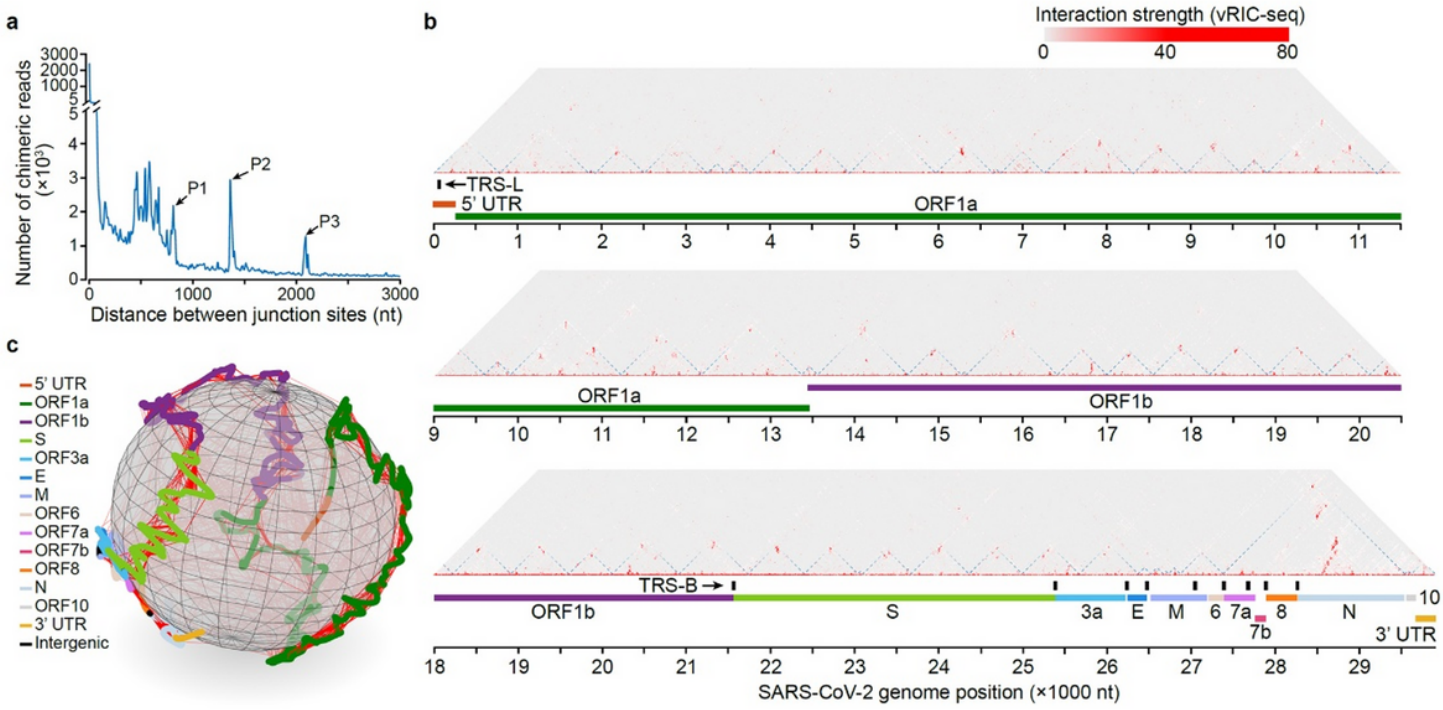

Figure 2

Global view of SARS-CoV-2 genome organization. a, Spanning distance of pairwise interacting RNAs. P1, P2, and P3 mark three peaks corresponding to chimeric interactions spanning 810, 1360, and 2090 nucleotides. b, RNA interaction map of the SARS-CoV-2 genome. The dashed triangles represent RNA topological domains. Transcription-regulatory sequence in the leader (TRS-L) and the body (TRS-B) are marked as black lines. c, The global configuration of the SARS-CoV-2 RNA genome in virions, modelled by the miniMDS software. The vRIC-seq detected RNA contact frequencies were used for the modelling. The solid red lines represent chimeric signals that support the local interactions, whereas the dashed red lines depict long-range interactions.

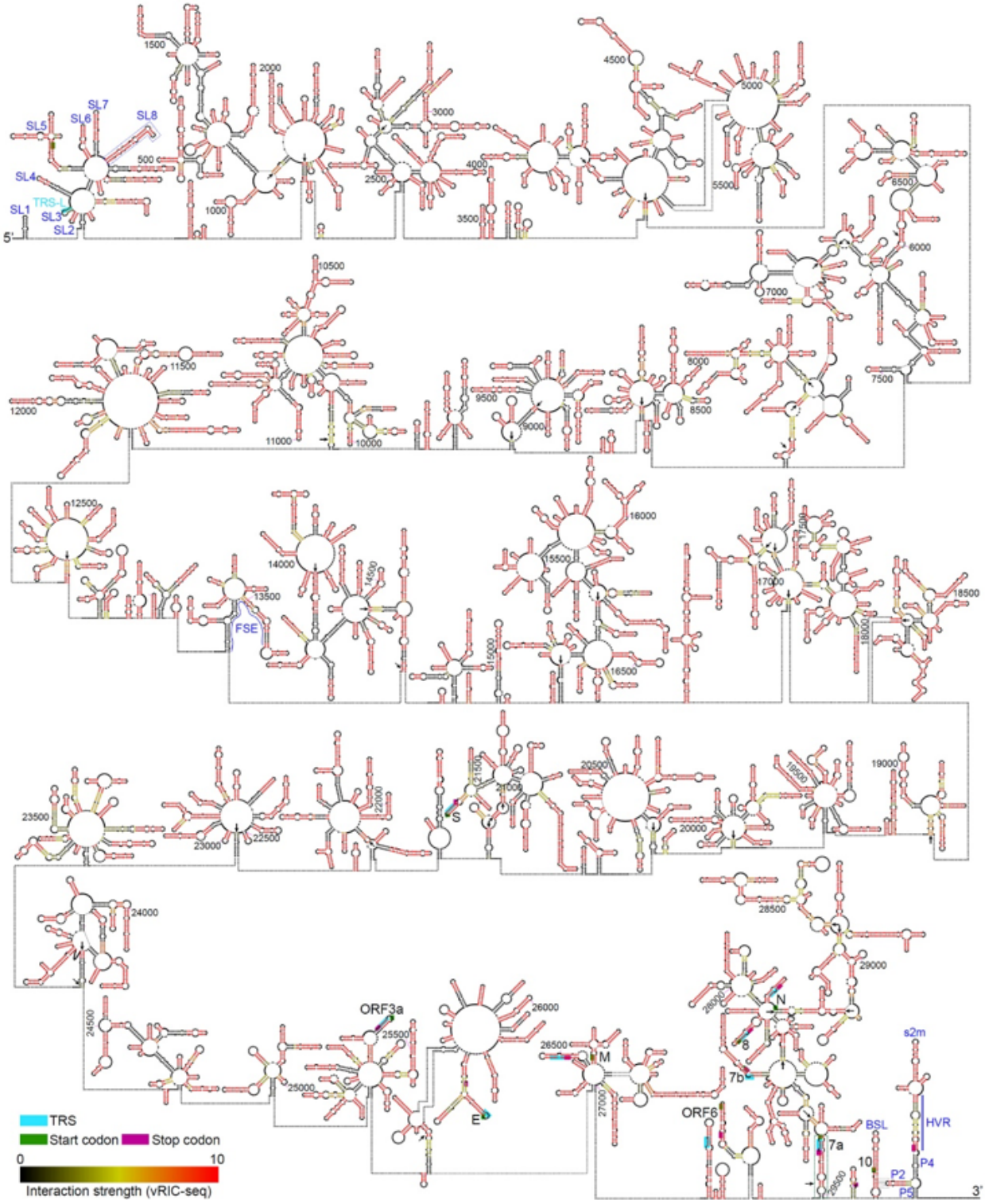

Figure 3

The secondary structure of the SARS-CoV-2 genome. The known structural elements in the 5’ UTR, the frame-shifting element (FSE), and the 3’ UTR are labelled or marked in blue. The pairwise interaction strength was quantified and shown in different colors. Black arrows highlight highly-confident long-range duplexes measured by vRIC-seq signals. Green and purple boxes mark the start and stop codons, respectively. A cyan box outlines the core sequence (CS) of each transcription-regulatory sequence (TRS).

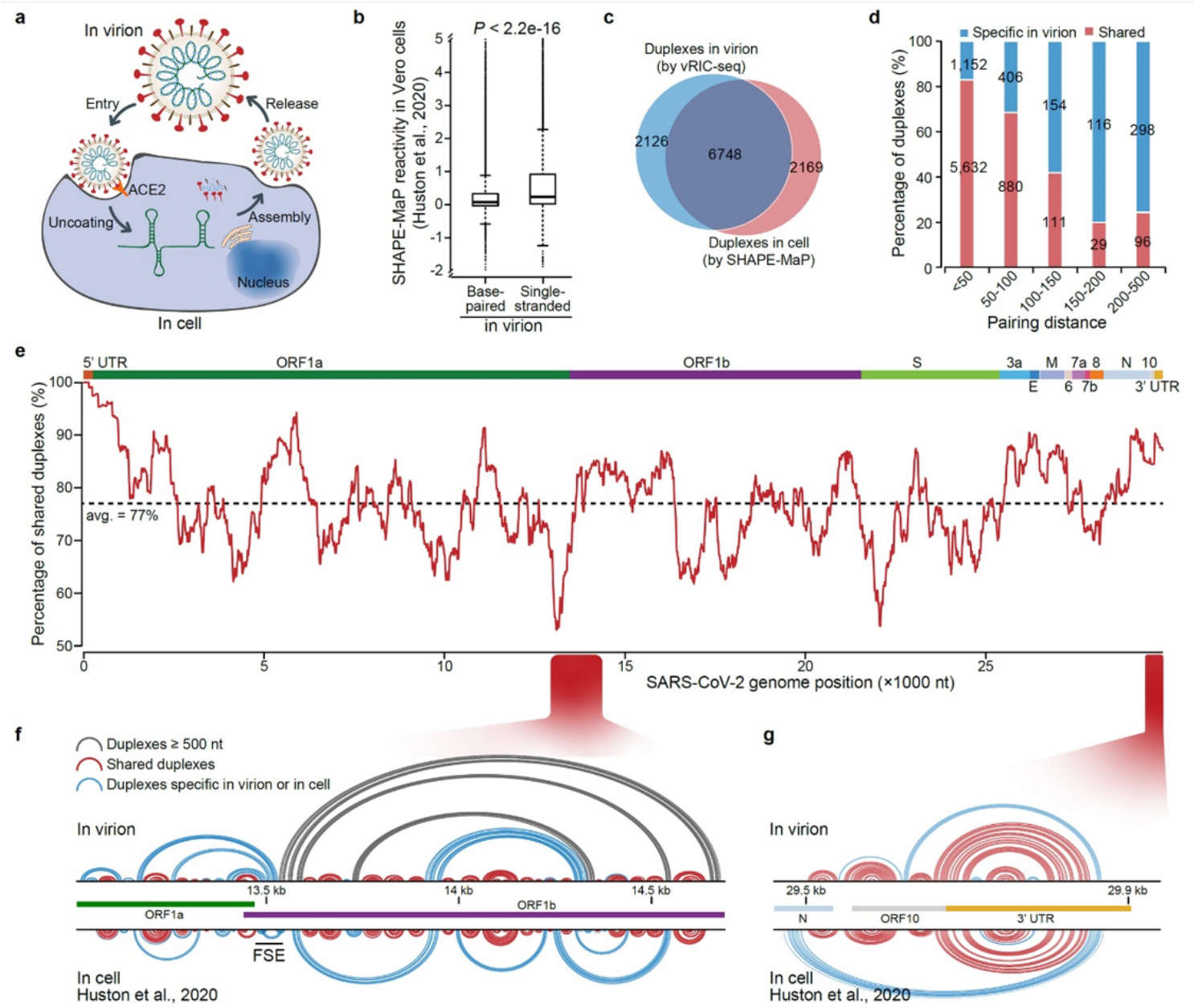

Figure 4

Comparison of SARS-CoV-2's structure in virions and host cells. a, Schematic diagram showing the life cycle of SARS-CoV-2. b, Single-stranded nucleotides identified by vRIC-seq in virions have higher SHAPE-MaP reactivities than base-paired nucleotides in cells. P-value was determined by two-tailed, unpaired Student's t-test. The centre line of the box plot represents the median, the box borders represent the first (Q1) and third (Q3) quartiles, and the whiskers are the most extreme data points within 1.5× the interquartile range (from Q1 to Q3). c, Venn diagram showing the overlap between duplexes revealed in virions and cells. d, The percentage of shared duplexes in virions and cells decreased along with the spanning distance. e, The percentage of shared duplexes along the SARS-CoV-2 genome in sliding 1 kb windows. The dashed line indicates the average percentage. f,g, In-virion (top) and in-cell (bottom) duplexes in the FSE (f) and 3' UTR (g) surrounded region. Arc lines colored in grey indicate base pairs

spanning more than 500 nt, arc lines in red indicate base pairs shared by duplexes in virions and cells, while arc lines in blue indicate base pairs specific in virions or cells.

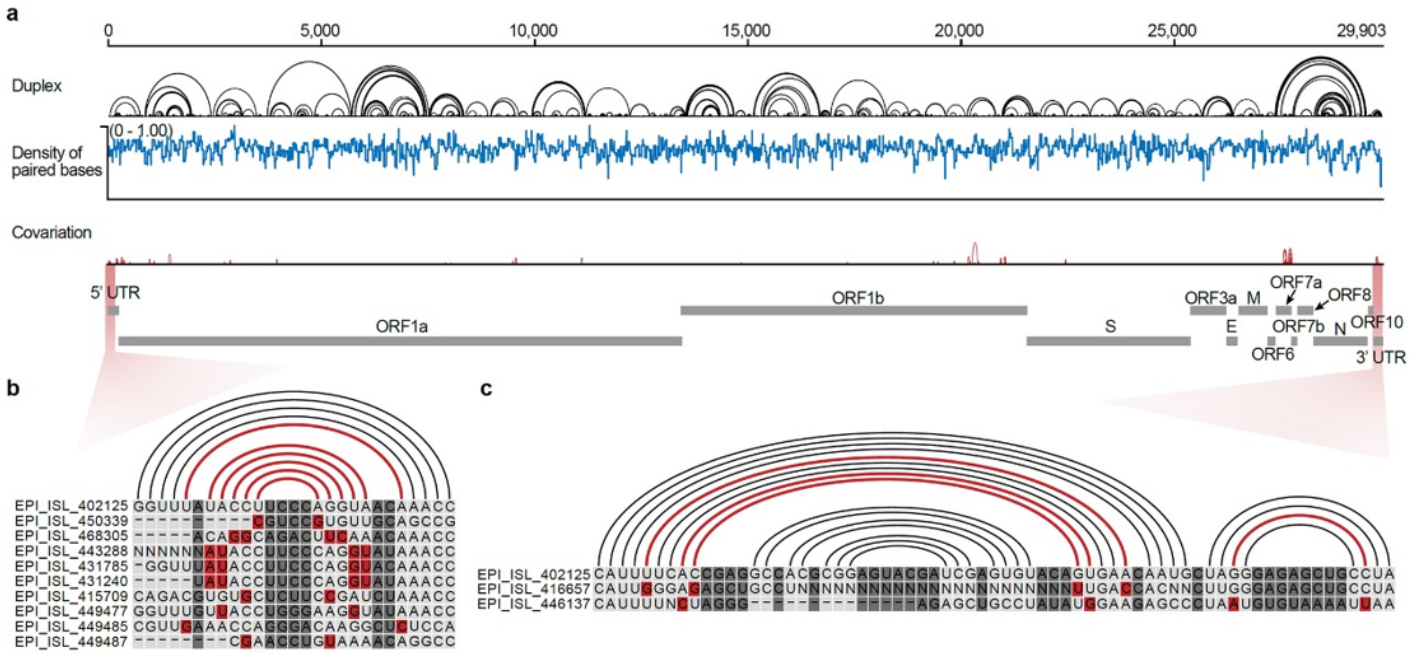

Figure 5

Co-variation among duplexes in different SARS-CoV-2 strains. a, Profile of structural elements and variants in the SARS-CoV-2 genome. The black arc lines stand for base pairs, the blue line indicates the density of base-paired nucleotides, and the red arc lines denote co-variant base pairs among different SARS-CoV-2 strains. b,c, Co-variant base pairs in the 5’ UTR (b) and 3’ UTR (c), respectively. Arc lines and nucleotides colored in red indicate co-variant base pairs. EPI_ISL_402125 is the reference SARS-CoV-2 sequence.

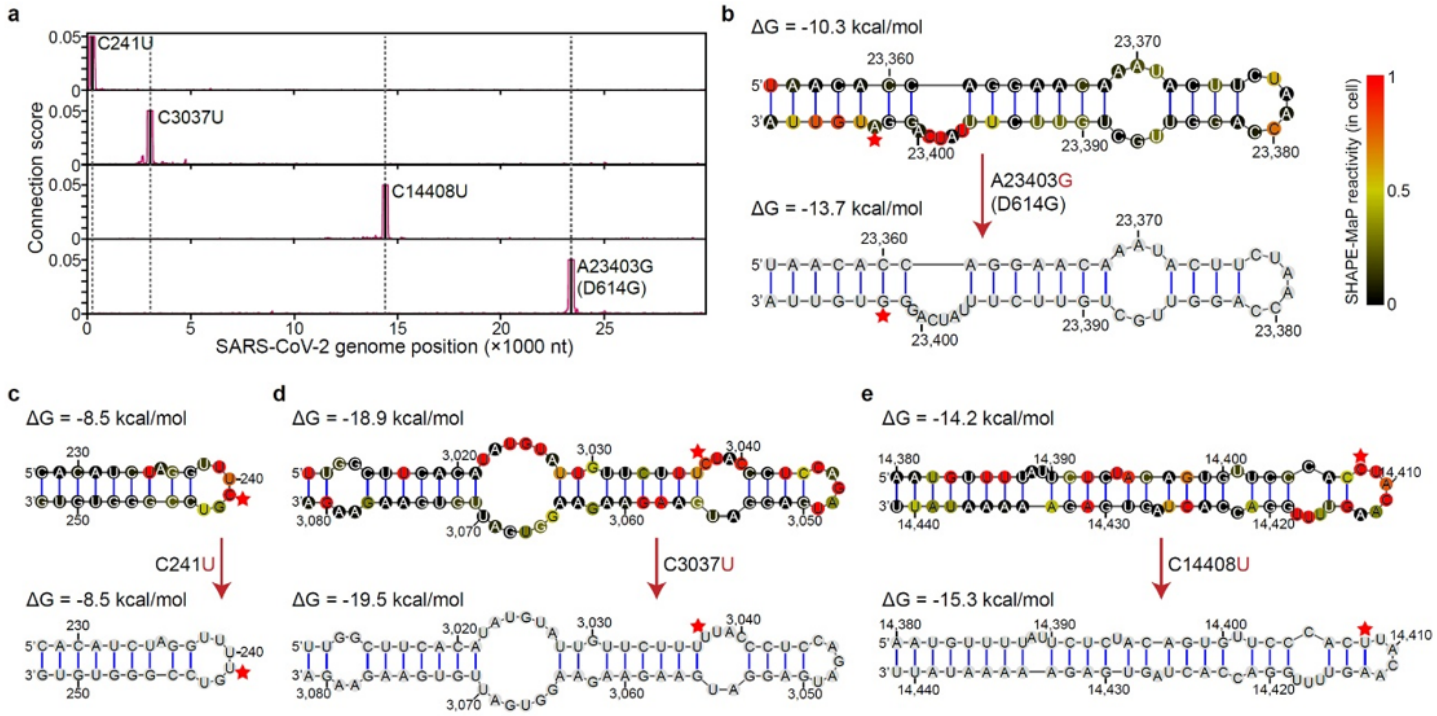

Figure 6

D614G and the accompanying mutations on structure remodeling. a, Line plot showing the point mutation resided regions tend to form local interactions. The C241U, C3037U, C14408U, and A23403G (D614G) mutants are marked as solid lines. b, The A-to-G transition (D614G mutant) at 23,403 nt remodels two bulge structures into a single six-nucleotide bulge. Red stars mark the mutated nucleotides. c-e, The D614G accompanying mutations have no influence on duplexes except for the C14408U transition (e).

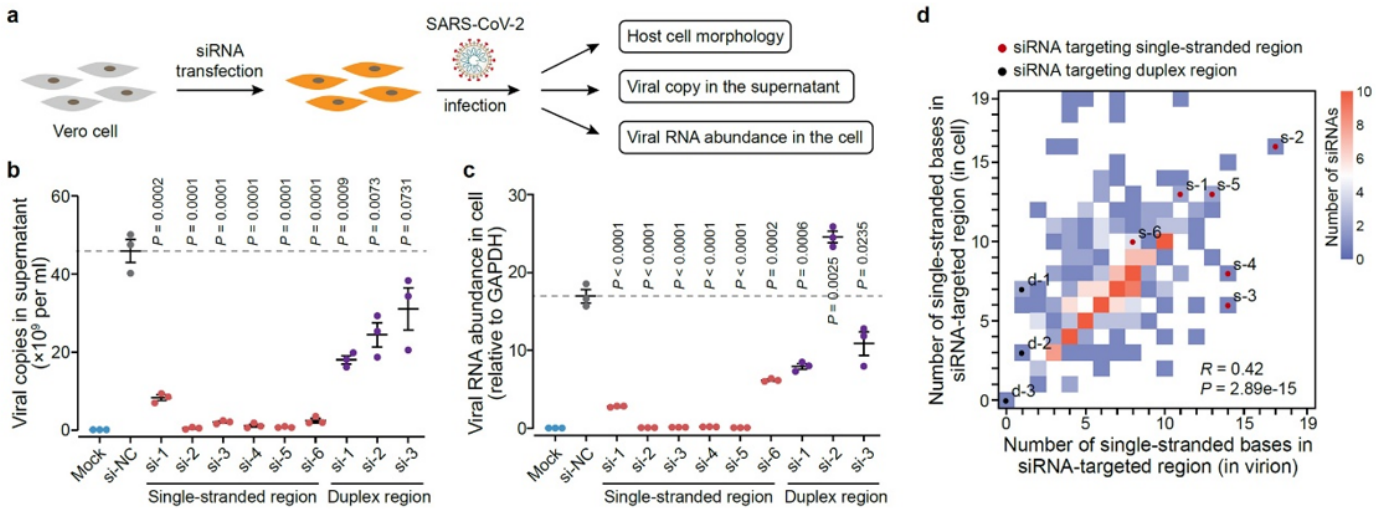

Figure 7

Structure-guided design of potent siRNAs as a cleavage agent to restrict SARS-CoV-2 infection. a, Diagram of strategy against SARS-CoV-2 infection in Vero cells. b, The SARS-CoV-2 copies in the supernatant were reduced to background level upon transfection with siRNAs targeting single-stranded regions (si-1 to si-6). Mock, uninfected cells; si-NC, non-targeting siRNA (control). c, qPCR showing the abundance of viral RNA in infected Vero cells. Data in b and c are mean ± s.e.m.; n = 3 biological replicates, two-tailed, unpaired Student's t-test. d, Number of single-stranded bases within the siRNA target regions identified in cells and virions. The colour intensity denotes the number of siRNAs. The six siRNAs targeting single-stranded regions and three siRNAs targeting duplex regions are labeled as s-1 to s-6 and d-1 to d-3, respectively. R, Pearson correlation coefficient. P-value was determined by the correlation test.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupportingInformation.docx