Discussion

In this work, we developed SEARCH-MaP and SEISMIC-RNA and applied them to detect structural ensembles involving long-range base pairs in SARS-CoV-2 and other coronaviruses. A previous study demonstrated that binding an ASO to one side of a long-range stem would perturb the chemical probing reactivities of the other side **?**. Here, we separated and identified the reactivities corresponding to long-range stems formed and unformed. This advance enables isolating the reactivities of the long-range stem formed – on not just one but both sides of the stem, linking corresponding alternative structures over distances much greater than the length of a read, which has not been possible in previous studies **??**. Using the linked reactivities from both sides of a long-range stem, its secondary structure can be modeled more accurately than would be possible using the ensemble average reactivities, as we have done for SARS-CoV-2 (Figure **??**) and TGEV (Figure **??**).

SEISMIC-RNA builds upon our previous work, the DREEM algorithm **?**. Here, we have optimized the algorithm to run approximately 10-30 times faster and built an entirely new workflow around it for aligning reads, calling mutations, masking data, and outputting a variety of graphs. SEISMIC-RNA can process data from any mutational profiling experiment, including DMS-MaPseq **?** and SHAPE-MaP **?**, not just SEARCH-MaP. The software is available from the Python Package Index (pypi.org/project/seismic-rna) or GitHub (github.com/rouskinlab/seismic-rna) and can be used as a command line executable program (`seismic`) or via its Python application programming interface (`import seismicrna`).

We envision SEARCH-MaP and SEISMIC-RNA bridging the gap between broad and detailed investigations of RNA structure. Other methods such as proximity ligation **?????** provide broad, transcriptome-wide information on RNA structure and could be used as a starting point to find structures of interest for deeper investigation with SEARCH-MaP/SEISMIC-RNA. Indeed, the first evidence of the FSE-arch in SARS-CoV-2 came from such a study **?**. To investigate RNA structures in detail, M2-seq **?** and related methods **?** can pinpoint base pairs with up to single-nucleotide

resolution and minimal need for structure prediction. However, base pairs are detectible only if the paired bases occur on the same sequencing read, which restricts their spans to at most the read length (typically 300 nt). Because the capabilities of M2-seq and SEARCH-MaP complement each other, they could be integrated: first SEARCH-MaP/SEISMIC-RNA to discover, quantify, and model long-range base pairs; then M2-seq for short-range base pairs. By providing the missing link – structure ensembles involving long-range base pairs – SEARCH-MaP and SEISMIC-RNA could combine broad and detailed views of RNA structure into one coherent model.

To understand structures of long RNA molecules, SEARCH-MaP and SEISMIC-RNA could also be used to validate predicted secondary structures and benchmark structure prediction algorithms. Algorithms that predict secondary structures achieve lower accuracies for longer sequences **??**, hence long-range base pairs in particular must be confirmed independently. We envision a workflow to determine the structure ensembles of an arbitrarily long RNA molecule that begins with DMS-MaPseq **?**. The DMS reactivities would be used **?** to predict two initial models of the structure: one with a limit to the base pair length (for short-range pairs), the other without (for long-range pairs). Sections of the RNA with potential long-range pairs would be flagged from the long-range model and from regions of the short-range model that disagreed with the DMS reactivities (as in Figure **??**d). Then, SEARCH-MaP/SEISMIC-RNA could be used to validate, quantify, and refine the potential long-range base pairs; and other methods such as M2-seq **?** to do likewise for short-range base pairs. This integrated workflow could characterize the secondary structures of RNA molecules that have evaded existing methods (e.g. messenger RNAs **?**) as well provide much-needed benchmarks for secondary structure prediction algorithms **?**.

In this study, we focused on the genomes of coronaviruses, specifically long-range base pairs involving the frameshift stimulating element (FSE). Long-range base pairs implicated in frameshifting also occur in several plant viruses of the family *Tombusviridae* **???**. However, in *Tombusviridae* species, the frameshift pseudoknots themselves are made of long-range base pairs; in coronaviruses, the pseudoknots are

local structures **????** and (at least in SARS-CoV-2) compete with long-range base pairs. Consequently, the long-range base pairs are necessary for frameshifting in *Tombusviridae* species **???** but dispensable in coronaviruses: even the 80-90 nt core FSE of SARS-CoV-2 has stimulated 15-40% of ribosomes to frameshift in dual luciferase constructs **??????**. Surprisingly, frameshifting has appeared to be nearly twice as frequent (50-70%) in live SARS-CoV-2 **???**; whether this discrepancy is due to long-range base-pairing, methodological artifacts, or *trans* factors **?** is unknown **?**.

If, how, and why the long-range base pairs affect frameshifting in coronaviruses are open questions. For *Tombusviridae*, one study **?** suggested that the long-range stem regulates viral RNA synthesis by negative feedback: without RNA polymerase, the long-range stem would form and stimulate frameshifting to produce polymerase, which would then unwind the long-range stem while replicating the genome. However, this mechanism seems implausible in coronaviruses, where RNA synthesis and translation occur in separate subcellular compartments (the double-membrane vesicles and the cytosol, respectively) **?**. Another study on *Tombusviridae* **?** hypothesized that after the ribosome has frameshifted, long-range stems destabilize the FSE so the ribosome can unwind it and continue translating. As the long-range base pairs in SARS-CoV-2 do compete with the pseudoknot, they might also have this role, which – for coronaviruses – could not be strictly necessary for frameshifting. One study **?** of translation in SARS-CoV-2 at different time points measured frameshifting around 20% at 4 hours post infection but 60-80% at 12-36 hours. This result is consistent with a previous hypothesis **?** that coronaviruses use frameshifting to time protein synthesis: first translating ORF1a to suppress the immune system, then translating ORF1b containing the RNA polymerase. We surmise the long-range base pairs would form in virions and persist when the virus released its genome into a host cell, where they would initially suppress frameshifting. Once host protein synthesis had been inhibited and the double-membrane vesicles formed, a signal specific to the cytosol would disassemble the long-range base pairs so that frameshifting could occur efficiently and produce the replication machinery from ORF1b. The long-range base pairs would form in viral progeny but not in genomic RNA released into

the cytosol for translation, so that more ORF1b could be translated. This possible role of long-range base pairs in the coronaviral life cycle could be tested by probing the RNA structure in subcellular compartments and virions, identifying cytosolic factors that could disassemble the long-range base pairs, and quantifying how they affect frameshifting in the context of a live coronavirus.

Future studies could also expand the scope of SEARCH-MaP and SEISMIC-RNA. While all SEARCH-MaP experiments in this study were performed *in vitro*, the method would likely also be feasible *in cellulo*: DMS-MaPseq can detect ASOs binding to RNAs within cells **?**. The main challenges would likely involve optimizing the ASO probes and transfection protocols to maximize the signal while minimizing unwanted side effects such as immunogenicity. SEARCH-MaP can screen an entire transcript (as in Figure **??**), but scaling up to an entire transcriptome could prove challenging. One strategy for probing many RNAs simultaneously could involve adding a pool of ASOs – with no more than one ASO capable of binding each RNA – rather than one ASO at a time. In this manner, a similar number of samples would be needed to search all RNAs as would be needed for the longest RNA. Distinguishing direct from indirect base pairing is another area for development: if segment Q could base-pair with either P or R, then blocking P could perturb R (and vice versa) as a consequence of perturbing Q, even though P and R could not base-pair directly. A solution could be to first block Q with one ASO; then, if blocking P with another ASO caused no change in R (and vice versa), it would suggest that they could only interact indirectly (through Q).

We imagine that SEARCH-MaP and SEISMIC-RNA will make it practical to determine accurate secondary structure ensembles of entire messenger, long noncoding, and viral RNAs. Collected in a database of long RNA structures, these results would facilitate subsequent efforts to predict RNA structures and benchmark algorithms, culminating in a real "AlphaFold for RNA" **?** in the hands of every biologist.