Genome-scale deconvolution of RNA structure ensembles

Edoardo Morandi^{1,2,5}, Ilaria Manfredonia^{1,5}, Lisa M. Simon^{2,5}, Francesca Anselmi²,
Martijn J. van Hemert³, Salvatore Oliviero^{2,4} and Danny Incarnato¹

RNA structure heterogeneity is a major challenge when querying RNA structures with chemical probing. We introduce DRACO, an algorithm for the deconvolution of coexisting RNA conformations from mutational profiling experiments. Analysis of the SARS-CoV-2 genome using dimethyl sulfate mutational profiling with sequencing (DMS-MaPseq) and DRACO, identifies multiple regions that fold into two mutually exclusive conformations, including a conserved structural switch in the 3' untranslated region. This work may open the way to dissecting the heterogeneity of the RNA structure.

RNA structural analysis by means of chemical probing is powerful but it suffers from the intrinsic limitation of providing only an averaged measurement of the base reactivities of all coexisting conformations that are simultaneously sampled by an RNA species in a biological sample^{1,2}. The development of mutational profiling has enabled multiple single-stranded residues to be recorded as mutations within the same complementary DNA product^{3,4}. Recently, an expectation–maximization approach for the deconvolution of alternative conformations from mutational profiling experiments (DREEM, detection of RNA folding ensembles using expectation–maximization), has been developed⁵. Even if powerful in principle, this method has two major limitations: (1) the maximum number of RNA conformations that can be searched for is user defined (two by default), to reduce the risk of overestimating the number of conformations (overclustering, a common problem with expectation–maximization approaches); and (2) it can handle only experiments in which the sequencing reads cover the entire length of the target RNA. Although DREEM can, in theory, be applied to longer transcripts by manual window sliding, it cannot handle the automatic merging of overlapping RNA segments, a non-trivial computational problem that makes it poorly suited for transcriptome-scale analyses.

To address these issues, we introduce DRACO (v1.0), a method for the deconvolution of alternative RNA conformations from mutational profiling experiments, based on a combination of spectral clustering and fuzzy clustering (Supplementary Note 1). DRACO analysis, as illustrated in Extended Data Fig. 1, is performed (by default) using sliding windows with a size equal to 90% of the median length of reads, and an offset of 5%. Spectral clustering, previously proposed as a suitable approach for the identification of structurally heterogeneous regions from mutational profiling experiments⁶, is performed for each window, which enables the automatic identification of the optimal number of conformations (clusters). DRACO then merges overlapping windows that have the same number of clusters and reconstructs overall mutational profiles. Consecutive sets of windows that have

different numbers of clusters are reported separately by DRACO. To validate the algorithm, we generated in silico dimethyl sulfate (DMS) mutational profiling reads of different lengths (50–150 nucleotides), for 10 sets of 100 RNAs of increasing length (300–1,500 nucleotides), designed to form up to four distinct conformations. DMS-induced mutations in reads were modeled as a binomial distribution, well-approximating the observed distribution of a previously published dataset⁶ (Supplementary Figs. 1,2). Analyses of in silico data (Supplementary Figs. 3–14) showed that DRACO accuracy relies on two main factors, read length and depth of coverage, which can be explained by DRACO's dependency on co-mutation information. Although greater depths of coverage partially compensated for the reduced amount of mutational information in shorter reads, best results were obtained with a read length of 150 nucleotides and a minimum depth of coverage of 5,000×. Under these conditions, DRACO correctly identified the expected number of conformations in nearly 100% of cases (Fig. 1a and Extended Data Fig. 2), accurately deconvoluted the individual conformation mutational profiles (median Pearson correlation coefficient, PCC>0.85; Fig. 1b) and precisely estimated relative conformation stoichiometries (PCC≈0.99; Fig. 1c).

Given that in silico-generated data might not completely capture the complexity of real DMS mutational profiling with sequencing (DMS-MaPseq) experiments, we then tested DRACO using published in vitro data for the *Escherichia coli* *cspA* 5' untranslated region (UTR)⁷. The *cspA* 5' UTR acts as an RNA thermometer, regulating the accessibility of the Shine–Dalgarno sequence in response to temperature changes, switching between a translationally repressed conformation at 37°C and a translationally competent conformation at 10°C (ref. ⁸). After mapping DMS-MaPseq data from in vitro folding experiments at 10°C and 37°C, reads from the two experiments were pooled at different percentages and analyzed using DRACO (Extended Data Fig. 3). DRACO successfully reconstructed the expected reactivity profiles with high accuracy, even with a conformation abundance of only 10% (PCC=0.88). Interestingly, the *cspA* protein can act as an RNA chaperone on its own 5' UTR, mediating the switch from the 10°C translationally competent conformation to the 37°C translationally repressed conformation. In the same study⁷, the *cspA* 5' UTR was folded and probed at 10°C, in the presence of increasing concentrations of the *cspA* protein. In the presence of 0.1 mM *cspA* the conformation of the 5' UTR resembled that observed at 37°C (ref. ⁷). The use of 50 μM *cspA* protein resulted in a reactivity profile that only partially correlated with either the 10°C or 37°C conformations. Strikingly, analysis with DRACO identified two nearly equimolar conformations (51.4% and 48.6%, respectively; Fig. 2a),

¹Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Groningen, the Netherlands. ²Dipartimento di Scienze della Vita e Biologia dei Sistemi, Università di Torino, Torino, Italy. ³Department of Medical Microbiology, Molecular Virology Laboratory, Leiden University Medical Center, Leiden, the Netherlands. ⁴Italian Institute for Genomic Medicine (IIGM), Candiolo, Italy. ⁵These authors contributed equally: Edoardo Morandi, Ilaria Manfredonia, Lisa M. Simon. ✉e-mail: salvatore.oliviero@unito.it; d.incarnato@rug.nl

Fig. 1 | In silico validation of DRACO. **a**, Maximum number of conformations detected for 10 sets of 100 simulated 300-nt-long RNAs expected to form 1–4 conformations, at a depth of coverage of 5,000 $\times$ and a read length of 150 nt. (Extended Data Fig. 2 shows the maximum number of conformations detected for RNAs with length ranging from 600 to 1,500 nt.) The individual data points represent the mean of each set. **b**, Box plot of median PCC of reconstructed reactivity profiles for the 10 sets. When DRACO detected more than one window with different numbers of clusters, only the largest window, spanning >50% of the RNA length, was used. Boxes span the 25th–75th percentiles. The horizontal center line represents the median. Whiskers span from the 25th percentile – 1.5-fold the interquartile range (IQR) to the 75th percentile + 1.5-fold the IQR. Data points falling outside this range represent outliers and are reported as dots. **c**, Violin plot of the distribution of expected versus reconstructed conformation abundances for 10 sets of 100 simulated 300-nt-long RNAs, expected to form 2 conformations with varying relative abundances. When multiple windows were detected, only the largest window was used. Whiskers span the 25th–75th percentiles. The central dot represents the median. nt, nucleotide; R , Pearson correlation.

the profiles of which were highly correlated with the 37°C and the 10°C conformations (respectively, $PCC=0.85$ and 0.83 ; Fig. 2b). Data-driven RNA structure prediction using these profiles produced secondary structure models nearly identical to the reference structures expected for the 10°C and 37°C conformations (positive predictive value, PPV, 1.00 and 0.91; sensitivity, 0.87 and 0.97, respectively; Fig. 2c). We further analyzed a recently published DMS–MaPseq dataset, originally generated to validate the DREEM algorithm⁵, by probing the structure of the adenosine deaminase riboswitch encoded by the *add* gene from *Vibrio vulnificus*, either in the absence or presence of 5 mM adenine. Although DREEM identified three conformations under both conditions⁵, DRACO showed that a single conformation is present in the absence of adenine, and that the addition of adenine triggers the conformation switch towards the translation-competent conformation on ~65.6% of the RNA molecules (Fig. 2d). The remaining ~34.4% represent instead the translation-incompetent conformation, as demonstrated by the high correlation to the adenine-free sample ($PCC=0.96$, Fig. 2e), as well as by the agreement

between the predicted and the expected secondary structures of the two conformations (Fig. 2f). These results support the robustness of the DRACO algorithm, as well as its lower propensity to overclustering, as compared with expectation–maximization-based approaches (Supplementary Note 2). Encouraged by these results, we applied DRACO to the analysis of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) RNA genome structure. In a recent report, we defined the secondary structure of the full SARS-CoV-2 genome using SHAPE–MaP (selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling)⁹. Although powerful, that approach was limited to the analysis of regions folding into a single well-defined conformation. We therefore queried (in two independent experiments) the full in vitro refolded SARS-CoV-2 genome using DMS–MaPseq analysis. Sequencing and assembly of 150-nucleotide-long paired-end reads produced more than 2.2×10^7 mappable fragments, resulting in a median coverage of $\sim 9.9 \times 10^4$ per experiment (Supplementary Fig. 15a,b). The data showed exceptional between-experiment correlation ($PCC=0.99$, Supplementary Fig. 15c), and were in agreement with conserved SARS-CoV-2 structures⁹ (Supplementary Fig. 16). In each of the two experiments, analysis with DRACO unambiguously identified 22 windows, which accounted for ~15.5% of the SARS-CoV-2 genome, that folded consistently into two conformations (Supplementary Fig. 17a). We observed an exceptional overall correlation of reactivity profiles for reconstructed conformations across experiments ($PCC=0.86$; Supplementary Fig. 17b), as well as highly consistent relative conformation abundance estimates (Supplementary Fig. 17c), with an average variation of only $\pm 1.9\%$. On inspection of the distribution of these windows, we noticed an enrichment of 50% at open reading frame (ORF) and at peptide boundaries (11 out of 22 windows) as compared with the ~19% expected by chance ($P=1.0 \times 10^{-3}$, one-sided binomial test; see Methods). One of these windows spanned the ORF1a–ORF1b boundary, overlapping with the frameshifting element (FSE, positions 13369–13542; Supplementary Fig. 18). Importantly, the data do not support the existence of a pseudoknotted structure at the level of the FSE. Instead, this region is likely to fold into either a single extended stem-loop structure or two stem-loop structures. This observation is further supported by a recently proposed structure analysis, using DMS–MaPseq, of the SARS-CoV-2 genome in living infected host cells¹⁰. It is conceivable that the identified RNA switches might be involved in controlling either the translation of SARS-CoV-2 proteins, or the discontinuous transcription of subgenomic messenger RNAs (or both), but additional experiments will be needed to investigate their functional relevance. One of the identified windows encompassed the 3′ UTR (positions 29546–29767), and showed consistent abundance estimates and reactivity profiles for the two identified conformations across the two experiments (Fig. 3a). The major conformation ($63.4 \pm 1.7\%$) had a reactivity pattern compatible with the known phylogenetically inferred 3′ UTR structure of sarbecoviruses, while the minor conformation ($36.6 \pm 1.7\%$) was predicted to form an alternative three-way junction, sequestering both the bulged stem-loop (BSL) and distal stem-loop (designated P2) helices (Fig. 3b). We further evaluated the conservation of this alternative conformation using an approach we recently developed to automatically identify regions of the SARS-CoV-2 genome that show significant covariation⁹ (Methods). Only those coronavirus sequences supporting both 3′ UTR conformations were retained in order to evaluate the covariation (Methods), and significant covariation was then identified in this alternative 3′ UTR three-way junction conformation (Fig. 3c), hinting at its functional relevance. When the same analysis was performed on the two conformations independently, additional significantly covarying base pairs were detected (Supplementary Fig. 19). Re-analysis of a recently published dataset of RNA–RNA interaction capture in SARS-CoV-2-infected cells¹¹ with COMRADES (crosslinking of matched RNAs and deep sequencing)¹² provided support for the presence of both conformations in vivo.

Fig. 2 | In vitro validation of DRACO. **a**, Original DMS-MaPseq profile and DRACO-deconvoluted profiles for the *cspA* 5' UTR folded at 10 °C in the presence of 50 μ M *cspA* recombinant protein, from Zhang et al., 2018⁷. Schematic representation of the structures is reported together with the estimated relative abundances. **b**, Heatmap of PCCs showing the correlation between the conformations deconvoluted by DRACO and the reactivity profiles of the *cspA* 5' UTR folded at either 10 °C or 37 °C, in the absence of the recombinant *cspA* protein. **c**, Arc plots depicting the secondary structure inferred from the DRACO-deconvoluted profiles, as compared to the reference *cspA* 5' UTR structures at 10 °C and 37 °C. **d**, DRACO-deconvoluted profiles for the *V. vulnificus add* riboswitch, in the absence (one conformation detected) or presence (two conformations detected) of 5 mM adenine, from Tomezko et al., 2020⁵. Schematic representation of the structures, and the estimated relative abundances are reported. **e**, Heatmap of PCCs showing the correlation between the conformations deconvoluted by DRACO. **f**, Arc plots depicting the secondary structure inferred from the DRACO-deconvoluted profiles, as compared to the reference *add* structure in the absence of adenine.

Altogether, these data demonstrate the ability of DRACO to capture otherwise hidden structural features, and reveal the presence of a conserved RNA switch at the level of a regulatory region in the SARS-CoV-2 genome.

We anticipate that DRACO will enable exploration of the RNA structurome at unprecedented resolution, and the identification of transient and dynamic features of cellular transcriptomes.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41592-021-01075-w>.

Fig. 3 | A conserved structural switch in the 3' UTR of SARS-CoV-2. a, Heat scatter plot of base reactivities for DRACO-deconvoluted reactivity profiles, in experiment 1 versus experiment 2, for conformations A and B. **b**, Secondary structure models with overlaid base reactivities for conformations A and B. Base pairs whose existence is supported by significant enrichment of RNA-RNA chimeras from in vivo COMRADES analysis (Ziv et al., 2020¹¹) are boxed in light blue. **c**, Structure models for conformations A and B, inferred by simultaneous phylogenetic analysis. Structures have been generated using R2R v1.0.6. Base pairs showing significant covariation (as determined by R-scape v1.4.0) are boxed in green ($E < 0.05$) or violet ($E < 0.1$). HVR, hypervariable region; s2m, stem-loop II-like motif.

Received: 17 September 2020; Accepted: 19 January 2021;
Published online: 22 February 2021

References

- Incarnato, D. & Oliviero, S. The RNA epistruome: uncovering RNA function by studying structure and post-transcriptional modifications. *Trends Biotechnol.* **35**, 318–333 (2017).
- Strobel, E. J., Yu, A. M. & Lucks, J. B. High-throughput determination of RNA structures. *Nat. Rev. Genet.* **19**, 615–634 (2018).
- Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11**, 959–965 (2014).
- Zubradt, M. et al. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods* **14**, 75–82 (2017).
- Tomezsko, P. J. et al. Determination of RNA structural diversity and its role in HIV-1 RNA splicing. *Nature* **582**, 438–442 (2020).
- Homan, P. J. et al. Single-molecule correlated chemical probing of RNA. *Proc. Natl Acad. Sci. USA* **111**, 13858–13863 (2014).
- Zhang, Y. et al. A stress response that monitors and regulates mRNA structure is central to cold shock adaptation. *Mol. Cell* **70**, 274–286.e7 (2018).
- Giuliodori, A. M. et al. The cspA mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. *Mol. Cell* **37**, 21–33 (2010).
- Manfredonia, I. et al. Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res.* **48**, 12436–12452 (2020).
- Lan, T. C. T. et al. Structure of the full SARS-CoV-2 RNA genome in infected cells. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.06.29.178343> (2020).
- Ziv, O. et al. The short- and long-range RNA-RNA interactome of SARS-CoV-2. *Mol. Cell* **80**, 1067–1077.e5 (2020).
- Ziv, O. et al. COMRADES determines in vivo RNA structures and interactions. *Nat. Methods* **15**, 785–788 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

DRACO algorithm. The DRACO algorithm is implemented in C++ and exploits the Armadillo library (<http://arma.sourceforge.net>), built on top of the BLAS (Basic Linear Algebra Subprograms, <http://www.netlib.org/blas/>) and LAPACK (Linear Algebra Package, <http://www.netlib.org/lapack/>) libraries for fast matrix manipulation and eigenvalue decomposition. As input, DRACO takes Mutation Map (MM) format files. These files store the relative coordinates of mutations for each read mapping on a given transcript and can be generated by processing a Sequence Alignment/Map (SAM, or its compressed, binary version, BAM) alignment file with the rf-count tool of RNA Framework v2.7.5 (ref. ¹³) (parameter: -mm). With default parameters, DRACO takes ~8–10 h, on a single thread, to analyze ~17 million reads mapping to the SARS-CoV-2 genome. This runtime should not be taken as representative of DRACO performances, because it is biased by the non-uniform distribution of reads across the SARS-CoV-2 genome due to a large fraction of the reads mapping at the 5' and 3' ends, which extends the computation time when analyzing genome boundaries. A complete description of the algorithm, including pseudocodes, is provided in Supplementary Note 1.

In silico generation of DMS-MaPseq data. One thousand RNA sequences with an average A or C content of 50% and varying lengths (300, 600, 900 or 1,500 nucleotides) were randomly generated. DMS modification profiles for one to four different conformations were then generated by randomly setting A or C residues represents an underestimate of what is expected for real RNAs (~51.3% of single-stranded A or C residues for *E. coli* 16S and 23S ribosomal RNAs). Mutated reads matching these modification profiles were then generated (in MM format) to obtain a median depth of coverage per base of 2,000×, 5,000×, 10,000× or 20,000×, using the generate_mm tool (available from the DRACO repository). Distribution of DMS-induced mutations in reads was empirically learnt from a previously published dataset⁶ (Supplementary Fig. 1) and well-approximated by a binomial distribution, with $P = 0.01927$ and $n = \text{length of the transcript}$ (Supplementary Fig. 2).

Analysis of in silico-generated DMS-MaPseq data. In silico-generated MM files were analyzed using DRACO (parameters: --set-all-uninformative-to-one --set-uninformative-clusters-to-surrounding --max-collapsing-windows <variable> --first-eigengap-threshold 0.9). Owing to the fact that A and C residues are non-uniformly distributed along transcripts, certain regions of the RNA can give rise to reads bearing a lower mutational information content, possibly leading to a local under- (or over-) estimate of the number of conformations. To account for this, DRACO can ignore a small set of windows (the number of which is controlled by the --max-collapsing-windows parameter) showing a discordant number of conformations with respect to surrounding windows. Given that the window size is determined by the read length (by default, 90% of the median read length), the number of discordant windows is expected to increase with decreasing read lengths. Therefore, the --max-collapsing-window parameter was linearly decreased from 5 to 2 with increasing read lengths from 50 to 150 nucleotides. By default, windows slide by 5% of the median read length, hence these --max-collapsing-window values imply that only 12.5 (for 50 nucleotide reads) to 15 bases (for 150 nucleotide reads) are ignored in such situations.

Analysis of DMS-MaPseq data. All of the relevant analysis steps, from reads alignment to data normalization and structure modeling, were performed using RNA Framework¹³. All tools referenced in the following paragraphs are distributed as part of the RNA Framework suite (<https://github.com/dincarnato/RNAFramework>). Specific analysis parameters are detailed in the respective paragraphs.

Optimization of folding parameters. For structure predictions, optimal slope (2.4) and intercept (−0.2) values were identified using jackknifing, with a DMS-MaPseq dataset for ex vivo deproteinized *E. coli* rRNAs that we previously published¹⁴ (accession, SRR8172706), the rf-jackknife tool (parameters: -rp -md 600 -nlp -x) and ViennaRNA v2.4.14 (ref. ¹⁵).

Analysis of *cspA* 5' UTR DMS-MaPseq data. Reads for DMS-MaPseq data of in vitro folded *cspA* 5' UTR at 37°C and 10°C were obtained from the Sequence Read Archive (accessions, SRR6123773 and SRR6123774) and mapped to the first 171 bases of the *cspA* transcript using the rf-map tool (parameters: -cq5 20 -cgo -mp --very-sensitive-local). Given that a lower fraction of reads aligned to the reference for the experiment conducted at 37°C, we chose to randomly shuffle the BAM file from the experiment conducted at 10°C, and a matching number of reads was extracted. Resulting BAM files for both samples were then randomly shuffled, and reads were extracted and combined to achieve final percentage stoichiometries of 90:10, 80:20, 70:30, 60:40 and 50:50 of the 10°C and 37°C conformations, respectively. Resulting BAM files were then analyzed with the rf-count tool to produce MM files (parameters: -m -mm -ds 75 -na -ni -md 3). MM files were analyzed with DRACO (parameters: --max-collapsing-windows 3 --set-all-uninformative-to-one --min-cluster-fraction 0.1 --set-uninformative-clusters-to-surrounding) and deconvoluted mutation

profiles were extracted from the resulting JSON (JavaScript Object Notation) files and converted into RNA Framework's RNA Count (RC) format. Starting from RC files, normalized reactivity profiles were obtained by first calculating the raw reactivity scores as the per-base ratio of the mutation count and the read coverage at each position, and then by normalizing values by box-plot normalization, using the rf-norm tool (parameters: -sm 4 -nm 3 -rb AC -mm 1 -n 1000). Data-driven RNA structure inference was performed using the rf-fold tool and the normalized reactivity profiles (parameters: -sl 2.4 -in -0.2 -nlp). DMS-MaPseq data for the *cspA* 5' UTR folded in the presence of 0.05 μM cspA protein (accession, SRR6507969) were analyzed using the same parameters. Comparison between the deconvoluted conformations and the *cspA* 5' UTR folded at either 10°C or 37°C was performed using the rf-compare tool.

Analysis of *V. vulnificus* add riboswitch DMS-MaPseq data. DMS-MaPseq data for the *add* riboswitch from *V. vulnificus*, in vitro folded either in the presence or absence of 5 mM adenine, were obtained from the Sequence Read Archive (accessions, SRR10850890 and SRR10850891). Forward and reverse reads were merged, prior to mapping, using PEAR (Illumina paired-end read merger) v0.9.11 (ref. ¹⁶), and then mapped to the *add* riboswitch using the rf-map tool (parameters: -cq5 20 -cgo -ctn -cmn 0 -mp --very-sensitive-local). Resulting BAM files were then analyzed with the rf-count tool to produce MM files (parameters: -m -mm -na -ni). MM files were analyzed with DRACO (parameters: --max-collapsing-windows 1 --set-all-uninformative-to-one --set-uninformative-clusters-to-surrounding), and deconvoluted mutation profiles were extracted from the resulting JSON files and converted into RC format. Starting from RC files, normalized reactivity profiles were obtained by first calculating the raw reactivity scores as the per-base ratio of the mutation count and the read coverage at each position, and then by normalizing values by box-plot normalization, using the rf-norm tool (parameters: -sm 4 -nm 3 -rb AC -mm 1 -n 1000). Data-driven RNA structure inference was performed using the rf-fold tool and the normalized reactivity profiles (parameters: -sl 2.4 -in -0.2 -nlp).

Cell culture and SARS-CoV-2 infection. Vero E6 cells were cultured in DMEM (Lonza, 12-604F), supplemented with 8% FCS (Bodinco), 2 mM L-glutamine, 100 U ml^{−1} penicillin and 100 μg ml^{−1} streptomycin (Sigma Aldrich, P4333-20ML) at 37°C, in the presence of 5% CO₂. Cells were passaged twice and then infected at a multiplicity of infection of 1.5 with SARS-CoV-2/Leiden-0002 (GenBank accession, MT510999). Infections were performed in EMEM (Lonza, 12-611F) supplemented with 25 mM HEPES, 2% FCS, 2 mM L-glutamine and antibiotics. Sixteen hours after infection, cells were trypsinized, resuspended in EMEM supplemented with 2% FCS, and then pelleted and washed with 50 ml 1X PBS.

Total RNA extraction and in vitro folding. Approximately 5 × 10⁶ SARS-CoV-2-infected cells were resuspended in 1 ml TriPure Isolation Reagent (Sigma Aldrich, 11667157001). After adding 0.2 volumes chloroform and vigorously vortexing for 15 s, the sample was centrifuged for 15 min at 12,500×g (4°C). The upper aqueous phase was mixed with 1 ml 100% ethanol, and then loaded on an RNA Clean & Concentrator-25 column (Zymo Research, R1017). In vitro folding was carried out as previously described^{21,14}. In brief, ~7.5 μg RNA was subjected to rRNA depletion using the RiboMinus Eukaryote System v2 (ThermoFisher Scientific, A15026). The rRNA-depleted RNA was denatured at 95°C for 2 min, then transferred immediately to ice for 1 min. Ice-cold 5X RNA folding buffer (500 mM HEPES pH 7.9; 500 mM NaCl) supplemented with 20 U SUPERase•In RNase Inhibitor (ThermoFisher Scientific, AM2696) was then added, and RNA was incubated for 10 min at 37°C to enable secondary structure formation. MgCl₂ was then added to a final concentration of 10 mM and the sample was further incubated for 20 min at 37°C.

Probing of SARS-CoV-2 RNA. For probing of RNA, DMS was pre-diluted in a ratio of 1:6 in 100% ethanol and added to a final concentration of 150 mM. Samples were then incubated at 37°C for 2 min. Reactions were quenched by the addition of 1 volume dithiothreitol 1.4 M and then purified on an RNA Clean & Concentrator-5 column (Zymo Research, R1013).

DMS-MaPseq analysis of SARS-CoV-2 RNA. DMS-MaPseq of SARS-CoV-2 was conducted as previously described¹⁴, with minor changes. First, probed RNA was fragmented to a median size of 150 nucleotides by incubation at 94°C for 8 min in RNA fragmentation buffer (65 mM Tris-HCl pH 8.0, 95 mM KCl, 4 mM MgCl₂), then purified with NucleoMag NGS Clean-up and Size Select beads (Macherey Nagel, 744970), supplemented with 10 U SUPERase•In RNase Inhibitor and eluted in 8 μl NF H₂O. Eluted RNA was supplemented with 1 μl 50 μM random hexamers and 2 μl deoxyribonucleoside triphosphates (10 mM each), then incubated at 70°C for 5 min and immediately transferred to ice for 1 min. Reverse transcription reactions were conducted in a final volume of 20 μl. Reactions were supplemented with 4 μl 5X RT buffer (250 mM Tris-HCl pH 8.3, 375 mM KCl, 15 mM MgCl₂), 1 μl dithiothreitol 0.1 M, 20 U SUPERase•In RNase Inhibitor and 200 U TGIRT-III Enzyme (InGex, TGIRT50). Reactions were incubated at 25°C for 10 min to enable partial primer extension, followed by 2 h at 57°C. TGIRT-III was degraded by the addition of 2 μg Proteinase K (Sigma Aldrich, P2308), followed by incubation at

37°C for 20 min. Proteinase K was inactivated by addition of Protease Inhibitor Cocktail (Sigma Aldrich, P8340). Reverse transcription reactions were then used as input for the NEBNext Ultra II Non-Directional RNA Second Strand Synthesis Module (New England Biolabs, E6111L). Second strand synthesis was performed by incubation for 1 h at 16°C, as per the manufacturer's instructions. Double-stranded DNA was purified using NucleoMag NGS Clean-up and Size Select beads and used as input for the NEBNext Ultra II DNA Library Prep Kit for Illumina, following the manufacturer's instructions.

Analysis of SARS-CoV-2 DMS-MaPseq data. Following sequencing, samples were demultiplexed using the bcl2fastq v2.20.0.422 utility. After clipping adapter sequences using Cutadapt v2.1 (ref. ¹⁷) (parameters: -a AGATCGGAAGAGC -A AGATCGGAAGAGC -O 1 -m 100:100), paired-end reads were merged using PEAR v0.9.11 (ref. ¹⁶) and then mapped to the SARS-CoV-2 reference using the rf-map tool and the Bowtie2 algorithm (v2.3.5.1), with soft-clipping enabled (parameters: -b2 -cq5 20 -ctn -cmn 0 -cl 150 -mp '--very-sensitive-local'). Alignments in SAM format were sorted and converted to BAM format using Samtools v1.9. An MM file was then generated from the resulting BAM alignment using the rf-count tool, by keeping only those reads that covered at least 150 bases. Insertions and deletions were ignored (because they account for <6% of DMS-induced mutations when using TGIRT-III¹⁸), and only mutations with a Phred quality score > 20 were considered. Furthermore, mutations were considered only when the two surrounding bases also had a Phred quality score > 20. Reads with more than 10% mutated bases were excluded (parameters: -m -ds 150 -es -nd -ni -mm -me 0.1). DRACO was invoked with default parameters. Following DRACO analysis, windows in which the median coverage (calculated on reads passing DRACO's filtering) was above 10,000× were selected. To select windows consistently folding into multiple conformations in both experiments, we retained windows predicted to have the same number of conformations in the two experiments, and which overlapped by at least 75% of their length, and considered only their intersection. Deconvoluted reactivity profiles for matching conformations from the two experiments were then averaged and used for secondary structure modeling. The correlation between reconstructed conformations from the two experiments was calculated using 90% of the reactivity values in the window, after exclusion of the first and last 5% of the A and C bases, to avoid terminal bias. RNA secondary structures for SARS-CoV-2 elements (FSE and 3' UTR) were generated using VARNA v3.93 (ref. ¹⁸).

Identification of conserved RNA structure elements. To evaluate the conservation of the alternative 3' UTR structure, we implemented a modified version of an automated pipeline that we have previously introduced⁹ (cm-builder; <https://github.com/dincarnato/labtools>), built on top of Infernal v1.1.3 (ref. ¹⁹). In brief, we first built two covariance models from Stockholm files containing only the SARS-CoV-2 sequence and the two alternative 3' UTR structures, using the cmbuild module. After calibrating the covariance models using the cmcalibrate module, we then used them to search for RNA homologs in a database composed of all of the non-redundant coronavirus complete genome sequences from the ViPR database²⁰ (<https://www.viprbrc.org/brc/home.spg?decorator=corona>), as well as a set of representative coronavirus genomes from the NCBI database, using the cmsearch module. Only matches from the sense strand were kept and a very relaxed *E* value threshold of 10 was used at this stage to select potential homologs. Three additional filtering criteria were used. First, we took advantage of the extremely conserved architecture of coronavirus genomes²¹ and restricted the selection to matches falling at the same relative position within their genome, with a tolerance of 3.5% (corresponding approximately to a maximum allowed shift of 1,050 nucleotides in a 30 kilobase genome). Through this more conservative selection, we kept only matches likely to represent true structural homologs, although at the cost of probably losing some true matches. Second, we filtered out matches retaining less than 55% of the canonical base pairs from the original structure elements. Third, truncated hits covering less than 50% of the structure were discarded. A fourth filtering step was also applied when simultaneously analyzing the two structures, by retaining only the set of sequences matched by both structures. The resulting set of homologs was then aligned to the original covariance models using the cmalign module and the resulting alignments were used to build new covariance models. The whole process was repeated three times. The alignment was then refactored, removing gap-only positions and including only bases spanning the first to the last base-paired residue. The alignment file was then analyzed using R-scape 1.4.0 (ref. ²²) and average product corrected G-test statistics to identify motifs showing significantly covarying base pairs.

Testing for significant overlap with ORF boundaries. To test for significant overlap between the windows folding into two mutually exclusive conformations and the ORF boundaries within the SARS-CoV-2 genome, we generated 10,000 random windows of matching size for each window identified by DRACO. For each DRACO-identified window, as well as for each random window, we calculated the number of windows overlapping the start and end positions of the SARS-CoV-2 ORFs, including each of the individual proteins within the polyprotein ORF1ab (positions 266, 806, 2720, 8555, 10055, 10973, 11843, 12092, 12686, 13025, 13442, 13468, 16237, 18040, 19621, 20659, 21563, 25393, 26245,

26523, 27202, 27394, 27756, 27894, 28274, 29558, 29674). Resulting values were used to perform a one-sided binomial test, with parameters $k = 11$ (number of windows identified by DRACO, overlapping with ORF boundaries), $n = 22$ (total number of windows identified by DRACO) and p , the ratio of the number of random windows overlapping with ORF boundaries, divided by the total number of random windows (220,000).

Validation of the alternative SARS-CoV-2 3' UTR conformation by COMRADES. COMRADES data for the SARS-CoV-2 virus in living infected host cells¹¹ were obtained from the Gene Expression Omnibus (GEO, [GSE154662](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154662)). The dataset consisted of two biological replicates, each one composed of a control (C) and the actual COMRADES sample (S). A reference was built to include all human transcripts from refGene, plus the sequence of the SARS-CoV-2 genome, using STAR (Spliced Transcripts Alignment to a Reference) v2.7.1a (ref. ²³) (parameters: --runMode genomeGenerate --genomeSAindexNbases 12), and reads were also aligned to the reference using STAR (parameters: --runMode alignReads --outFilterMultimapNmax 100 --outSAMattributes All --alignIntronMin 1 --scoreGapNoncan -4 --scoreGapATAC -4 --chimSegmentMin 15 --chimJunctionOverhangMin 15). Resulting alignments (as well as chiasmic alignments from the junctions file) were filtered, discarding ungapped reads, reads with more than one gap, and reads aligning to the human transcriptome, and the total number of reads per experiment was calculated (C_{tot} and S_{tot}). Each chimeric read was described as a set of two numeric intervals (I1 and I2), corresponding to the two halves of the chimera. To assess whether a base pair $i-j$ was enriched in the COMRADES sample with respect to the control sample, we calculated the number of reads in which base i overlapped interval I1 and base j overlapped interval I2, for both samples (C_{i-j} and S_{i-j}). Significance of the enrichment was then assessed using a one-tailed binomial test, with parameters $k = S_{i-j}$, $n = S_{\text{tot}}$ and $p = C_{i-j} / C_{\text{tot}}$. Only base pairs with $P < 0.05$ in both replicates were considered to have in vivo support.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequencing data have been deposited to the Gene Expression Omnibus (GEO) database under the accession [GSE158052](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158052). Additional processed files are available at http://www.incarnatolab.com/datasets/DRACO_Morandi_2021.php. Source data are provided with this paper.

Code availability

The source code of DRACO is freely available from GitHub under the GPLv3 license (<https://github.com/dincarnato/draco>). A complete list of the software used for data analysis is available from the Nature Research Reporting Summary.

References

13. Incarnato, D., Morandi, E., Simon, L. M. & Oliviero, S. RNA Framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications. *Nucleic Acids Res.* **46**, e97 (2018).
14. Simon, L. M. et al. In vivo analysis of influenza A mRNA secondary structures identifies critical regulatory motifs. *Nucleic Acids Res.* **47**, 7003–7017 (2019).
15. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
16. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End read merger. *Bioinformatics* **30**, 614–620 (2014).
17. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
18. Darty, K., Denise, A. & Ponty, Y. VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics* **25**, 1974–1975 (2009).
19. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
20. Pickett, B. E. et al. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40**, D593–D598 (2012).
21. Lauber, C. et al. The footprint of genome architecture in the largest genome expansion in RNA viruses. *PLoS Pathog.* **9**, e1003500 (2013).
22. Rivas, E., Clements, J. & Eddy, S. R. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nat. Methods* **14**, 45–48 (2017).
23. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

Acknowledgements

D.I. was supported by the Dutch Research Council (Netherlands Organisation for Scientific Research, NWO) as part of the research programme NWO Open Competitie ENW-XS (project number OCENW.XS3.044), and by the Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen. S.O. was supported by the Associazione Italiana per la Ricerca sul Cancro (AIRC), grant AIRC IG 2017 Id. 20240 and PRIN 2017. M.J.H. was supported by the Leiden University Fund

(LUF), the Bontius Foundation, and donations from the crowdfunding initiative 'wake up to corona'.

Author contributions

E.M. and D.I. conceived the project; I.M., L.M.S. and F.A. carried out the wet-lab work; M.J.H. carried out SARS-CoV-2 manipulations; E.M. and D.I. designed and implemented the DRACO algorithm; E.M. and D.I. carried out bioinformatics, structure modeling and data analysis; D.I. and S.O. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41592-021-01075-w>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-021-01075-w>.

Correspondence and requests for materials should be addressed to S.O. or D.I.

Peer review information *Nature Methods* thanks Walter N. Moss and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Lei Tang was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

Extended Data Fig. 1 | Overview of the DRACO algorithm. By default, a window of a size equal to 90% of the median read is slid along the transcript, in 5% increments. For each window, a mutation map is generated using only the reads covering the entire window. Bases that are mutated with respect to the reference are assigned a value of 1, while not mutated bases are assigned a value of 0. By using this map, a graph is built, in which each vertex is a base of the transcript, and edges connecting two vertices are weighted proportionally to the number of reads in which the two connected bases have been observed to co-mutate. Starting from the adjacency matrix of the graph, the normalized Laplacian matrix is calculated and used for spectral deconvolution. A null-model is derived by repeating the same procedure after shuffling the mutations in the original mutation map. Analysis of the distance between consecutive eigenvalues (eigengaps) for the experimental data with respect to the null model allows identifying the number of informative eigengaps, corresponding to the number of coexisting RNA conformations (clusters). Once the number of clusters has been defined, fuzzy clustering is performed using a custom graph cut approach, that enables the weighting of vertices in accordance with their affinity to each cluster. This analysis is repeated across the whole transcript. Consecutive windows showing a compatible number of clusters are merged. Then, reads are re-assigned to the respective cluster, allowing the deconvolution of the cluster reactivity profiles and relative abundances.

Extended Data Fig. 2 | In silico validation of DRACO. **a**, Maximum number of conformations detected for 10 sets of 100 simulated RNAs, with length ranging from 600 to 1,500 nt, expected to form 1 to 4 conformations, at a coverage of 5,000X and a read length of 150 nt. Data are presented as mean values $\pm$ SD of the 10 sets. The individual data points, representing the mean of each set, are shown. **b**, Box-plot of median Pearson correlation coefficients (PCC) of reconstructed reactivity profiles for 10 sets of 100 simulated RNAs, with length ranging from 600 to 1,500 nt, expected to form 1 to 4 conformations, at a coverage of 5,000X and a read length of 150 nt. When DRACO detected more than one window with different numbers of clusters, only the largest window, spanning $>50\%$ of the RNA length, was considered. Boxes span the 25th to the 75th percentile. The center represents the median. Whiskers span from the 25th percentile - 1.5 times the IQR, to above the 75th percentile + 1.5 times the IQR. Data points falling outside of this range represent outliers and are reported as dots. **c**, Violin plot depicting the distribution of expected versus reconstructed conformation abundances for 10 sets of 100 simulated RNAs, with length ranging from 600 to 1,500 nt, expected to form 2 conformations with varying relative abundances, at a coverage of 5,000X and with a read length of 150 nt. When multiple windows were detected, only the largest window was considered. The Pearson correlation is indicated in the bottom-right corner of each plot. Whiskers span the 25th to the 75th percentile. The central dot represents the median.

Extended Data Fig. 3 | Validation of DRACO on in silico-merged in vitro-generated profiles. **a**, DRACO-deconvoluted profiles for *cspA* RNA folded and probed in vitro at either 37 °C or 10 °C (from Zhang et al., 2018), pooled at different percentages. The percentage of pooling is indicated next to each reconstructed profile. **b**, Heatmap of Pearson correlation coefficient for DRACO-deconvoluted profiles for each pool, compared to the expected profiles at 37 °C or 10 °C.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i>
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection Illumina bcl2fastq v2.20.0.422

Data analysis
 PEAR 0.9.11 (<https://cme.h-its.org/exelixis/web/software/pear/>)
 Cutadapt v2.1 (<https://pypi.org/project/cutadapt/>)
 Bowtie v2.3.5.1 (<http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>)
 STAR v2.7.1a (<https://github.com/alexdobin/STAR>)
 Samtools v1.9 (<http://www.htslib.org/download/>)
 RNA Framework v2.7.5 (<https://github.com/dincarnato/RNAFramework>)
 ViennaRNA v2.4.14 (<https://www.tbi.univie.ac.at/RNA/>)
 R-scape v1.4.0 (<http://eddylib.org/R-scape/>)
 INFERNAL v1.1.3 (including the cmalign, cmsearch, cmbuild, and cmcalibrate programs; <http://eddylib.org/infernal/>)
 DRACO v0.9b (<https://github.com/dincarnato/draco>)
 VARNA v3.93 (<http://varna.lri.fr/index.php?lang=en&page=home&css=varna>)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data has been deposited to the Gene Expression Omnibus (GEO) database, under the accession GSE158052. Additional processed files are available at http://www.incarnatolab.com/datasets/DRACO_Morandi_2021.php. The source code of DRACO is available from GitHub (<https://github.com/dincarnato/draco>).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For spectral clustering analysis, a median post-filtering coverage of 10000X per analyzed window was used. The minimum coverage of 5,000X was determined by 10 set of 100 simulations conducted on simulated RNAs of lengths ranging from 300 to 1500nt.
Data exclusions	Bases with mutation frequencies below 0.005 were ignored. This threshold was empirically selected to distinguish real DMS-MaPseq signals from the background.
Replication	SARS-CoV-2 structure was assayed by DMS-MaPseq in 2 independent replicates. Only 2 replicates were performed, and replication was successful, as proven by the high correlation between the two (Pearson R > 0.99)
Randomization	Randomization is not typically performed or required for in vitro experiments.
Blinding	No comparison between different (biological) conditions was involved, hence no blinding was required.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	Vero E6 cells obtained from ATCC.
Authentication	Cell line was not authenticated.
Mycoplasma contamination	Cells were tested negative for Mycoplasma using the MycoAlertTM Mycoplasma Detection Kit (Lonza, cat. LT07-318).
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.