Genome-scale deconvolution of RNA structure ensembles

Published version:

DOI:10.1038/s41592-021-01075-w

Genome-scale deconvolution of RNA structure ensembles

Edoardo Morandi[1,2§], Ilaria Manfredonia[2§], Lisa M. Simon[1§], Francesca Anselmi[1], Martijn J. van Hemert[3], Salvatore Oliviero[1*] and Danny Incarnato[2*]

[1] Dipartimento di Scienze della Vita e Biologia dei Sistemi, Università di Torino, Via Accademia Albertina 13, 10123 Torino, Italy

[2] Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, Nijenborgh 7, 9747 AG, Groningen, the Netherlands

[3] Molecular Virology Laboratory, Department of Medical Microbiology, Leiden University Medical Center, Leiden, the Netherlands

[§] These authors contributed equally to this work

* To whom correspondence should be addressed:
Danny Incarnato (d.incarnato@rug.nl) and Salvatore Oliviero (salvatore.oliviero@unito.it)

RNA structure heterogeneity represents the major challenge for the study of RNA structures by chemical probing. To solve this, we developed DRACO (Deconvolution of RNA Alternative COnformations), an algorithm for the reconstruction of individual reactivity profiles and relative stoichiometries of coexisting alternative RNA conformations from mutational profiling (MaP) experiments. After extensively validating the robustness of DRACO on both *in silico* and *in vitro* data, we applied it to DMS-MaPseq data from the full SARS-CoV-2 genome, identifying multiple regions folding into two mutually-exclusive conformations. Our work opens the way to dissecting the heterogeneity of the RNA structurome.

Although powerful, RNA structure analyses by means of chemicals probing with dimethyl sulfate (DMS) and Selective 2'-Hydroxyl Acylation analyzed by Primer Extension (SHAPE) reagents, suffer of the intrinsic limitation of only being able to provide an averaged measurement of the base reactivities of all the coexisting conformations simultaneously sampled by the RNA molecules in a biological sample[1,2]. Over the years, several computational approaches have been proposed to deal with the problem of RNA structure heterogeneity, many of which based on the attempt to identify a parsimonious subset of structures from the Boltzmann ensemble, that would justify the experimentally-measured reactivity profile for an RNA[3,4]. Main limitation of these approaches is the impossibility to identify the correct set of RNA conformations if these have a low probability of occurring within the Boltzmann ensemble, hence to be sampled. With the advent of mutational profiling (MaP) methods, based on the recording of DMS/SHAPE modification sites as mutations in the resulting cDNA molecules[5–7], it has become possible to record multiple modification sites, corresponding to residues that were simultaneously single-stranded in the same original RNA molecule, within the same cDNA product. In an early attempt to deconvolute multiple alternative conformations from MaP experiments, spectral clustering was proposed as a suitable approach to identify the number of coexisting RNA structures in a heterogeneous mixture[8]. More recently, an alternative approach named DREEM, based on expectation maximization, has been proposed[9]. This tool represents the first concrete attempt to deconvolute alternative structures from MaP experiments. Even if powerful in principle, it suffers of two major limitations. Particularly, (1) the maximum number of RNA conformations to search for is user-defined (two by default, maximum four), to reduce the risk of overestimating the number of conformations (also known as *overclustering*, a common problem with expectation maximization approaches), and (2) it can only handle experiments in which each sequencing read covers the entire length of the target RNA. The latter makes it only suitable for the analysis of short transcripts (within the maximum read length achievable on Illumina platforms, ~600 nt), or for targeted analyses, but not for transcriptome-scale analyses, characterized by short reads tiling long transcripts. Although DREEM can be in theory applied to longer transcripts by manual window sliding, it cannot handle the merging of overlapping RNA segments, a non-trivial computational problem.

To address these issues, we here introduce DRACO (Deconvolution of RNA Alternative COnformations), a fast and accurate algorithm for the deconvolution of alternative RNA conformations, and of their relative stoichiometries, from MaP experiments, based on combination of spectral clustering and fuzzy clustering (Supplementary Note 1). We sought to design an approach suitable for transcriptome-scale analyses, usually characterized by short tiling reads, covering only partly the analyzed transcripts. To this end, DRACO analysis is performed (by default) in sliding windows with a size of 90% the median length of reads, and an offset of 5% (Fig. 1a). Spectral clustering is performed for each window, allowing the automatic identification of the optimal number of conformations (clusters). The algorithm then merges overlapping windows (for which the same number of clusters have been detected), reconstructing overall mutational profiles. In case a large set of windows is found to form a discordant number of conformations with respect to surrounding windows, this set is merged in a single window and reported separately from surrounding window sets. To validate the algorithm, we first generated *in silico* DMS-MaPseq data (with read lengths varying from 50 to 150 nt), for 1,000 RNAs (with lengths ranging from 300 to 1,500 nt), designed to form up to 4 distinct conformations. DMS-induced mutations in reads were modeled as a binomial distribution, well approximating the observed distribution of a previously published dataset[8] (Supplementary Fig. 1-2). Analysis of *in silico* data (Supplementary Fig. 3-14) showed that DRACO accuracy relies on two main factors: read length and coverage. This can be easily explained by DRACO dependency on co-mutation information. Although higher coverages can partially compensate for the reduced amount of mutational information in shorter reads, best results were obtained with a read length of 150 nt and a minimum coverage of 5,000X. Under these conditions, DRACO correctly identified the expected number of conformations in nearly 100% of the cases (Fig. 1b), accurately deconvoluted the individual conformation mutational profiles (median PCC > 0.85; Fig. 1c) and precisely estimated relative conformation stoichiometries (PCC ≈ 0.99; Fig. 1d).

As *in silico*-generated data might not completely capture the complexity of a real DMS-MaPseq experiments, we further sought to test DRACO using *in vitro* data for *E. coli cspA* 5' UTR from a previous study[10]. *cspA* 5' UTR acts as an RNA thermometer, regulating the accessibility of the Shine-Dalgarno in response to the environment temperature, switching between a translationally-repressed conformation at 37°C and a translationally-competent conformation at 10°C[11]. After mapping DMS-MaPseq data from *in vitro* folding experiments at either 10°C or 37°C, reads from the two experiments were pooled at different percentages and analyzed using DRACO (Supplementary Fig. 15). Notably, DRACO successfully reconstructed the expected reactivity profiles with high accuracy, even with a conformation abundance of as little as 10% (PCC = 0.88). Furthermore, the cspA protein has been previously shown to act as an RNA chaperone on its own 5' UTR, mediating the refolding of the 10°C translationally-competent conformation into the 37°C translationally-repressed conformation. In the same study[10], the *cspA* 5' UTR was folded at 10°C was in the presence of increasing concentrations of the cspA protein and analyzed by DMS-MaPseq. While in the presence of 0.1 μM cspA the conformation of the 5' UTR resembled that observed at 37°C[10], use of half this amount of the cspA protein resulted in a reactivity profile that only partially correlated with both the 10°C and 37°C conformations. Prompted by this observation, we hypothesized that this might have been the consequence of the coexistence of both conformations in the sample. Strikingly, DRACO reconstructed two nearly-equimolar conformations (48.6% and 51.4% respectively; Fig. 2a), whose profiles were highly correlated to either the 10°C or the 37°C conformation (respectively, PCC = 0.83 and 0.85; Fig. 2b). Accordingly, use of these profiles as constraints for data-driven RNA structure prediction produced secondary structure models nearly identical to those expected for the 10°C and 37°C conformations (respectively, PPV: 1.00 and 0.91, sensitivity: 0.87 and 0.97; Fig. 2c). We further analyzed a recently published DMS-MaPseq dataset, originally generated to validate the DREEM algorithm[9] by probing the structure of the *add* riboswitch from *V. vulnificus*, either in the absence or presence of 5 mM adenine. While DREEM identified three conformations under both conditions[9], analysis with 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% represents 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 higher robustness of the DRACO algorithm, as well as its lower propensity to overclustering, as compared to expectation maximization-based approaches, rather than a lower sensitivity (see Supplementary Note 2). Encouraged by the performances of DRACO on both *in silico* and *in vitro* data, we next sought to apply it to the analysis of the SARS-CoV-2 virus RNA genome structure. In a recent report[12], we have defined the secondary structure of the full SARS-CoV-2 genome by SHAPE-MaP, identifying conserved structure elements folding into single well-defined conformations and harboring potentially druggable pockets. Although powerful, our previous approach was limited to the analysis of regions folding into a single well-defined conformation, possibly overlooking important structure elements or transient pockets. We therefore sought to query (in duplicate) the full *in vitro* refolded SARS-CoV-2 genome by DMS-MaPseq analysis. Paired-end 150 bp sequencing and assembly of paired reads produced over 2.2×10^7 fragments (per each replicate), resulting in a median coverage of ~9.9×10^4 (Supplementary Fig. 16a-b), way above the minimum coverage requirement of DRACO. Our data showed exceptional correlation between replicates (PCC = 0.99, Supplementary Fig. 16c) and agreement with well-defined Sarbecovirus structures in the 5' UTR, as well as additional conserved RNA structure elements we have recently identified[12] (Supplementary Fig. 17). Analysis with DRACO unambiguously identified 22 windows, roughly accounting for ~15.5% of the SARS-CoV-2 genome, coherently reported to fold into 2 conformations in both replicates (Supplementary Fig. 18a). We observed an exceptional overall correlation of reactivity profiles for reconstructed conformations across replicates (PCC = 0.86; Supplementary Fig. 18b), as well as highly consistent relative conformation abundances (Supplementary Fig. 18c), with an average variation of just ±1.9%. By inspecting the distribution of these windows, we noticed an enrichment at ORF boundaries (11/22 windows (50%) spanning ORF starts/ends, versus just ~19% windows over 10,000 randomizations per window of matching size; $P = 1.0e\text{-}3$, one-sided Binomial test), including one window spanning the ORF1a/ORF1b boundary, overlapping with the frameshifting element (FSE, pos. 13369-13542; Supplementary Fig. 19). Strikingly, our data does not support the existence of a pseudoknotted structure at the level of the FSE. Rather, this region is likely to fold into either a single extended stem-loop or two stem-loop structures. This observation is further supported by a recently proposed structure analysis by DMS-MaPseq of the SARS-CoV-2 genome in living infected host cells[13]. It is conceivable that this and the other identified RNA switches might be involved in controlling either the translation of SARS-CoV-2 proteins, or the discontinuous transcription of subgenomic mRNAs (or both), but additional experiments will be needed to investigate their functional relevance. Interestingly, one of the identified windows encompassed the 3' UTR (pos. 29546-29767), showing consistent abundance estimates and reactivity profiles for the two identified conformations across the two analyzed replicates (Fig. 3a, b). The major conformation (63.4 ± 1.7%) showed a reactivity pattern compatible with the known phylogenetically-inferred 3' UTR structure of Sarbecoviruses, while the minor conformation (36.6 ± 1.7%) was predicted to form an alternative three-way junction structure, sequestering both the BSL and P2 helices (Fig. 3c). We further evaluated the conservation of this alternative conformation by using an approach we have recently exploited to automatically identify regions of the SARS-

CoV-2 genome showing significant covariation[12] (see Methods), based on the use of Infernal[14], to build a structurally-informed alignment of related coronavirus sequences, and R-scape[15] to evaluate the significance of the observed covariations. Only sequences simultaneously matching both structures were retained. Strikingly, formation of the alternative three-way junction structure showed significant covariation support (Fig. 3d), hinting at its functional relevance. Notably, when performing the same analysis on the two conformations independently, even more significantly covarying base-pairs were detected (Supplementary Fig. 20). Furthermore, re-analysis of a recently published dataset of RNA-RNA interaction capture in SARS-CoV-2 infected cells[16] provided support for the presence of both conformations *in vivo*. 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 an important regulatory region in the SARS-CoV-2 genome.

In summary, we have here introduced DRACO, the first algorithm enabling genome-scale deconvolution of RNA alternative conformations from MaP experiments. We can anticipate that use of DRACO will allow the exploration of the RNA structurome at unprecedented resolution, revealing transient and dynamic features of cellular transcriptomes.

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 (http://www.netlib.org/blas/) and LAPACK (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 SAM/BAM alignment file with the *rf-count* tool of the RNA Framework (parameter: *-mm*). With default parameters, DRACO takes ~8-10 hours, on a single thread, to analyze ~17 million reads mapping to the SARS-CoV-2 genome. A complete description of the algorithm, including pseudo-codes, is provided in Supplementary Note 1. DRACO source code is available from GitHub (https://github.com/dincarnato/draco).

***In silico* generation of DMS-MaPseq data.** 1,000 RNA sequences with an average A/C content of 50% and varying lengths (300, 600, 900 or 1,500 nt) were randomly generated. DMS modification profiles for one to four different conformations were then generated by randomly setting as single-stranded ~30% of the A/C residues. This fraction of single-stranded A/C residues represents an underestimate of what is expected for real RNAs (~51.3% of single-stranded A/C residues for *E. coli* 16S/23S rRNAs). Mutated reads matching these modification profiles were then generated (in MM format) to obtain a median coverage per base of 2,000X, 5,000X, 10,000X, or 20,000X, using the *generate_mm* tool (available from DRACO's repository). Distribution of DMS-induced mutations in reads was empirically learnt from a previously published dataset[8] (Supplementary Fig. 1) and well approximated by a binomial distribution with p = 0.01927 and n = 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*). As A/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 (whose number is controlled by the "*--max-collapsing-windows*" parameter) showing a discordant number of conformations with respect to surrounding windows. As 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 nt. Given that, by default, windows are slid by 5% the median read length, these "--max-collapsing-window" values imply that just 12.5 (for 50 nt reads) to 15 (for 150 nt reads) bases are ignored in such situations.

Analysis of DMS-MaPseq data. All the relevant analysis steps, from reads alignment to data normalization and structure modeling, were performed using RNA Framework[17]. 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 by jackknifing, using a DMS-MaPseq dataset for *ex vivo* deproteinized *E. coli* rRNAs we previously published[7] (accession: SRR8172706) and the *rf-jackknife* tool (parameters: *-rp '-md 600 -nlp' -x*).

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 -cqo -mp '--very-sensitive-local'*). As a lower fraction of reads aligned to the reference for the experiment conducted at 37°C, the BAM file from the experiment conducted at 10°C was randomly shuffled 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 stoichiometries (%) of 90-10, 80-20, 70-30, 60-40, or 50-50 of respectively the 10°C and 37°C conformations. 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 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 by then 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) was 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. Reads for DMS-MaPseq data of *in vitro* folded *add* riboswitch from *V. vulnificus*, 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 v0.9.11[18] and then mapped to the *add* riboswitch using the *rf-map* tool (parameters: *-cq5 20 -cqo -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 by then 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 T-175 flasks in Dulbecco's modified Eagle's medium (DMEM; Lonza, cat. 12-604F), supplemented with 8% fetal calf serum (FCS; Bodinco), 2 mM L-glutamine, 100 U/mL of penicillin and 100 μg/mL of streptomycin (Sigma Aldrich, cat. P4333-20ML) at 37°C in an atmosphere of 5% CO2 and 95%–99% humidity. Cells were infected at a MOI of 1.5 with SARS-CoV-2/Leiden-0002 (GenBank accession: MT510999), a clinical isolate obtained from a nasopharyngeal sample at LUMC, which was passaged twice in Vero E6 cells before use. Infections were performed in Eagle's minimal essential medium (EMEM; Lonza, cat. 12-611F) supplemented with 25 mM HEPES, 2% FCS, 2 mM L-glutamine, and antibiotics. At 16 h post-infection, infected cells were harvested by trypsinization, followed by resuspension in EMEM supplemented with 2% FCS, and then washed with 50 mL 1X PBS.

All experiments with infectious SARS-CoV-2 were performed in a biosafety level 3 facility at the LUMC.

Total RNA extraction and *in vitro* folding

Approximately $5x10^6$ of the harvested infected cells were resuspended in 1 mL of TriPure Isolation Reagent (Sigma Aldrich, cat. 11667157001) and 200 µl of chloroform were added. The sample was vigorously vortexed for 15 sec and then incubated for 2 min at room temperature, after which it was centrifuged for 15 min at 12,500 x g (4°C). The upper aqueous phase was collected in a clean 2 mL tube, supplemented with 1 mL (~2 volumes) of 100% ethanol, and then loaded on an RNA Clean & Concentrator-25 column (Zymo Research, cat. R1017). *In vitro* folding was carried out as previously described[7,12]. Briefly, ~5 µg of total RNA from infected Vero E6 cells was first depleted of ribosomal RNAs using the RiboMinus™ Eukaryote System v2 (ThermoFisher Scientific, cat. A15026), following manufacturer instructions. Ribo- RNA in a volume of 39 µl was denatured at 95°C for 2 min, then transferred immediately to ice and incubated for 1 min. 10 µl of ice-cold 5X RNA Folding Buffer [500 mM HEPES pH 7.9; 500 mM NaCl] supplemented with 20 U of SUPERase•In™ RNase Inhibitor (ThermoFisher Scientific, cat. AM2696) were added. RNA was then incubated for 10 min at 37°C to allow secondary structure formation. Subsequently, 1 µl of 500 mM $MgCl_2$ (pre-warmed at 37°C) was added and RNA was further incubated for 20 min at 37°C to allow tertiary structure formation.

Probing of SARS-CoV-2 RNA

For probing of RNA, DMS was pre-diluted 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 then quenched by the addition of 1 volume DTT 1.4 M and then purified on an RNA Clean & Concentrator-5 column (Zymo Research, cat. R1013).

DMS-MaPseq analysis of SARS-CoV-2 RNA

DMS-MaPseq of SARS-CoV-2 was conducted a previously described[7], with minor changes. First, probed RNA was fragmented to a median size of 150 nt by incubation at 94°C for 8 min in RNA Fragmentation Buffer [65 mM Tris-HCl pH 8.0; 95 mM KCl; 4 mM $MgCl_2$], then purified with NucleoMag NGS Clean-up and Size Select beads (Macherey Nagel, cat. 744970), supplemented with 10 U SUPERase•In™ RNase Inhibitor, and eluted in 8 µl NF H_2O. Eluted RNA was supplemented with 1 µl 50 µM random hexamers and 2 µl dNTPs (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 $MgCl_2$], 1 µl DTT 0.1 M, 20 U SUPERase•In™ RNase Inhibitor and 200 U TGIRT™-III Enzyme (InGex, cat. TGIRT50). Reactions were incubated at 25°C for 10 min to allow partial primer extension, followed by 2 h at 57°C. TGIRT-III was degraded by addition of 2 µg Proteinase K, followed by incubation at 37°C for 20 min. Proteinase K was inactivated by addition of Protease Inhibitor Cocktail (Sigma Aldrich, cat. P8340). Reverse transcription reactions were then used as input for the NEBNext® Ultra II Non-Directional RNA Second Strand Synthesis Module (New England Biolabs, cat. E6111L). Second strand synthesis was performed by incubating 1 h at 16°C, as per manufacturer instructions. DsDNA 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 manufacturer instructions.

Analysis of SARS-CoV-2 DMS-MaPseq data. After clipping adapter sequences using Cutadapt v2.1[19] (parameters: *-a AGATCGGAAGAGC -A AGATCGGAAGAGC -O 1 -m 100:100*), paired-end reads were merged using PEAR v0.9.11[18] and then mapped to the SARS-CoV-2 reference using the *rf-map* tool and the Bowtie2 algorithm, with soft-clipping enabled, (parameters: *-b2 -cq5 20 -ctn -cmn 0 -cl 150 -mp '--very-sensitive-local'*). An MM file was then generated from the resulting BAM alignment using the *rf-count* tool, by only keeping reads covering at least 150 bases. Insertions and deletions were ignored (as they account for less than 6% of DMS-induced mutations when using TGIRT-III[6]), considering only mutations having Phred qualities > 20. Furthermore, mutations were only considered when the two surrounding bases had Phred qualities > 20 as well. 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 10,000X were selected. To select windows that were coherently folding into multiple conformations in both replicates, we retained windows predicted to have the same number of conformations in the two replicates, overlapping by at least 75% of their length, and considered only their intersection. Deconvoluted reactivity profiles for matching conformations from the two replicates were then averaged and used for secondary structure modeling. Correlation between reconstructed conformations from the two replicates were calculated using 90% of the reactivity values in the window, after excluding the first and last 5% of the A/C bases, to avoid terminal biases.

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 we have previously introduced[12] (*cm-builder*; https://github.com/dincarnato/labtools), built on top of Infernal 1.1.3[14]. Briefly, we first built two covariance models (CMs) from Stockholm files containing only the SARS-CoV-2 sequence and the two alternative 3' UTR structures, using the *cmbuild* module. After calibrating the CMs using the *cmcalibrate* module, we used them to search for RNA homologs in a database composed of all the non-redundant coronavirus complete genome sequences from the ViPR database[20] (https://www.viprbrc.org/brc/home.spg?decorator=corona), as well as a set of representative coronavirus genomes from 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 genomes[21] and restricted the selection to matches falling at the same relative position within their genome, with a tolerance of 3.5% (roughly corresponding to a maximum allowed shift of 1050 nt in a 30 kb genome). Through this more "conservative" selection, we only kept 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 <50% of the structure were discarded. A fourth filtering step was also applied when analyzing simultaneously 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 CMs using the *cmalign* module and the resulting alignments were used to build new CMs. The whole process was repeated for a total of 3 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[15] and APC-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 windows folding into two mutually-exclusive conformations and 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/end positions of the SARS-CoV-2 ORFs, including each of the individual proteins within the polyprotein ORF1a/b (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 = ratio between 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 cells[16] was obtained from GEO (GSE154662). The dataset consisted of 2 biological replicates, each one composed of a control (C) and the actual COMRADES sample (S). A reference was built on all human transcripts from refGene, plus the sequence of the SARS-CoV-2 genome, using STAR v2.7.1a[22] (parameters: *--runMode genomeGenerate --genomeSAindexNbases 12*), and reads were aligned to the reference using the same (parameters: *--runMode alignReads --outFilterMultimapNmax 100 --outSAMattributes All --alignIntronMin 1 --scoreGapNoncan -4 --scoreGapATAC -4 --chimSegmentMin 15 --chimJunctionOverhangMin 15*). Resulting alignments (as well as chiastic alignments from the junctions file) were filtered, discarding ungapped reads, reads having 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 2 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 $k = S_{i-j}$, $n = S_{tot}$, and $p = C_{i-j} / C_{tot}$. Only base-pairs with $p\text{-}value < 0.05$ in both replicates were considered to have *in vivo* support.

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_2020.php.

Acknowledgements

D.I. was supported by the Dutch Research Council (NWO) as part of the research programme NWO Open Competitie ENW - XS with 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

Project conceptualization: E.M. and D.I.; Wet-lab: I.M., L.M.S., and F.A.; SARS-CoV-2 manipulations: M.J.H.; DRACO algorithm design and implementation: E.M. and D.I.; Bioinformatics, structure modeling and data analysis: E.M. and D.I.; Writing: D.I. and S.O.

Competing interests

The authors declare no competing interests.

References

1. Incarnato, D. & Oliviero, S. The RNA Epistructurome: Uncovering RNA Function by Studying Structure and Post-Transcriptional Modifications. *Trends in biotechnology* **35**, 318–333 (2017).

2. Strobel, E. J., Yu, A. M. & Lucks, J. B. High-throughput determination of RNA structures. *Nature reviews. Genetics* **19**, 615–634 (2018).

3. Spasic, A., Assmann, S. M., Bevilacqua, P. C. & Mathews, D. H. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. *Nucleic Acids Res* **46**, 314–323 (2017).

4. Li, H. & Aviran, S. Statistical modeling of RNA structure profiling experiments enables parsimonious reconstruction of structure landscapes. *Nat Commun* **9**, 606 (2018).

5. 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). *Nature Methods* **11**, 959–965 (2014).

6. Zubradt, M. *et al.* DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nature Methods* **14**, 75–82 (2017).

7. 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).

8. Homan, P. J. *et al.* Single-molecule correlated chemical probing of RNA. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 13858–13863 (2014).

9. Tomezsko, P. J. *et al.* Determination of RNA structural diversity and its role in HIV-1 RNA splicing. *Nature* **582**, 438–442 (2020).

10. Zhang, Y. *et al.* A Stress Response that Monitors and Regulates mRNA Structure Is Central to Cold Shock Adaptation. *Molecular cell* **70**, 274-286.e7 (2018).

11. 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).

12. Manfredonia, I. *et al.* Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res* gkaa1053- (2020) doi:10.1093/nar/gkaa1053.

13. Lan, T. C. T. *et al.* Structure of the full SARS-CoV-2 RNA genome in infected cells. *Biorxiv* 2020.06.29.178343 (2020) doi:10.1101/2020.06.29.178343.

14. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinform Oxf Engl* **29**, 2933–5 (2013).

15. 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).

16. Ziv, O. *et al.* The short- and long-range RNA-RNA Interactome of SARS-CoV-2. *Mol Cell* (2020) doi:10.1016/j.molcel.2020.11.004.

17. 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 Research* **46**, e97–e97 (2018).

18. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

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

20. Pickett, B. E. *et al.* ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res* **40**, D593-8 (2011).

21. Lauber, C. *et al.* The footprint of genome architecture in the largest genome expansion in RNA viruses. *Plos Pathog* **9**, e1003500 (2013).

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

Figure legends

Figure 1. Overview of the DRACO algorithm. (**a**) Schematic representation of the DRACO algorithm. (**b**) Maximum number of conformations detected for 10 sets of 100 simulated RNAs, with length ranging from 300 to 1500 nt, expected to form 1 to 4 conformations, at a coverage of 5,000X and a read length of 150 nt. Error bars represent SD of the 10 sets. (**c**) Box-plot of median Pearson correlation coefficients (PCC) of reconstructed reactivity profiles for 10 sets of 100 simulated 1000 simulated RNAs, with length ranging from 300 to 1500 nt, expected to form 1 to 4 conformations, at a coverage of 20,000X 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. (**d**) Violin plot depicting the distribution of expected versus reconstructed conformation abundances for 10 sets of 100 simulated RNAs, with length ranging from 300 to 1500 nt, expected to form 1 to 4 conformations, at a coverage of 20,000X and a read length of 150 nt. The Pearson correlation is indicated in the bottom-right corner of each plot.

Figure 2. *In vitro* validation of DRACO. (**a**) Original DMS-MaPseq profile, and DRACO-deconvoluted profiles for *cspA* 5' UTR folded at 10°C in the presence of 50 *μ*M cspA recombinant protein, from Zhang *et al.*, 2018[10]. Schematic representation of the structures, and the reconstructed relative abundances are indicated. (**b**) Heatmap of Pearson correlation coefficients 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. Positive predictive value (PPV) and sensitivity are indicated). (**d**) DRACO-deconvoluted profiles for *V. vulnificus add* riboswitch, in the absence (1 conformation detected) or presence (2 conformations detected) of 5 mM adenine, from Tomezsko *et al.*, 2020[9]. Schematic representation of the structures, and the reconstructed relative abundances are indicated. (**e**) Heatmap of Pearson correlation coefficients 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.

Figure 3. A conserved structural switch in the 3' UTR of SARS-CoV-2. (**a**) Relative abundances of the two alternative conformations (A and B) of the SARS-CoV-2 3' UTR. Error bars represent the standard deviation from two replicates. (**b**) Heat scatterplot of base reactivities for DRACO-deconvoluted reactivity profiles in replicate #1 versus replicate #2 of conformation A (left) and B (right). Base-pairs whose existence is supported by significant enrichment of RNA-RNA chimeras from *in vivo* COMRADES analysis (Ziv *et al.*, 2020[16]) are boxed in light blue. (**c**) Secondary structure models with overlaid base reactivities for conformation A and B. (**d**) Structure models for conformation A (top) and B (bottom), inferred by simultaneous phylogenetic analysis. Structures have been generated using the R2R software. Base-pairs showing significant covariation (as determined by R-scape) are boxed in green (E-value < 0.05) and violet (E-value < 0.1) respectively.