

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Genome-scale deconvolution of RNA structure ensembles

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/1771193> since 2021-03-10T17:45:36Z

Published version:

DOI:10.1038/s41592-021-01075-w

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

Genome-scale deconvolution of RNA structure ensembles

Edoardo Morandi^{1,2§}, Ilaria Manfredonia^{2§}, Lisa M. Simon^{1§}, Francesca Anselmi¹, Martijn J. van Hemert³, 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.

1 Although powerful, RNA structure analyses by means of chemicals probing with dimethyl
2 sulfate (DMS) and Selective 2'-Hydroxyl Acylation analyzed by Primer Extension (SHAPE)
3 reagents, suffer of the intrinsic limitation of only being able to provide an averaged
4 measurement of the base reactivities of all the coexisting conformations simultaneously
5 sampled by the RNA molecules in a biological sample^{1,2}. Over the years, several
6 computational approaches have been proposed to deal with the problem of RNA structure
7 heterogeneity, many of which based on the attempt to identify a parsimonious subset of
8 structures from the Boltzmann ensemble, that would justify the experimentally-measured
9 reactivity profile for an RNA^{3,4}. Main limitation of these approaches is the impossibility to
10 identify the correct set of RNA conformations if these have a low probability of occurring
11 within the Boltzmann ensemble, hence to be sampled. With the advent of mutational profiling
12 (MaP) methods, based on the recording of DMS/SHAPE modification sites as mutations in
13 the resulting cDNA molecules⁵⁻⁷, it has become possible to record multiple modification
14 sites, corresponding to residues that were simultaneously single-stranded in the same
15 original RNA molecule, within the same cDNA product. In an early attempt to deconvolute
16 multiple alternative conformations from MaP experiments, spectral clustering was proposed
17 as a suitable approach to identify the number of coexisting RNA structures in a
18 heterogeneous mixture⁸. More recently, an alternative approach named DREEM, based on
19 expectation maximization, has been proposed⁹. This tool represents the first concrete
20 attempt to deconvolute alternative structures from MaP experiments. Even if powerful in
21 principle, it suffers of two major limitations. Particularly, (1) the maximum number of RNA
22 conformations to search for is user-defined (two by default, maximum four), to reduce the
23 risk of overestimating the number of conformations (also known as *overclustering*, a
24 common problem with expectation maximization approaches), and (2) it can only handle
25 experiments in which each sequencing read covers the entire length of the target RNA. The
26 latter makes it only suitable for the analysis of short transcripts (within the maximum read
27 length achievable on Illumina platforms, ~600 nt), or for targeted analyses, but not for
28 transcriptome-scale analyses, characterized by short reads tiling long transcripts. Although
29 DREEM can be in theory applied to longer transcripts by manual window sliding, it cannot
30 handle the merging of overlapping RNA segments, a non-trivial computational problem.

31 To address these issues, we here introduce DRACO (Deconvolution of RNA Alternative
32 Conformations), a fast and accurate algorithm for the deconvolution of alternative RNA
33 conformations, and of their relative stoichiometries, from MaP experiments, based on

1 combination of spectral clustering and fuzzy clustering (Supplementary Note 1). We sought
2 to design an approach suitable for transcriptome-scale analyses, usually characterized by
3 short tiling reads, covering only partly the analyzed transcripts. To this end, DRACO analysis
4 is performed (by default) in sliding windows with a size of 90% the median length of reads,
5 and an offset of 5% (Fig. 1a). Spectral clustering is performed for each window, allowing the
6 automatic identification of the optimal number of conformations (clusters). The algorithm
7 then merges overlapping windows (for which the same number of clusters have been
8 detected), reconstructing overall mutational profiles. In case a large set of windows is found
9 to form a discordant number of conformations with respect to surrounding windows, this set
10 is merged in a single window and reported separately from surrounding window sets. To
11 validate the algorithm, we first generated *in silico* DMS-MaPseq data (with read lengths
12 varying from 50 to 150 nt), for 1,000 RNAs (with lengths ranging from 300 to 1,500 nt),
13 designed to form up to 4 distinct conformations. DMS-induced mutations in reads were
14 modeled as a binomial distribution, well approximating the observed distribution of a
15 previously published dataset⁸ (Supplementary Fig. 1-2). Analysis of *in silico* data
16 (Supplementary Fig. 3-14) showed that DRACO accuracy relies on two main factors: read
17 length and coverage. This can be easily explained by DRACO dependency on co-mutation
18 information. Although higher coverages can partially compensate for the reduced amount of
19 mutational information in shorter reads, best results were obtained with a read length of 150
20 nt and a minimum coverage of 5,000X. Under these conditions, DRACO correctly identified
21 the expected number of conformations in nearly 100% of the cases (Fig. 1b), accurately
22 deconvoluted the individual conformation mutational profiles (median PCC > 0.85; Fig. 1c)
23 and precisely estimated relative conformation stoichiometries (PCC ≈ 0.99; Fig. 1d).
24 As *in silico*-generated data might not completely capture the complexity of a real DMS-
25 MaPseq experiments, we further sought to test DRACO using *in vitro* data for *E. coli cspA*
26 5' UTR from a previous study¹⁰. *cspA* 5' UTR acts as an RNA thermometer, regulating the
27 accessibility of the Shine-Dalgarno in response to the environment temperature, switching
28 between a translationally-repressed conformation at 37°C and a translationally-competent
29 conformation at 10°C¹¹. After mapping DMS-MaPseq data from *in vitro* folding experiments
30 at either 10°C or 37°C, reads from the two experiments were pooled at different percentages
31 and analyzed using DRACO (Supplementary Fig. 15). Notably, DRACO successfully
32 reconstructed the expected reactivity profiles with high accuracy, even with a conformation
33 abundance of as little as 10% (PCC = 0.88). Furthermore, the *cspA* protein has been

1 previously shown to act as an RNA chaperone on its own 5' UTR, mediating the refolding of
2 the 10°C translationally-competent conformation into the 37°C translationally-repressed
3 conformation. In the same study¹⁰, the *cspA* 5' UTR was folded at 10°C was in the presence
4 of increasing concentrations of the *cspA* protein and analyzed by DMS-MaPseq. While in
5 the presence of 0.1 μ M *cspA* the conformation of the 5' UTR resembled that observed at
6 37°C¹⁰, use of half this amount of the *cspA* protein resulted in a reactivity profile that only
7 partially correlated with both the 10°C and 37°C conformations. Prompted by this
8 observation, we hypothesized that this might have been the consequence of the coexistence
9 of both conformations in the sample. Strikingly, DRACO reconstructed two nearly-equimolar
10 conformations (48.6% and 51.4% respectively; Fig. 2a), whose profiles were highly
11 correlated to either the 10°C or the 37°C conformation (respectively, PCC = 0.83 and 0.85;
12 Fig. 2b). Accordingly, use of these profiles as constraints for data-driven RNA structure
13 prediction produced secondary structure models nearly identical to those expected for the
14 10°C and 37°C conformations (respectively, PPV: 1.00 and 0.91, sensitivity: 0.87 and 0.97;
15 Fig. 2c). We further analyzed a recently published DMS-MaPseq dataset, originally
16 generated to validate the DREEM algorithm⁹ by probing the structure of the *add* riboswitch
17 from *V. vulnificus*, either in the absence or presence of 5 mM adenine. While DREEM
18 identified three conformations under both conditions⁹, analysis with DRACO showed that a
19 single conformation is present in the absence of adenine, and that the addition of adenine
20 triggers the conformation switch towards the translation-competent conformation on ~65.6%
21 of the RNA molecules (Fig. 2d). The remaining ~34.4% represents instead the translation-
22 incompetent conformation, as demonstrated by the high correlation to the adenine-free
23 sample (PCC = 0.96, Fig. 2e), as well as by the agreement between the predicted and the
24 expected secondary structures of the two conformations (Fig. 2f). These results support the
25 higher robustness of the DRACO algorithm, as well as its lower propensity to overclustering,
26 as compared to expectation maximization-based approaches, rather than a lower sensitivity
27 (see Supplementary Note 2). Encouraged by the performances of DRACO on both *in silico*
28 and *in vitro* data, we next sought to apply it to the analysis of the SARS-CoV-2 virus RNA
29 genome structure. In a recent report¹², we have defined the secondary structure of the full
30 SARS-CoV-2 genome by SHAPE-MaP, identifying conserved structure elements folding into
31 single well-defined conformations and harboring potentially druggable pockets. Although
32 powerful, our previous approach was limited to the analysis of regions folding into a single
33 well-defined conformation, possibly overlooking important structure elements or transient

1 pockets. We therefore sought to query (in duplicate) the full *in vitro* refolded SARS-CoV-2
2 genome by DMS-MaPseq analysis. Paired-end 150 bp sequencing and assembly of paired
3 reads produced over 2.2×10^7 fragments (per each replicate), resulting in a median coverage
4 of $\sim 9.9 \times 10^4$ (Supplementary Fig. 16a-b), way above the minimum coverage requirement of
5 DRACO. Our data showed exceptional correlation between replicates (PCC = 0.99,
6 Supplementary Fig. 16c) and agreement with well-defined Sarbecovirus structures in the 5'
7 UTR, as well as additional conserved RNA structure elements we have recently identified¹²
8 (Supplementary Fig. 17). Analysis with DRACO unambiguously identified 22 windows,
9 roughly accounting for $\sim 15.5\%$ of the SARS-CoV-2 genome, coherently reported to fold into
10 2 conformations in both replicates (Supplementary Fig. 18a). We observed an exceptional
11 overall correlation of reactivity profiles for reconstructed conformations across replicates
12 (PCC = 0.86; Supplementary Fig. 18b), as well as highly consistent relative conformation
13 abundances (Supplementary Fig. 18c), with an average variation of just $\pm 1.9\%$. By
14 inspecting the distribution of these windows, we noticed an enrichment at ORF boundaries
15 (11/22 windows (50%) spanning ORF starts/ends, versus just $\sim 19\%$ windows over 10,000
16 randomizations per window of matching size; $P = 1.0 \times 10^{-3}$, one-sided Binomial test), including
17 one window spanning the ORF1a/ORF1b boundary, overlapping with the frameshifting
18 element (FSE, pos. 13369-13542; Supplementary Fig. 19). Strikingly, our data does not
19 support the existence of a pseudoknotted structure at the level of the FSE. Rather, this
20 region is likely to fold into either a single extended stem-loop or two stem-loop structures.
21 This observation is further supported by a recently proposed structure analysis by DMS-
22 MaPseq of the SARS-CoV-2 genome in living infected host cells¹³. It is conceivable that this
23 and the other identified RNA switches might be involved in controlling either the translation
24 of SARS-CoV-2 proteins, or the discontinuous transcription of subgenomic mRNAs (or
25 both), but additional experiments will be needed to investigate their functional relevance.
26 Interestingly, one of the identified windows encompassed the 3' UTR (pos. 29546-29767),
27 showing consistent abundance estimates and reactivity profiles for the two identified
28 conformations across the two analyzed replicates (Fig. 3a, b). The major conformation (63.4
29 $\pm 1.7\%$) showed a reactivity pattern compatible with the known phylogenetically-inferred 3'
30 UTR structure of Sarbecoviruses, while the minor conformation ($36.6 \pm 1.7\%$) was predicted
31 to form an alternative three-way junction structure, sequestering both the BSL and P2
32 helices (Fig. 3c). We further evaluated the conservation of this alternative conformation by
33 using an approach we have recently exploited to automatically identify regions of the SARS-

1 CoV-2 genome showing significant covariation¹² (see Methods), based on the use of
2 Infernal¹⁴, to build a structurally-informed alignment of related coronavirus sequences, and
3 R-scape¹⁵ to evaluate the significance of the observed covariations. Only sequences
4 simultaneously matching both structures were retained. Strikingly, formation of the
5 alternative three-way junction structure showed significant covariation support (Fig. 3d),
6 hinting at its functional relevance. Notably, when performing the same analysis on the two
7 conformations independently, even more significantly covarying base-pairs were detected
8 (Supplementary Fig. 20). Furthermore, re-analysis of a recently published dataset of RNA-
9 RNA interaction capture in SARS-CoV-2 infected cells¹⁶ provided support for the presence
10 of both conformations *in vivo*. Altogether, these data demonstrate the ability of DRACO to
11 capture otherwise hidden structural features, and reveal the presence of a conserved RNA
12 switch at the level of an important regulatory region in the SARS-CoV-2 genome.

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

1 **Methods**

2
3 **DRACO algorithm.** The DRACO algorithm is implemented in C++ and exploits the
4 Armadillo library (<http://arma.sourceforge.net>), built on top of the BLAS
5 (<http://www.netlib.org/blas/>) and LAPACK (<http://www.netlib.org/lapack/>) libraries for fast
6 matrix manipulation and eigenvalue decomposition. As input, DRACO takes Mutation Map
7 (MM) format files. These files store the relative coordinates of mutations for each read
8 mapping on a given transcript and can be generated by processing a SAM/BAM alignment
9 file with the *rf-count* tool of the RNA Framework (parameter: *-mm*). With default parameters,
10 DRACO takes ~8-10 hours, on a single thread, to analyze ~17 million reads mapping to the
11 SARS-CoV-2 genome. A complete description of the algorithm, including pseudo-codes, is
12 provided in Supplementary Note 1. DRACO source code is available from GitHub
13 (<https://github.com/dincarnato/draco>).

14
15 ***In silico* generation of DMS-MaPseq data.** 1,000 RNA sequences with an average A/C
16 content of 50% and varying lengths (300, 600, 900 or 1,500 nt) were randomly generated.
17 DMS modification profiles for one to four different conformations were then generated by
18 randomly setting as single-stranded ~30% of the A/C residues. This fraction of single-
19 stranded A/C residues represents an underestimate of what is expected for real RNAs
20 (~51.3% of single-stranded A/C residues for *E. coli* 16S/23S rRNAs). Mutated reads
21 matching these modification profiles were then generated (in MM format) to obtain a median
22 coverage per base of 2,000X, 5,000X, 10,000X, or 20,000X, using the *generate_mm* tool
23 (available from DRACO's repository). Distribution of DMS-induced mutations in reads was
24 empirically learnt from a previously published dataset⁸ (Supplementary Fig. 1) and well
25 approximated by a binomial distribution with $p = 0.01927$ and $n =$ length of the transcript
26 (Supplementary Fig. 2).

27
28 **Analysis of *in silico*-generated DMS-MaPseq data.** *In silico*-generated MM files were
29 analyzed using DRACO (parameters: *--set-all-uninformative-to-one --set-uninformative-*
30 *clusters-to-surrounding --max-collapsing-windows <variable> --first-eigengap-threshold*
31 *0.9*). As A/C residues are non-uniformly distributed along transcripts, certain regions of the
32 RNA can give rise to reads bearing a lower mutational information content, possibly leading
33 to a local under (or over) estimate of the number of conformations. To account for this,

1 DRACO can ignore a small set of windows (whose number is controlled by the "*--max-*
2 *collapsing-windows*" parameter) showing a discordant number of conformations with
3 respect to surrounding windows. As the window size is determined by the read length (by
4 default, 90% of the median read length), the number of discordant windows is expected to
5 increase with decreasing read lengths. Therefore, the "*--max-collapsing-window*" parameter
6 was linearly decreased from 5 to 2 with increasing read lengths from 50 to 150 nt. Given
7 that, by default, windows are slid by 5% the median read length, these "*--max-collapsing-*
8 *window*" values imply that just 12.5 (for 50 nt reads) to 15 (for 150 nt reads) bases are
9 ignored in such situations.

10

11 **Analysis of DMS-MaPseq data.** All the relevant analysis steps, from reads alignment to
12 data normalization and structure modeling, were performed using RNA Framework¹⁷. All
13 tools referenced in the following paragraphs are distributed as part of the RNA Framework
14 suite (<https://github.com/dincarnato/RNAFramework>). Specific analysis parameters are
15 detailed in the respective paragraphs.

16

17 **Optimization of folding parameters.** For structure predictions, optimal *slope* (2.4) and
18 *intercept* (-0.2) values were identified by jackknifing, using a DMS-MaPseq dataset for *ex*
19 *vivo* deproteinized *E. coli* rRNAs we previously published⁷ (accession: SRR8172706) and
20 the *rf-jackknife* tool (parameters: *-rp '-md 600 -nlp' -x*).

21

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

1 *fraction 0.1 --set-uninformative-clusters-to-surrounding*) and deconvoluted mutation profiles
2 were extracted from the resulting JSON files and converted into RC format. Starting from
3 RC files, normalized reactivity profiles were obtained by first calculating the raw reactivity
4 scores as the per-base ratio of the mutation count and the read coverage at each position
5 and by then normalizing values by box-plot normalization, using the *rf-norm* tool
6 (parameters: *-sm 4 -nm 3 -rb AC -mm 1 -n 1000*). Data-driven RNA structure inference was
7 performed using the *rf-fold* tool and the normalized reactivity profiles (parameters: *-sl 2.4 -*
8 *in -0.2 -nlp*). DMS-MaPseq data for the *cspA* 5' UTR folded in the presence of 0.05 μ M *cspA*
9 protein (accession: SRR6507969) was analyzed using the same parameters. Comparison
10 between the deconvoluted conformations and the *cspA* 5' UTR folded at either 10°C or 37°C
11 was performed using the *rf-compare* tool.

12

13 **Analysis of *V. vulnificus add* riboswitch DMS-MaPseq data.** Reads for DMS-MaPseq
14 data of *in vitro* folded *add* riboswitch from *V. vulnificus*, either in the presence or absence of
15 5 mM adenine, were obtained from the Sequence Read Archive (accessions: SRR10850890
16 and SRR10850891). Forward and reverse reads were merged prior to mapping using PEAR
17 v0.9.11¹⁸ and then mapped to the *add* riboswitch using the *rf-map* tool (parameters: *-cq5 20*
18 *-cqo -ctn -cmn 0 --mp '--very-sensitive-local'*). Resulting BAM files were then analyzed with
19 the *rf-count* tool to produce MM files (parameters: *-m -mm -na -ni*). MM files were analyzed
20 with DRACO (parameters: *--max-collapsing-windows 1 --set-all-uninformative-to-one --set-*
21 *uninformative-clusters-to-surrounding*) and deconvoluted mutation profiles were extracted
22 from the resulting JSON files and converted into RC format. Starting from RC files,
23 normalized reactivity profiles were obtained by first calculating the raw reactivity scores as
24 the per-base ratio of the mutation count and the read coverage at each position and by then
25 normalizing values by box-plot normalization, using the *rf-norm* tool (parameters: *-sm 4 -nm*
26 *3 -rb AC -mm 1 -n 1000*). Data-driven RNA structure inference was performed using the *rf-*
27 *fold* tool and the normalized reactivity profiles (parameters: *-sl 2.4 -in -0.2 -nlp*).

28

29 **Cell culture and SARS-CoV-2 infection**

30 Vero E6 cells were cultured in T-175 flasks in Dulbecco's modified Eagle's medium (DMEM;
31 Lonza, cat. 12-604F), supplemented with 8% fetal calf serum (FCS; Bodinco), 2 mM L-
32 glutamine, 100 U/mL of penicillin and 100 μ g/mL of streptomycin (Sigma Aldrich, cat. P4333-
33 20ML) at 37°C in an atmosphere of 5% CO₂ and 95%–99% humidity. Cells were infected

1 at a MOI of 1.5 with SARS-CoV-2/Leiden-0002 (GenBank accession: MT510999), a clinical
2 isolate obtained from a nasopharyngeal sample at LUMC, which was passaged twice in
3 Vero E6 cells before use. Infections were performed in Eagle's minimal essential medium
4 (EMEM; Lonza, cat. 12-611F) supplemented with 25 mM HEPES, 2% FCS, 2 mM L-
5 glutamine, and antibiotics. At 16 h post-infection, infected cells were harvested by
6 trypsinization, followed by resuspension in EMEM supplemented with 2% FCS, and then
7 washed with 50 mL 1X PBS.

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

10

11 **Total RNA extraction and *in vitro* folding**

12 Approximately 5×10^6 of the harvested infected cells were resuspended in 1 mL of TriPure
13 Isolation Reagent (Sigma Aldrich, cat. 11667157001) and 200 μ l of chloroform were added.
14 The sample was vigorously vortexed for 15 sec and then incubated for 2 min at room
15 temperature, after which it was centrifuged for 15 min at 12,500 x g (4°C). The upper
16 aqueous phase was collected in a clean 2 mL tube, supplemented with 1 mL (~2 volumes)
17 of 100% ethanol, and then loaded on an RNA Clean & Concentrator-25 column (Zymo
18 Research, cat. R1017). *In vitro* folding was carried out as previously described^{7,12}. Briefly,
19 ~5 μ g of total RNA from infected Vero E6 cells was first depleted of ribosomal RNAs using
20 the RiboMinus™ Eukaryote System v2 (ThermoFisher Scientific, cat. A15026), following
21 manufacturer instructions. Ribo- RNA in a volume of 39 μ l was denatured at 95°C for 2 min,
22 then transferred immediately to ice and incubated for 1 min. 10 μ l of ice-cold 5X RNA Folding
23 Buffer [500 mM HEPES pH 7.9; 500 mM NaCl] supplemented with 20 U of SUPERase•In™
24 RNase Inhibitor (ThermoFisher Scientific, cat. AM2696) were added. RNA was then
25 incubated for 10 min at 37°C to allow secondary structure formation. Subsequently, 1 μ l of
26 500 mM MgCl₂ (pre-warmed at 37°C) was added and RNA was further incubated for 20 min
27 at 37°C to allow tertiary structure formation.

28

29 **Probing of SARS-CoV-2 RNA**

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

1

2 **DMS-MaPseq analysis of SARS-CoV-2 RNA**

3 DMS-MaPseq of SARS-CoV-2 was conducted a previously described⁷, with minor changes.
4 First, probed RNA was fragmented to a median size of 150 nt by incubation at 94°C for 8
5 min in RNA Fragmentation Buffer [65 mM Tris-HCl pH 8.0; 95 mM KCl; 4 mM MgCl₂], then
6 purified with NucleoMag NGS Clean-up and Size Select beads (Macherey Nagel, cat.
7 744970), supplemented with 10 U SUPERase•In™ RNase Inhibitor, and eluted in 8 µl NF
8 H₂O. Eluted RNA was supplemented with 1 µl 50 µM random hexamers and 2 µl dNTPs (10
9 mM each), then incubated at 70°C for 5 min and immediately transferred to ice for 1 min.
10 Reverse transcription reactions were conducted in a final volume of 20 µl. Reactions were
11 supplemented with 4 µl 5X RT Buffer [250 mM Tris-HCl pH 8.3; 375 mM KCl; 15 mM MgCl₂],
12 1 µl DTT 0.1 M, 20 U SUPERase•In™ RNase Inhibitor and 200 U TGIRT™-III Enzyme
13 (InGex, cat. TGIRT50). Reactions were incubated at 25°C for 10 min to allow partial primer
14 extension, followed by 2 h at 57°C. TGIRT-III was degraded by addition of 2 µg Proteinase
15 K, followed by incubation at 37°C for 20 min. Proteinase K was inactivated by addition of
16 Protease Inhibitor Cocktail (Sigma Aldrich, cat. P8340). Reverse transcription reactions
17 were then used as input for the NEBNext® Ultra II Non-Directional RNA Second Strand
18 Synthesis Module (New England Biolabs, cat. E6111L). Second strand synthesis was
19 performed by incubating 1 h at 16°C, as per manufacturer instructions. DsDNA was purified
20 using NucleoMag NGS Clean-up and Size Select beads, and used as input for the
21 NEBNext® Ultra™ II DNA Library Prep Kit for Illumina, following manufacturer instructions.

22

23 **Analysis of SARS-CoV-2 DMS-MaPseq data.** After clipping adapter sequences using
24 Cutadapt v2.1¹⁹ (parameters: *-a AGATCGGAAGAGC -A AGATCGGAAGAGC -O 1 -m*
25 *100:100*), paired-end reads were merged using PEAR v0.9.11¹⁸ and then mapped to the
26 SARS-CoV-2 reference using the *rf-map* tool and the Bowtie2 algorithm, with soft-clipping
27 enabled, (parameters: *-b2 -cq5 20 -ctn -cmn 0 -cl 150 -mp '--very-sensitive-local'*). An MM
28 file was then generated from the resulting BAM alignment using the *rf-count* tool, by only
29 keeping reads covering at least 150 bases. Insertions and deletions were ignored (as they
30 account for less than 6% of DMS-induced mutations when using TGIRT-III⁶), considering
31 only mutations having Phred qualities > 20. Furthermore, mutations were only considered
32 when the two surrounding bases had Phred qualities > 20 as well. Reads with more than
33 10% mutated bases were excluded (parameters: *-m -ds 150 -es -nd -ni -mm -me 0.1*).

1 DRACO was invoked with default parameters. Following DRACO analysis, windows in
2 which the median coverage (calculated on reads passing DRACO's filtering) was above
3 10,000X were selected. To select windows that were coherently folding into multiple
4 conformations in both replicates, we retained windows predicted to have the same number
5 of conformations in the two replicates, overlapping by at least 75% of their length, and
6 considered only their intersection. Deconvoluted reactivity profiles for matching
7 conformations from the two replicates were then averaged and used for secondary structure
8 modeling. Correlation between reconstructed conformations from the two replicates were
9 calculated using 90% of the reactivity values in the window, after excluding the first and last
10 5% of the A/C bases, to avoid terminal biases.

11

12 **Identification of conserved RNA structure elements**

13 To evaluate the conservation of the alternative 3' UTR structure, we implemented a modified
14 version of an automated pipeline we have previously introduced¹² (*cm-builder*;
15 <https://github.com/dincarnato/labtools>), built on top of Infernal 1.1.3¹⁴. Briefly, we first built
16 two covariance models (CMs) from Stockholm files containing only the SARS-CoV-2
17 sequence and the two alternative 3' UTR structures, using the *cmbuild* module. After
18 calibrating the CMs using the *cmcalibrate* module, we used them to search for RNA
19 homologs in a database composed of all the non-redundant coronavirus complete genome
20 sequences from the ViPR database²⁰
21 (<https://www.viprbrc.org/brc/home.spg?decorator=corona>), as well as a set of
22 representative coronavirus genomes from NCBI database, using the *cmsearch* module.
23 Only matches from the sense strand were kept and a very relaxed E-value threshold of 10
24 was used at this stage to select potential homologs. Three additional filtering criteria were
25 used. First, we took advantage of the extremely conserved architecture of coronavirus
26 genomes²¹ and restricted the selection to matches falling at the same relative position within
27 their genome, with a tolerance of 3.5% (roughly corresponding to a maximum allowed shift
28 of 1050 nt in a 30 kb genome). Through this more "conservative" selection, we only kept
29 matches likely to represent true structural homologs, although at the cost of probably losing
30 some true matches. Second, we filtered out matches retaining less than 55% of the
31 canonical base-pairs from the original structure elements. Third, truncated hits covering
32 <50% of the structure were discarded. A fourth filtering step was also applied when
33 analyzing simultaneously the two structures, by retaining only the set of sequences matched

1 by both structures. The resulting set of homologs was then aligned to the original CMs using
2 the *cmalign* module and the resulting alignments were used to build new CMs. The whole
3 process was repeated for a total of 3 times. The alignment was then refactored, removing
4 gap-only positions and including only bases spanning the first to the last base-paired
5 residue. The alignment file was then analyzed using R-scape 1.4.0¹⁵ and APC-corrected G-
6 test statistics to identify motifs showing significantly covarying base-pairs.

7

8 **Testing for significant overlap with ORF boundaries**

9 To test for significant overlap between windows folding into two mutually-exclusive
10 conformations and ORF boundaries within the SARS-CoV-2 genome, we generated 10,000
11 random windows of matching size for each window identified by DRACO. For each DRACO-
12 identified window, as well as for each random window, we calculated the number of windows
13 overlapping the start/end positions of the SARS-CoV-2 ORFs, including each of the
14 individual proteins within the polyprotein ORF1a/b (positions: 266, 806, 2720, 8555, 10055,
15 10973, 11843, 12092, 12686, 13025, 13442, 13468, 16237, 18040, 19621, 20659, 21563,
16 25393, 26245, 26523, 27202, 27394, 27756, 27894, 28274, 29558, 29674). Resulting
17 values were used to perform a one-sided binomial test, with parameters $k = 11$ (number of
18 windows identified by DRACO, overlapping with ORF boundaries), $n = 22$ (total number of
19 windows identified by DRACO), and $p =$ ratio between the number of random windows
20 overlapping with ORF boundaries, divided by the total number of random windows
21 (220,000).

22

23 **Validation of the alternative SARS-CoV-2 3' UTR conformation by COMRADES**

24 COMRADES data for the SARS-CoV-2 virus in living infected host cells¹⁶ was obtained from
25 GEO (GSE154662). The dataset consisted of 2 biological replicates, each one composed
26 of a control (C) and the actual COMRADES sample (S). A reference was built on all human
27 transcripts from refGene, plus the sequence of the SARS-CoV-2 genome, using STAR
28 v2.7.1a²² (parameters: `--runMode genomeGenerate --genomeSAindexNbases 12`), and
29 reads were aligned to the reference using the same (parameters: `--runMode alignReads --`
30 `outFilterMultimapNmax 100 --outSAMattributes All --alignIntronMin 1 --scoreGapNoncan -4`
31 `--scoreGapATAC -4 --chimSegmentMin 15 --chimJunctionOverhangMin 15`). Resulting
32 alignments (as well as chiasitic alignments from the junctions file) were filtered, discarding
33 ungapped reads, reads having more than one gap, and reads aligning to the human

1 transcriptome, and the total number of reads per experiment was calculated (C_{tot} and S_{tot}).
2 Each chimeric read was described as a set of 2 numeric intervals (I1 and I2), corresponding
3 to the two halves of the chimera. To assess whether a base-pair $i-j$ was enriched in the
4 COMRADES sample with respect to the control sample, we calculated the number of reads
5 in which base i overlapped interval I1 and base j overlapped interval I2, for both samples
6 (C_{i-j} and S_{i-j}). Significance of the enrichment was then assessed using a one-tailed binomial
7 test, with parameters $k = S_{i-j}$, $n = S_{tot}$, and $p = C_{i-j} / C_{tot}$. Only base-pairs with p -value < 0.05
8 in both replicates were considered to have *in vivo* support.

9

10 **Data availability.** Sequencing data has been deposited to the Gene Expression Omnibus
11 (GEO) database, under the accession GSE158052. Additional processed files are available
12 at http://www.incarnatolab.com/datasets/DRACO_Morandi_2020.php.

13

14 **Acknowledgements**

15 D.I. was supported by the Dutch Research Council (NWO) as part of the research
16 programme NWO Open Competitie ENW - XS with project number OCENW.XS3.044 and
17 by the Groningen Biomolecular Sciences and Biotechnology Institute (GBB, University of
18 Groningen). S.O. was supported by the Associazione Italiana per la Ricerca sul Cancro
19 (AIRC), grant AIRC IG 2017 Id. 20240, and PRIN 2017. M.J.H. was supported by the Leiden
20 University Fund (LUF), the Bontius Foundation and donations from the crowdfunding
21 initiative "wake up to corona".

22

23 **Author contributions**

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

27

28 **Competing interests**

29 The authors declare no competing interests.

30

1 **References**

- 2
- 3 1. Incarnato, D. & Oliviero, S. The RNA Epistrurome: Uncovering RNA Function by
4 Studying Structure and Post-Transcriptional Modifications. *Trends in biotechnology* **35**,
5 318–333 (2017).
- 6 2. Strobel, E. J., Yu, A. M. & Lucks, J. B. High-throughput determination of RNA structures.
7 *Nature reviews. Genetics* **19**, 615–634 (2018).
- 8 3. Spasic, A., Assmann, S. M., Bevilacqua, P. C. & Mathews, D. H. Modeling RNA secondary
9 structure folding ensembles using SHAPE mapping data. *Nucleic Acids Res* **46**, 314–323
10 (2017).
- 11 4. Li, H. & Aviran, S. Statistical modeling of RNA structure profiling experiments enables
12 parsimonious reconstruction of structure landscapes. *Nat Commun* **9**, 606 (2018).
- 13 5. Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E. & Weeks, K. M. RNA motif
14 discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods* **11**, 959–965
15 (2014).
- 16 6. Zubradt, M. *et al.* DMS-MaPseq for genome-wide or targeted RNA structure probing in
17 vivo. *Nature Methods* **14**, 75–82 (2017).
- 18 7. Simon, L. M. *et al.* In vivo analysis of influenza A mRNA secondary structures identifies
19 critical regulatory motifs. *Nucleic Acids Res* **47**, 7003–7017 (2019).
- 20 8. Homan, P. J. *et al.* Single-molecule correlated chemical probing of RNA. *Proceedings of*
21 *the National Academy of Sciences of the United States of America* **111**, 13858–13863
22 (2014).
- 23 9. Tomezsko, P. J. *et al.* Determination of RNA structural diversity and its role in HIV-1 RNA
24 splicing. *Nature* **582**, 438–442 (2020).
- 25 10. Zhang, Y. *et al.* A Stress Response that Monitors and Regulates mRNA Structure Is
26 Central to Cold Shock Adaptation. *Molecular cell* **70**, 274-286.e7 (2018).
- 27 11. Giuliadori, A. M. *et al.* The cspA mRNA Is a Thermosensor that Modulates Translation
28 of the Cold-Shock Protein CspA. *Mol Cell* **37**, 21–33 (2010).
- 29 12. Manfredonia, I. *et al.* Genome-wide mapping of SARS-CoV-2 RNA structures identifies
30 therapeutically-relevant elements. *Nucleic Acids Res* gkaa1053- (2020)
31 doi:10.1093/nar/gkaa1053.
- 32 13. Lan, T. C. T. *et al.* Structure of the full SARS-CoV-2 RNA genome in infected cells.
33 *Biorxiv* 2020.06.29.178343 (2020) doi:10.1101/2020.06.29.178343.
- 34 14. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches.
35 *Bioinform Oxf Engl* **29**, 2933–5 (2013).

- 1 15. Rivas, E., Clements, J. & Eddy, S. R. A statistical test for conserved RNA structure
2 shows lack of evidence for structure in lncRNAs. *Nat Methods* **14**, 45–48 (2017).
- 3 16. Ziv, O. *et al.* The short- and long-range RNA-RNA Interactome of SARS-CoV-2. *Mol Cell*
4 (2020) doi:10.1016/j.molcel.2020.11.004.
- 5 17. Incarnato, D., Morandi, E., Simon, L. M. & Oliviero, S. RNA Framework: an all-in-one
6 toolkit for the analysis of RNA structures and post-transcriptional modifications. *Nucleic*
7 *Acids Research* **46**, e97–e97 (2018).
- 8 18. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina
9 Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
- 10 19. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing
11 reads. *EMBnet.journal* **17**, 10–12 (2011).
- 12 20. Pickett, B. E. *et al.* ViPR: an open bioinformatics database and analysis resource for
13 virology research. *Nucleic Acids Res* **40**, D593-8 (2011).
- 14 21. Lauber, C. *et al.* The footprint of genome architecture in the largest genome expansion
15 in RNA viruses. *Plos Pathog* **9**, e1003500 (2013).
- 16 22. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21
17 (2013).
- 18

1 **Figure legends**

2

3 **Figure 1. Overview of the DRACO algorithm.** (a) Schematic representation of the DRACO
4 algorithm. (b) Maximum number of conformations detected for 10 sets of 100 simulated
5 RNAs, with length ranging from 300 to 1500 nt, expected to form 1 to 4 conformations, at a
6 coverage of 5,000X and a read length of 150 nt. Error bars represent SD of the 10 sets. (c)
7 Box-plot of median Pearson correlation coefficients (PCC) of reconstructed reactivity profiles
8 for 10 sets of 100 simulated 1000 simulated RNAs, with length ranging from 300 to 1500 nt,
9 expected to form 1 to 4 conformations, at a coverage of 20,000X and a read length of 150
10 nt. When DRACO detected more than one window with different numbers of clusters, only
11 the largest window, spanning >50% of the RNA length, was considered. (d) Violin plot
12 depicting the distribution of expected versus reconstructed conformation abundances for 10
13 sets of 100 simulated RNAs, with length ranging from 300 to 1500 nt, expected to form 1 to
14 4 conformations, at a coverage of 20,000X and a read length of 150 nt. The Pearson
15 correlation is indicated in the bottom-right corner of each plot.

16

17 **Figure 2. *In vitro* validation of DRACO.** (a) Original DMS-MaPseq profile, and DRACO-
18 deconvoluted profiles for *cspA* 5' UTR folded at 10°C in the presence of 50 μM *cspA*
19 recombinant protein, from Zhang *et al.*, 2018¹⁰. Schematic representation of the structures,
20 and the reconstructed relative abundances are indicated. (b) Heatmap of Pearson
21 correlation coefficients showing the correlation between the conformations deconvoluted by
22 DRACO, and the reactivity profiles of the *cspA* 5' UTR folded at either 10°C or 37°C, in the
23 absence of the recombinant *cspA* protein. (c) Arc plots depicting the secondary structure
24 inferred from the DRACO-deconvoluted profiles, as compared to the reference *cspA* 5' UTR
25 structures at 10°C and 37°C. Positive predictive value (PPV) and sensitivity are indicated).
26 (d) DRACO-deconvoluted profiles for *V. vulnificus add* riboswitch, in the absence (1
27 conformation detected) or presence (2 conformations detected) of 5 mM adenine, from
28 Tomezsko *et al.*, 2020⁹. Schematic representation of the structures, and the reconstructed
29 relative abundances are indicated. (e) Heatmap of Pearson correlation coefficients showing
30 the correlation between the conformations deconvoluted by DRACO. (f) Arc plots depicting
31 the secondary structure inferred from the DRACO-deconvoluted profiles, as compared to
32 the reference *add* structure in the absence of adenine.

33

1 **Figure 3. A conserved structural switch in the 3' UTR of SARS-CoV-2.** (a) Relative
2 abundances of the two alternative conformations (A and B) of the SARS-CoV-2 3' UTR.
3 Error bars represent the standard deviation from two replicates. (b) Heat scatterplot of base
4 reactivities for DRACO-deconvoluted reactivity profiles in replicate #1 versus replicate #2 of
5 conformation A (left) and B (right). Base-pairs whose existence is supported by significant
6 enrichment of RNA-RNA chimeras from *in vivo* COMRADES analysis (Ziv *et al.*, 2020¹⁶) are
7 boxed in light blue. (c) Secondary structure models with overlaid base reactivities for
8 conformation A and B. (d) Structure models for conformation A (top) and B (bottom), inferred
9 by simultaneous phylogenetic analysis. Structures have been generated using the R2R
10 software. Base-pairs showing significant covariation (as determined by R-scape) are boxed
11 in green (E-value < 0.05) and violet (E-value < 0.1) respectively.