sgDI-tector: defective interfering viral genome bioinformatics for detection of coronavirus subgenomic RNAs

Andrea Di Gioacchino, Rachel Legendre, Yannis Rahou, Valérie Najburg, Pierre Charneau, Benjamin Greenbaum, Frédéric Tangy, Sylvie van Der Werf, Simona Cocco, Anastassia Komarova

sgDI-tector: defective interfering viral genome bioinformatics for detection of coronavirus subgenomic RNAs

ANDREA DI GIOACCHINO,[1] RACHEL LEGENDRE,[2] YANNIS RAHOU,[3] VALÉRIE NAJBURG,[4] PIERRE CHARNEAU,[5] BENJAMIN D. GREENBAUM,[6,7] FRÉDÉRIC TANGY,[4] SYLVIE VAN DER WERF,[3] SIMONA COCCO,[1] and ANASTASSIA V. KOMAROVA[3]

[1]Sorbonne Université, Université de Paris, Laboratoire de Physique de l'Ecole Normale Supérieure, PSL & CNRS UMR8063, 75005, Paris, France
[2]Institut Pasteur, Université de Paris, Hub de Bioinformatique et Biostatistique - Département Biologie Computationnelle, 75015, Paris, France
[3]Institut Pasteur, Université de Paris, CNRS UMR3569, Génétique Moléculaire des Virus à ARN, F-75015 Paris, France
[4]Institut Pasteur, Université de Paris, Laboratory of Innovative Vaccines, 75015, Paris, France
[5]Institut Pasteur, Pasteur-TheraVectys joined unit, 75015, Paris, France
[6]Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York 10065, USA
[7]Physiology, Biophysics and Systems Biology, Weill Cornell Medicine, Weill Cornell Medical College, New York 10065, USA

ABSTRACT

Coronavirus RNA-dependent RNA polymerases produce subgenomic RNAs (sgRNAs) that encode viral structural and accessory proteins. User-friendly bioinformatic tools to detect and quantify sgRNA production are urgently needed to study the growing number of next-generation sequencing (NGS) data of SARS-CoV-2. We introduced sgDI-tector to identify and quantify sgRNA in SARS-CoV-2 NGS data. sgDI-tector allowed detection of sgRNA without initial knowledge of the transcription-regulatory sequences. We produced NGS data and successfully detected the nested set of sgRNAs with the ranking M > ORF3a > N>ORF6 > ORF7a > ORF8 > S > E>ORF7b. We also compared the level of sgRNA production with other types of viral RNA products such as defective interfering viral genomes.

Keywords: subgenomic RNA; SARS-CoV-2; defective viral genomes; user-friendly bioinformatics

INTRODUCTION

Viral RNA-dependent RNA polymerases (RdRp) ensure multiple molecular mechanisms to produce a large spectrum of viral RNA products inside infected cells. Some of these molecular mechanisms have already been described in detail and others are yet to be uncovered. Mechanisms of RNA virus replication and transcription, cap-snatching, and RNA editing have been relatively well described (Strauss and Strauss 2007), whereas molecular mechanisms underlying defective viral genome (DVG) production have yet to be discovered. DVGs are truncated forms of and/or rearranged viral genomes generated by most viruses during viral replication. DVG can also be called defective interfering (DI) genomes when viral particles containing them are able to interfere with standard virus replication (Pathak and Nagy 2009). DVG and viral genomes share the minimum essential characteristics for replication: a competent initiation site at the 3′-end and its complementary sequence at the 5′-end. However, DVG genomes are defective for replication in the absence of the complete functional virus genome that provides the missing functions. Four main classes of DVGs exist: deletions, insertions, snap-back DI genomes or "hairpin" structures, and copy-back or "panhandle" structure DI genomes (see Fig. 1; Lazzarini et al. 1981; Dimmock et al. 2014).

SARS-CoV-2 is an enveloped, positive-sense, single-stranded RNA virus with a genome of nearly 30,000 nts (Lu et al. 2020). Similar to other coronaviruses, SARS-CoV-2 replication involves the synthesis by the viral RdRp of positive and negative sense full-length genomes as well as the production of a nested set of subgenomic RNA (sgRNAs). SARS-CoV-2 sgRNAs encode four structural

Corresponding authors: anastasia.komarova@pasteur.fr, andrea.digioacchino@phys.ens.fr

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.078969.121. Freely available online through the *RNA* Open Access option.

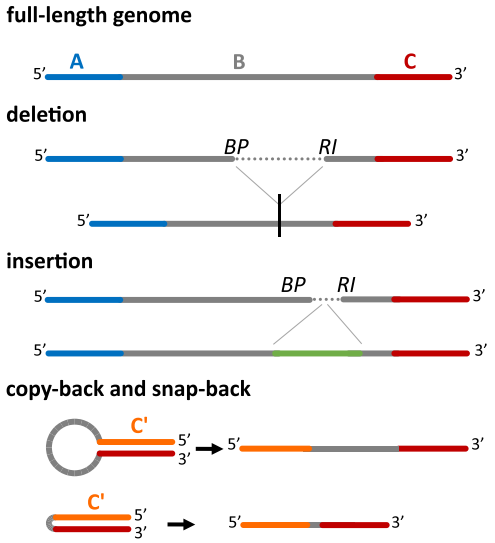

FIGURE 1. Four main classes of DI genomes can be detected by DI-tector. The full-length genome (divided here in three regions, A, B and C) is shown first. When a part of the region B is missed then DI-tector detects a deletion; if instead a region is added, DI-tector detects an insertion. Copy-backs and snap-backs are formed through junctions involving the two strands (positive-sense and negative-sense). C′ is the region complementary to C. (BP) Breakpoint site, (RI) reinitiation site.

proteins (S, spike; E, envelope; M, membrane; N, nucleocapsid) and several accessory factors (ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8, and ORF10) (Davidson et al. 2020; Kim et al. 2020; Wu et al. 2020; Zhou et al. 2020).

sgRNAs are produced by the yet to be determined complex mechanisms assigned as discontinuous transcription that includes two steps (Pasternak et al. 2006). The first consists in the viral RdRp pausing the negative-sense RNA synthesis at a specific 6–7 nt in length sequence (body transcription regulatory sequence, TRS-B) at the 3′-end of the viral genome and then performing a long-range jump at the 5′-end of the genome to join a common sequence of ~70 nt encompassing another, identical, 6–7 nt in length sequence. This second hexanucleotide sequence is named leader transcription-regulatory sequence (TRS-L), and the part of the viral genome starting at the 5′-end and ending just after the TRS-L is called leader sequence. In our study we use generic TRS to indicate short sequences which are found both in the final part of the leader sequence, and in the 3′ part of the genome around the site of the junction. The second step is the replication of the positive sense sgRNA from the negative-sense RNA template produced at the first step. This way, the molecular organization of coronavirus sgRNAs is similar to the deletion type of the DI genomes, and methods applied for DI genome detection in NGS data become suitable for the detection of coronavirus sgRNA. However, the subgenomic biogenesis mechanism is still under intensive study, and several important questions have been addressed only very recently: For instance, it has been suggested that in addition to the full-length SARS-CoV-2 genomic template, the sgRNAs themselves can give rise to shorter sgRNA, through additional RdRp pause-and-jump events (Wang et al. 2021). Another question that has been addressed recently, with a particular focus on SARS-CoV-2, regards the role in sgRNA expression of the secondary structure within the 5′ untranslated region (Sola et al. 2015; Miao et al. 2021).

Several sgRNA-oriented NGS studies have already provided various ratios of the nested SARS-CoV-2 sgRNA concentrations. Kim et al.'s study applied nanopore direct RNA sequencing validated by DNA nanoball sequencing (MGI NGS platform) to demonstrate that Vero cells infection with SARS-CoV-2 produces a complexed transcriptome with nine canonical sgRNAs and noncanonical viral ORFs (Kim et al. 2020). To detect sgRNA reads from short length NGS data, Kim et al. applied Spliced Transcripts Alignment to a Reference (STAR) free open source software (Dobin et al. 2012), and performed a quantitative analysis of sgRNA transcription. They revealed that N RNA was the most abundantly expressed transcript, followed by S, 7a, 3a 8, M, E, 6, and 7b RNAs (Kim et al. 2020). Finkel et al. used two approaches to calculate the abundance of sgRNAs in NGS data from SARS-CoV-2 infected cell cultures (Finkel et al. 2021). The first was the STAR-based assessment of the relative abundances of RNA reads spanning leader–body junctions for the canonical sgRNAs. The second approach used deconvolution of RNA densities. In the deconvolution or "decumulation" approach, the RNA expression of each ORF is calculated by subtracting the RNA-read density upstream of the ORF region (inter-TRS region) (Irigoyen et al. 2016). Finkel et al.'s study has described the N transcript as the most abundant followed by M, ORF7a and ORF3a. Finally, the first bioinformatic pipeline for detection and quantification of sgRNA specifically in SARS-CoV-2 genomic sequencing data has been proposed and called Periscope (Parker et al. 2021). Periscope deals with various types of SARS-CoV-2 sequencing analysis including ARTIC Nanopore-generated and Illumina metagenomic sequencing. In order to identify sgRNAs, Periscope requires previous knowledge of the coronavirus leader sequence (Parker et al. 2021). However, the above approaches are rather complex, requiring several intermediate steps or previous experience with a number of other bioinformatic tools, especially when applied to SARS-CoV-2 short-read NGS data to detect and characterize sgRNAs.

We have recently developed a user-friendly bioinformatic pipeline called DI-tector to detect various types of DVGs from NGS data (Beauclair et al. 2018). In this report, we extend the functionality of DI-tector by introducing sgDI-tector, a tool that, after running DI-tector, can properly detect and quantify coronavirus sgRNAs without previous knowledge of the TRS. We use the output of sgDI-

tector to compute the SARS-CoV-2 sgRNA ratios and discuss the comparison with results obtained through previously developed approaches (Kim et al. 2020; Finkel et al. 2021; Parker et al. 2021), showing its largest robustness and sensitivity. To confirm sgDI-tector's potential to detect rare noncanonical sgRNA populations from NGS data, the RT-qPCR approach was applied. Moreover, contrary to other methods, sgDI-tector does not impose a junction sequence, allowing us to investigate the sequences found at the junction and their variability.

RESULTS

DI-tector detects various types of SARS-CoV-2 DI genomes

First, we generated an NGS data set on total RNA from human cells (HEK293 transduced with ACE) infected with SARS-CoV-2. Then we ran DI-tector on the RNA-seq data to characterize and quantify DVGs. The results of RNA-seq and alignment to reference genomes, and those of DI-tector, are presented in Figure 2 (the raw number of counts used for this figure are provided in Supplemental Table 1). We observed a relevant number of reads mapped to the SARS-CoV-2 genome (respectively 33%, 40%, and 40% in the three biological replicates, see Supplemental Table 1), and a much smaller quantity of DVG reads. The most represented type of DVG read was deletions, which accounted for ~74% of the DVG reads (Table 1). While expected, given the coronaviruses' mechanism of production of sgRNA, this large fraction of deletions suggests that

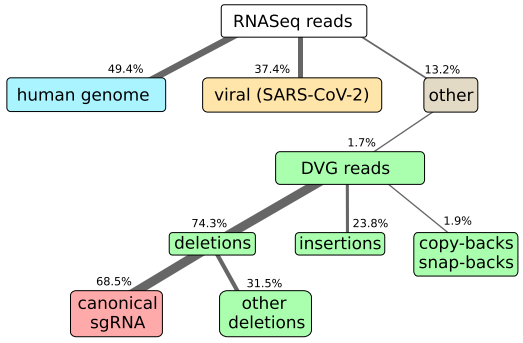

FIGURE 2. Most of the DVG reads can be associated to canonical sgRNAs. Here we show the results of RNA-seq and alignment of the reads to the human and SARS-CoV-2 genomes. NGS library preparation was performed with a ribodepletion step. The unmapped reads have been further processed with DI-tector, and the resulting characterization of the DVG reads into deletions, insertions and copy-backs/snap-backs is given. The percentage of junction reads corresponding to canonical sgRNA that is standard annotated subgenomic ORFs for SARS-CoV-2 (S, 3A, E, M, 6, 7a, 7b, N, 10), is specified. All percentages are averaged over three biological replicates. Cells colors are given to classify reads in host-related reads (blue), viral reads (yellow), DVG reads (green), canonical sgRNA reads (red), other reads (gray).

TABLE 1. Overview on DVG observed in this study

	Number of junctions (percentage)	Number of unique junctions
Deletions (sgRNA)	34,957 (51%)	169
Deletions (others)	15,772 (23%)	8604
Insertions	16,341 (24%)	8657
Copy-backs	1294 (2%)	655

All the values are obtained as averages of three biological replicates. Unique junctions are defined as different BP-RI positions spanned by observed reads. Copy-back DVGs also include snap-back DVGs.

DI-tector results can be used to identify and quantify viral sgRNA from NGS data in a simple and efficient way. Indeed, we were able to associate 68% of these DVG reads to canonical SARS-CoV-2 sgRNA transcripts (coding for ORFs: S, 3A, E, M, 6, 7a, 7b, N) (see Materials and Methods for details about the pipeline used). We also found a quite large number of insertions, accounting for ~24% of the total DVG reads. Finally, we observed ~2% of the total DVG of the copy-back or snap-back type of DI genomes in SARS-CoV-2 infected cells (Table 1; Supplemental Table 1).

DVGs of deletion type can be used to characterize the nested set of sgRNAs

The leader-body junctions formed during the transcription of sgRNA in coronaviruses are detected by DI-tector as deletion DVGs, and the importance of such sgRNAs is reflected by their abundance with respect to other DVG types. We assessed here the possibility of using DI-tector to characterize the nested set of ORFs transcribed as sgRNAs, and to compare the levels of expression of different sgRNAs. Our pipeline, which we named sgDI-tector, starts from the DI-tector output and is based on two assumptions: (i) sgRNA coding for expressed ORFs are transcribed more frequently (on the overall) than classical DVGs; and (ii) there is a leader sequence shared among all sgRNAs. The detailed pipeline is described in the Materials and Methods section. We stress, however, that differently from other methods we do not need to explicitly know the leader/junction TRS to apply our algorithm, but the fact that there is one is necessary for the algorithm to work. We run our analysis by using three biological replicates, and all of the results presented here are almost unchanged in each of them (Fig. 3). As can be seen in Figure 3, in each of our samples we observed clear signals of several ORFs, in particular M, 3a, N, 6, 7a, 8, S, and E. We also found some signal of direct transcription of ORF 7b, which could also be translated from the same sgRNA coding for ORF 7a as suggested for SARS-CoV in Schaecher et al. (2007). Moreover, we observed a leader-

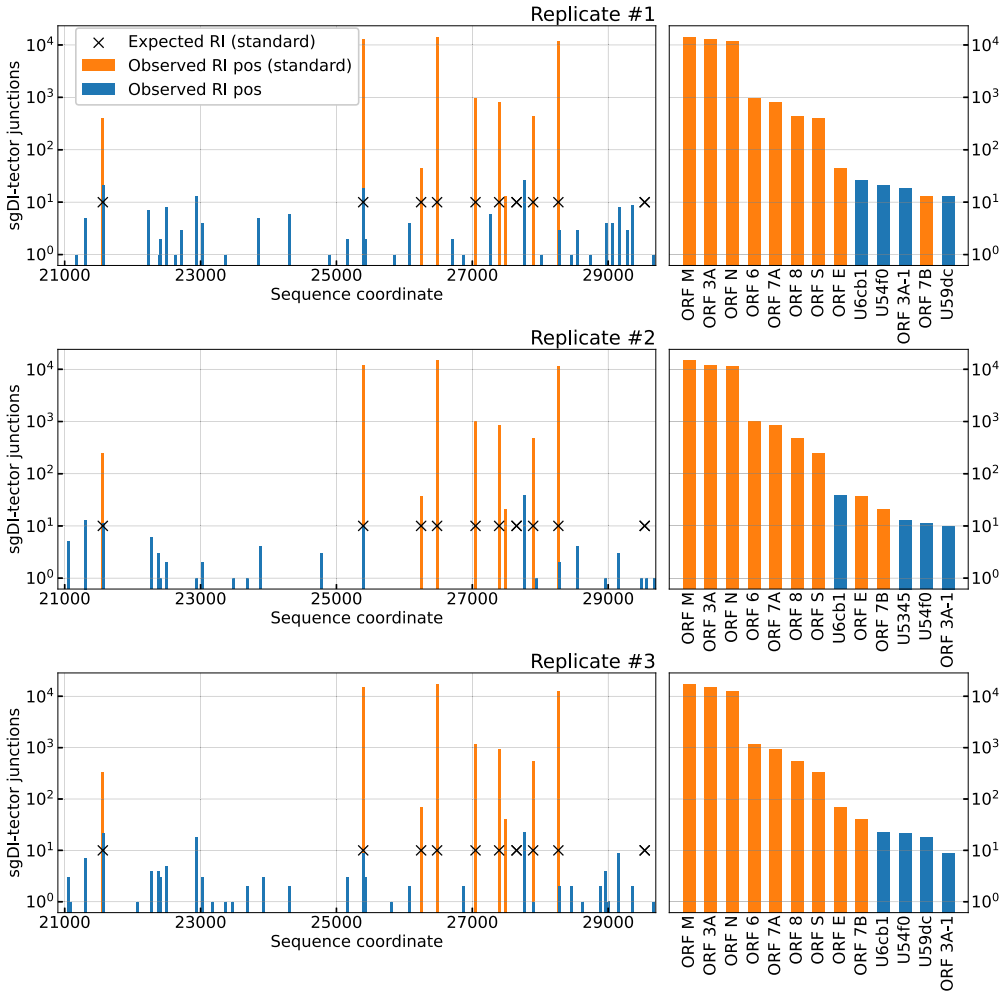

FIGURE 3. sgDI-tector detects most canonical sgRNAs in all replicates with a high number of counts. (*Left* panels) Deletion DVGs distribution across the last 10 kb positions of the SARS-CoV-2 genome (GISAID ID: EPI_ISL_414631). Each blue or orange bar corresponds to a deletion, the position of the bar being the starting point of the "body" part of the junction (called RI position in sgDI-tector). Crosses are expected RI positions from Alexandersen et al. (2020), and bars are colored in orange if the deletion is observed in that position in our data. (*Right* panels) Number of counts and ORF name for the 13 deletions with most counts observed in each replicate. Orange bars correspond to orange crosses in the *left* panel and represent canonical TRS. Blue bars correspond to putative noncanonical ORFs detected in our data. Names for noncanonical ORFs are hexadecimal numbers representing the position of the corresponding start codon (AUG) in the standard 5′-to-3′ sense in the reference sequence (GISAID ID: EPI_ISL_414631), see also Supplemental Table 2.

body junction resulting in a potential expression of a noncanonical ORF. Remarkably, its number of counts was comparable with ORF E and larger than ORF 7b when counts were averaged over replicates, see Figure 4. This ORF, which is referred to as "U6cb1" in Figure 3 and Figure 4 (in Supplemental Table 2 we reported the AUG codon position in our reference SARS-CoV-2 genome, for each noncanonical sgRNA we detected), codes for a 20 amino acid long protein and has been previously identified in Finkel et al. (2021) and referred to as 7b.iORF1.

Additional leader-body junctions in our data gave rise to different ORFs which were found in all three biological replicates. Those found with highest count numbers are presented in Figure 3 and Figure 4 (blue bars). To provide a control for false positives using sgDI-tector, we analyzed as input the RNA-seq data from mock-infected HEK293-ACE2 cells. sgDI-tector did not give any sgRNA read in this case, confirming the robustness and sensitivity of our approach.

Validation of noncanonical sgRNA produced inside ORF1ab

We used the RT-qPCR approach in order to test the accuracy of our sgDI-detector algorithm. We validated the noncanonical ORF U3dc4 that has its body TRS sequence located in ORF1ab. ORF U3dc4 was less present in our data than the majority of other noncanonical ORFs. Of note, U3dc4 was previously detected in the Kim et al. (2020) study by the STAR approach, but to our knowledge was never

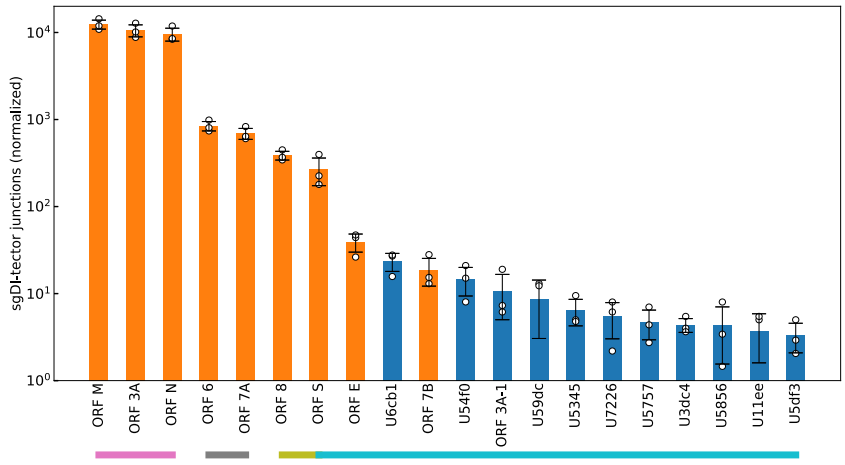

FIGURE 4. Canonical sgRNAs and some noncanonical sgRNAs are consistently observed across the three biological replicates. ORF names and numbers of counts of the 20 deletions with the most counts were observed in NGS data in three biological replicates. Data from different replicates have been normalized so that the number of viral reads observed in each replicate is constant (see Materials and Methods). Bar heights are given by the average of the three replicates (shown as white dots after normalization), and error bars represent the standard deviation. The colored bars *below* the ORF names indicate statistical significance of the count differences; ORFs *above* bars of different colors have statistically different junction counts (*P*-value ≤0.05, from a two-sample, two-tailed, Welch's unequal-variance *t*-test).

validated by a conventional approach. We performed RT-qPCR analysis on total RNA extracted from HEK293 (transduced with ACE) infected with SARS-CoV-2 or mock-infected (negative control) cells. As expected, ORF U3dc4 was detected only in RNA extracted from infected cells similar to the detection of canonical ORF encoding the N protein (Table 2; Supplemental Table 3). These experiments validate the presence of ORF U3dc4 in SARS-CoV-2-infected cells by a RT-qPCR approach and thus highlight the power of sgDI-tector to reveal the landscape of the sgRNA population from NGS data.

Comparison of sgDI-tector results on our data with existing bioinformatic tools

Several techniques have been used so far to detect sgRNA expression levels in coronaviruses. Firstly, the reads per kilobase of transcript per million mapped reads (RPKM) have been directly used for several coronaviruses, once the RPKM of each sgRNA is properly "decumulated" from downstream ORFs, which are present in each ORF transcript (Irigoyen et al. 2016). Despite its simplicity, we show in Figure 5A that, in our case, the decumulation approach gave results which failed to correlate with those obtained with the junction analysis performed by sgDI-tector. Moreover, the decumulation resulted in several ORFs having a negative number of counts, and this can happen for two reasons: the (unavoidable) errors in the estimation of transcription levels from the NGS data (especially for short ORFs) and the presence of other DVGs which are not considered in the decumulation procedure. Similar results have been independently reported in Finkel et al. (2021). We believe that the absence of correlation between junction counts and decumulated RPKM in our sample cannot be explained by a failure of sgDI-tector and/or of our pipeline. Indeed we recovered, with the same pipeline, a much higher correlation between junction counts and decumulated RPKM in another NGS data set, collected by Finkel et al. (2021) (see Supplemental Fig. 1). In Figure 5B,C we compare sgDI-tector with other tools which exploit chimeric reads to search for putative sgRNA junctions: Periscope (Parker et al. 2021) and STAR 2.7.3a as used in Kim et al. (2020), Finkel et al. (2021) and Wang et al. (2021). The most abundant sgRNAs (M, 3A, and N) are detected by all tools with many counts, although the ranking of them is different in the output of STAR. Periscope, on the other hand, cannot detect at all ORF E, and ORF 7b (ORF 6 is detected with only one count), suggesting that our tool might be more accurate for Illumina data than Periscope, which has been developed to deal with ARTIC Nanopore-generated sequencing data. DI-tector and STAR results are almost perfectly correlated for most of the sgRNAs. However, junctions associated with ORF 8, and most of the junctions associated with ORF 3A, were detected by STAR only when NGS reads mapped on the negative-sense viral genome were included in the data analysis, thus after performing manual curation of the data. The only ORF for which there is a sensible difference in sgDI-tector's and STAR's results is ORF 3A, as STAR can detect significantly fewer junctions with respect to sgDI-tector (see Table 3). Although it is difficult to clearly assess whether the error is on the STAR or sgDI-tector side, the former hypothesis seems more supported since

TABLE 2. RT-qPCR validation of U3dc4 sgRNA transcript

Target detected	SARS-CoV-2	Noninfected
U3dc4	27.0 ± 0.1	ND
ORF N	17.0 ± 0.1	ND
GAPDH	20.65 ± 0.05	20.6 ± 0.1

RT-qPCR C_t values for U3dc4, ORF N, and GAPDH detection in cDNA equivalent of 50 ng of total RNA extracted from SARS-CoV-2 or mock-infected HEK293T cells are shown. Samples were analyzed in duplicates. ND: Not determined ($C_t > 34$). Data are given as average ± standard deviation. Results obtained from the other biological replicates are given in Supplemental Table 4.

TABLE 3. Comparison of the sgRNA abundance obtained by sgDI-tector, STAR, and Periscope

	STAR	sgDI-tector	Periscope
ORF S	220 ± 74	267 ± 94	85 ± 16
ORF 3A	4239 ± 624	10573 ± 1665	8663 ± 1375
ORF E	41 ± 7	39 ± 9	0 ± 0
ORF M	12175 ± 1461	12396 ± 1466	12207 ± 1433
ORF 6	817 ± 103	843 ± 104	0.6 ± 0.4
ORF 7A	695 ± 104	691 ± 99	188 ± 35
ORF 7B	16 ± 7	19 ± 7	0 ± 0
ORF 8	382 ± 46	387 ± 45	290 ± 31
ORF N	9633 ± 1640	9578 ± 1619	5767 ± 806

We compare the average and standard deviation of the counts observed by sgDI-tector, STAR and Periscope for each canonical sgRNA junction (given as mean ± standard deviation from our three biological replicates). A color code is used to indicate the statistical significance as follows: A cell has a green background if it is not statistically different from at least another tool, an orange background if it is statistically different from the other tools, and a red background if no junctions are found in the three replicates. ORF 10 is not present as none of the three tools detected the corresponding sgRNA junctions. The *P*-values used to assign colors are obtained through two-sample, two-tailed, Welch's unequal-variance *t*-tests. Counts from different replicates have been normalized so that the number of SARS-CoV-2 reads observed in each replicate is constant (see Materials and Methods).

the junction counts of ORF 3A obtained by STAR is statistically different from both the counts obtained by sgDI-tector and Periscope, while the junction counts detected by sgDI-tector is compatible with the result obtained from Periscope. In Table 3 we present a more systematic comparison among the three tools based on the abundance ranking of the canonical sgRNA, whose outcome is that in the only case where the results of sgDI-tector and STAR were not compatible, Periscope's result was compatible with sgDI-tector's result and not with STAR's result. In Figure 5D–F, we compared our results with those obtained through STAR for two other RNA-sequencing data sets, (Kim et al. 2020; Finkel et al. 2021). In both publications the authors infected Vero cells, but with different protocols: Finkel et al. used a multiplicity of infection (MOI) of 0.2, harvested cells after 5 h post-infection (hpi) and 24 hpi, and the sequencing was conducted on the Illumina Miseq platform; Kim et al. used a MOI of 0.05, harvested cells 24 hpi, and used nanopore and nanoballs RNA sequencing. It is apparent that ORF N is consistently one of the most transcribed sgRNAs. However, several variations between experiments can be noticed: For instance, ORF E seems to be more transcribed in Finkel et al.'s analysis (Finkel et al. 2021), and ORF S is the second most transcribed sgRNA in Kim et al.'s analysis (Kim et al. 2020). In addition to these differences, it is apparent that for most of the ORFs HEK293 cells present less junctions than those observed in previous experiments (see dashed lines in Fig. 5D–F) performed on Vero cells. Differences in efficiency of infection of the Vero cell line, that is routinely used to amplify SARS-CoV-2, and of ACE-transduced HEK293 (ST-CH^{ACE-2}) cells could explain the observed inequality in number of junctions.

Applying sgDI-tector on previously published SARS-CoV-2 NGS data sets

Next, we run our tool starting from the raw RNA-seq data obtained by Finkel et al. (2021) to make a comparison with their results, which is presented in Figure 6. We observed that the results obtained by sgDI-tector and STAR 2.7.3a are well correlated, showing the effectiveness of the approach proposed here to quantify sgRNA transcription from NGS data. The results of the same test done with Periscope showed a much lower correlation with the original Finkel et al.'s junction counts. In particular, Periscope completely missed junctions with sgRNAs 6, E, and 7b (they are shown with 1 pseudocount in Fig. 6C,D) at

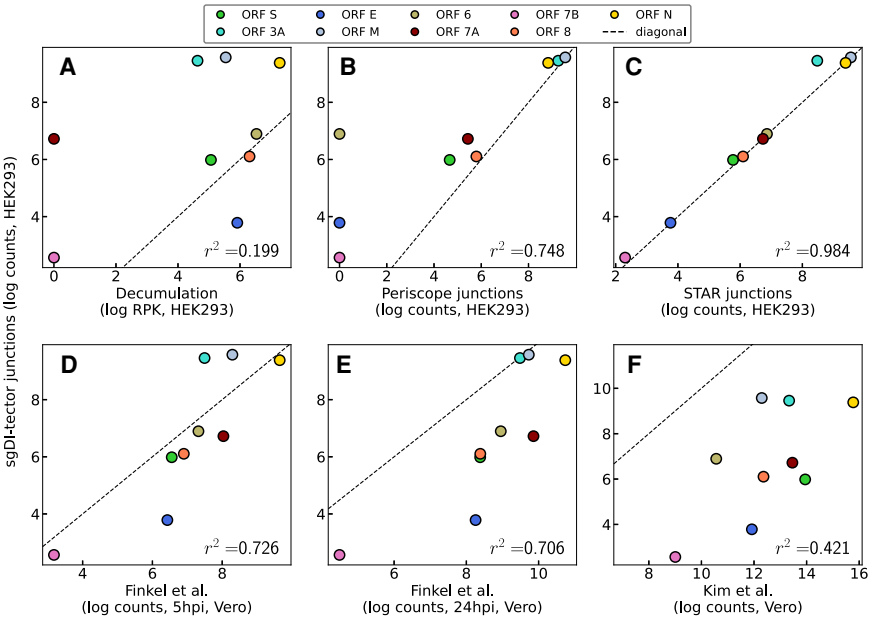

FIGURE 5. sgDI-tector results are not correlated with decumulation results, while agreeing with other tools applied on the same and on other data. The first row (A–C) presents only results obtained in our experiments and analyzed with several bioinformatic tools, while the second row (D–F) presents a comparison between our data and other data present in the literature. All given results are for one replicate. One pseudocount has been added when necessary to visualize the same number of ORFs in each plot. The black dashed line is the diagonal line, added to ease the comparison between the different methods. Notice that the plots on the *bottom* row compare results on different cell lines: HEK293 on the *y*-axis, and Vero on the *x*-axis.

5 hpi, which on the opposite were found with a striking correlation by sgDI-tector and by Finkel et al.

Leader-body conserved TRSs can be obtained from sgDI-tector output data

sgDI-tector does not require knowing a priori the identity of the leader sequence, as the observed junction-spanning reads are used in the pipeline to recover its position. For the viral strain used in our experiment, the leader-body junctions start, from the 5′ (leader) side, between position 60 and 80. This subsequence can be interpreted as the final part of the leader sequence. Focusing on the nucleotides around the RI positions for the reads detected by DI-tector, our tool allows for a completely automatic discovery of the leader-body conserved TRSs, as shown in Table 4, Figure 7, Supplemental Table 3 and Supplemental Figure 2. As apparent from the table and from the logo, the previously reported TRS 5′-ACGAAC-3′ (Wang et al. 2021) has clearly a special role, being present in most cases (and in the junctions with the highest number of observed counts). In particular, the nucleotides AAC in positions 71–73 (last part of 5′-ACGAAC-3′) are perfectly conserved within the body

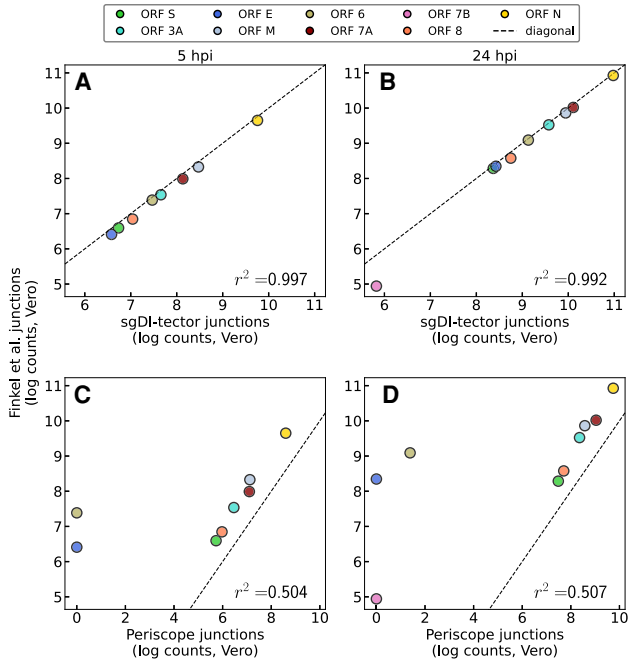

FIGURE 6. sgRNA junction counts obtained with sgDI-tector correlate with Finkel et al.'s junction counts obtained with STAR (panels *A,B*) while Periscope results show a lower correlation (panels *C,D*). *Left* (*right*) column contains the results for data at 5 (24) hpi. Only data for the first biological replicate are presented here. ORF 10 junctions are never found by both STAR and by DI-tector, while a single read has been detected by Periscope at 5 hpi. One pseudocount has been added to junctions which are not detected by one tool while being detected by the other. The black dashed line is the diagonal line, added to ease the comparison between the different methods.

TABLE 4. TRS and putative TRS detected by sgDI-tector

Junction ORF	Motif
ORF M	UAA**ACGAAC**U
ORF 3A	AA**ACGAAC**UU
ORF N	CUAA**ACGAAC**
ORF 7A	UAA**ACGAAC**
ORF 8	CUAA**ACGAAC**
ORF S	CUAA**ACGAAC**
ORF E	**ACGAAC**UU
U6cb1	GAACUUU
U54f0	A**ACGAAC**
U5f5b	UUCUCUA
U3dc4	GAACUUUAA
U382d	AACUUUAA
U6894	AACUUUA
U744b	AACUUUAA

For SARS-CoV-2, sgDI-tector fixes the minimum length to have a TRS hit to seven (see Materials and Methods), so only junctions having a subsequence identical in the leader of length larger or equal to seven are reported here. We highlighted the well-conserved canonical hexanucleotide ACGAAC motif with bold font. The results shown here have been obtained from the data coming from one biological replicate. The tables for the two additional biological replicates are given as Supplemental Table 3.

partners of the leader-body junctions, for all the junctions analyzed here. Remarkably, however not all the junctions present the 5′-ACGAAC-3′ motif. Moreover, this analysis shows that all the putative TRSs tend to be A- and U-rich. The tables and junction logos obtained for the other two biological replicates gave similar results, and are reported in Supplemental Table 3 and Supplemental Figure 2. The strategy used to collect the putative TRSs and to plot the logo are described in Materials and Methods.

DISCUSSION

Based on suggested mechanisms of production for coronavirus sgRNAs (Pasternak et al. 2006; Sola et al. 2015; Wang et al. 2021), several bioinformatic approaches can be applied to detect them in NGS data: decumulation, detection of exon–exon junction reads (such as the STAR algorithm), alignment to leader sequence and TRS (such as the Periscope tool) and DVG recognition tools. We earlier developed DI-tector to characterize different forms of DVGs in NGS data. In this study we compared the other existing techniques for estimation and characterization of SARS-CoV-2 sgRNAs protocols with a pipeline based on the DI-tector's output, which we named sgDI-tector.

A well-known strategy to quantify sgRNAs in coronaviruses is the so-called decumulation procedure, which has been introduced to characterize the sgRNA expression in cells infected by Murine coronavirus (also known as mouse

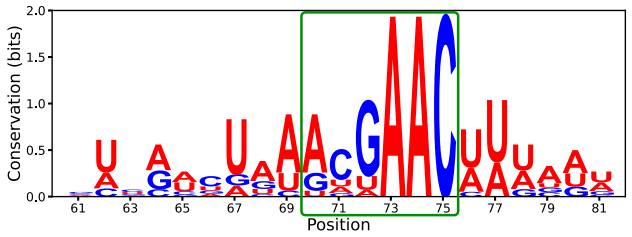

FIGURE 7. Logo of the RI positions around the TRS-putative sequences obtained from DI-tector. The conservation plotted as total height of the letters representing nucleotides is obtained as $log_2(4) - \Sigma_n f_i(n) log_2(f_i(n))$, where $f_i(n)$ is the frequency of nucleotide n in position i. Therefore a height equal to 2 corresponds to perfect conservation. The horizontal axis is the position with respect to the reference sequence (GISAID ID: EPI_ISL_414631). The green box highlights the canonical TRS. The alignment step to obtain this logo is described in the Materials and Methods section. Color code used: red for adenine and uracil, blue for cytosine and guanine.

hepatitis virus) (Irigoyen et al. 2016), and has been shown to give, in that case, consistent results with another standard approach consisting in finding the chimeric sequences that span the TRS.

However, the correlation between the two approaches decreased when the results obtained by Finkel et al. on SARS-CoV-2-infected cells were analyzed. Moreover, several sgRNA had negative RPKM after the decumulation procedure, highlighting another drawback of this method. Moreover, when the decumulation procedure was applied to our data, the results were completely inconsistent with any of the other tools used. We suggest that the low performance of the decumulation method for our data may be due to the large number of DVG observed, in particular insertions and deletions not resulting in canonical ORFs. This very large family of transcripts should be, in principle, accounted for in the decumulation procedure. Moreover, as observed in Kim et al. (2020) and Finkel et al. (2021), the canonical ORFs seem to be produced together with other, noncanonical ORFs and DVGs that should be considered also during decumulation. The remarkable correlation observed by Irigoyen et al. (2016) could be explained by the fact that most of DVGs and noncanonical ORFs are produced in relevant amounts later in infection, as was also discussed in Finkel et al. (2021). This was confirmed when DI-tector was run on Finkel et al.'s data: ~65% of the DVG at 5 h post infection were associated with a canonical ORF and this number reduced to 46% at 24 h post infection. Moreover, the total number of detected DVG (including canonical ORF junctions) compared with the number of non-DVG, mapped viral reads increased from 1% (5 hpi) to 2% (24 hpi).

Applying the STAR algorithm for detection of SARS-CoV-2 sgRNAs as a type of "exon–exon junction" was successfully performed in Kim et al. (2020), Finkel et al. (2021), and Wang et al. (2021). We observed a strong correlation between NGS data analyzed by sgDI-tector and STAR algorithms with only a few relevant differences. On the contrary, Periscope analysis on short-read data gave qualitatively different results, suggesting that this tool may have a decrease in performance when applied to a data set not obtained through the ARTIC sequencing protocol for which the tool has been developed, especially when the reads have short lengths. When compared to previously published bioinformatic tools, the advantages of sgDI-tector are mainly two: Firstly, sgDI-tector does not need to have as an input the leader sequence or the TRS, differently from Periscope and from other TRS-based approaches; secondly, sgDI-tector has been designed specifically for addressing the sgRNA level expression in viruses, and for this reason it works without the need for unconventional parameter choices, as is the case for STAR. In addition to these two technical advantages, sgDI-tector is user-friendly. We consider sgDI-tector to be the most user-friendly bioinformatic tool to estimate SARS-CoV-2 sgRNA from NGS data. Indeed, although STAR is a well-known mapping tool widely used in transcriptomics, it requires a large number of nontrivial options and parameters, which must be tuned accurately to detect sgRNAs. Periscope is a pipeline based on a workflow management system (snakemake) that is easy to install but, similarly to STAR, it has several parameters that the user must know and the full list of mandatory options is not provided.

sgDI-tector is a Python script that is easy to run. It only needs the DI-tector script in the working directory (together with the tools required to run DI-tector, that is BWA and samtools), and all efforts have been made so that sgDI-tector operates with the lowest possible number of settings that can be easily selected based on virology knowledge. Finally, although the differences between STAR and sgDI-tector were small in most of the cases, for one canonical junction (used to express ORF 3A) the results of these two tools were statistically different, whereas Periscope's result was for ORF 3A compatible with sgDI-tector's result and not compatible with STAR's result, suggesting that sgDI-tector might be more accurate than STAR in some cases (Table 3).

We showed here that the large fraction of deletion DVG detected by DI-tector can be used to identify and quantify viral sgRNAs from NGS data. Notice that, although all these ORFs have in principle the potential to express an ORF (5′ and 3′ identical to the full viral genome and an AUG start codon), we do not expect all of them to be translated, as their AUG codons could be within a poor Kozak context to serve as translation sites. Another possibility is that for some noncanonical ORFs the initiation of translation could be driven by non-AUG start codons (Kearse and Wilusz 2017) and thus escape the sgDI-tector algorithm. Moreover, from our analysis we had access to the full set of DVG produced by SARS-CoV-2 during its life cycle (see Table 1). We used RT-qPCR to confirm the presence of noncanonical sgRNAs detected by sgDI-tector in

cells infected with SARS-CoV-2. ORF U3dc4 is the noncanonical sgRNA that has its body TRS located in the ORF1ab and was detected in our three biological replicates specifically from infected cells (Figs. 3, 4). This noncanonical RNA caught our attention as if transcribed it should encode a part of NSP12 protein which is SARS-CoV-2 RdRp. However additional experiments are required to validate that ORF U3dc4 is indeed translated.

In addition to the deletion DVGs, we also detected an important number of insertion DVGs. sgDI-tector algorithm can only detect insertions of viral origin. Notice that the observed number of insertion DVG reads (average over replicates: 23.8% of all the DVG reads) is very close to the number of deletion DVG reads that cannot be associated with canonical sgRNAs (average over replicates: 23.4% of all the DVG reads). Further studies are needed to understand whether these insertion events correspond to the real DVGs produced during viral replication or represent viral genome recombination events previously described for coronaviruses (Simon-Loriere and Holmes 2011).

Finally, we detected a very low number of DVG of the type 5′/3′-copy-backs/snap-backs. Of note, 5′/3′-copy-back DVG are largely described for negative-sense RNA viruses (Lazzarini et al. 1981; Dimmock et al. 2014).

Exact mechanisms of production of DVG are unknown. The central question is whether their production is induced by host factors, aiming at introducing interferences with viral replication and allowing virus detection by the host's innate immune system. As stated, DVG are truncated and/or rearranged forms of viral genomes generated by most viruses during viral replication and sharing the minimum essential characteristics for replication: a competent initiation site at the 3′-end, its complementary sequence at the 5′-end. It is intriguing that the above description can also be applied on SARS-CoV-2 sgRNAs. This similarity can further be used to suggest mechanisms for production of deletion forms of DVG arguing for an internal property of viral RdRp to produce DVGs. Thus, further comparison of molecular structures and kinetics of coronavirus sgRNAs and DVGs accumulation will be of strong interest for future studies.

MATERIALS AND METHODS

Cells

HEK293 (human embryonic kidney cells) cell lines expressing One-STrEP-tagged Cherry (ST-CH) (Schaecher et al. 2007), Vero-E6 (African green monkey kidney cells, ATCC CRL-1586), were maintained at 37°C, 5% CO_2 in Dulbecco's modified Eagle medium (DMEM; Thermo Fisher Scientific) supplemented with 5% heat-inactivated fetal calf serum (FCS; GE Healthcare) and 1% PS (Penicillin 10,000 U/mL; Streptomycin 10,000 µg/mL). ST-CH cell line was supplemented with G418 (Sigma) at 500 g/mL. The absence of mycoplasma was regularly checked by PCR in all cell lines. For the generation of ST-CH overexpressing ACE-2 (ST-CH$^{ACE\text{-}2}$), lentivirus transduction of hACE2 was performed. Cells were screened by FACS for ACE2 expression and susceptibility to SARS-CoV-2 replication.

Viral titers, infection with SARS-CoV-2 and total RNA extraction

The SARS-CoV-2 hCoV-19/France/GES-1973/2020 GISAID ID: EPI_ISL_414631 strain was supplied by the National Reference Centre for Respiratory Viruses hosted by Institut Pasteur (Paris, France). Vero-E6 cells were used for the amplification and titration of viral stocks. Vero-E6 monolayers were infected with SARS-CoV-2 in the presence of 0.1% TPCK trypsin (Sigma) at a multiplicity of infection (MOI) of 0.0001 plaque-forming units (pfu) per cell. When the cytopathogenic effect was apparent, the culture supernatant was collected and centrifuged for 5 min at 850*g*. The efficiency of virus amplification was evaluated by titrating the supernatant on Vero-E6 cells, in a standard plaque assay adapted from Matrosovich et al. (2006). For SARS-CoV-2 infection ST-CH$^{ACE\text{-}2}$, cells were seeded into polylysine-coated (SIGMA) T150 flasks 1 d before infection (20×10^6 cells/flask). Virus infections were carried out at an MOI of 1. Viruses were diluted with DMEM 0%FCS to obtain a final inoculum volume of 5 mL. Cells were incubated with virus for 1 h at 37°C with gentle shaking. Twenty-five milliliters of DMEM containing 0% FCS was added to each T150 flask, and cells were incubated at 37°C until infections were stopped by cell lysis 24 h later. Total RNA was extracted from either SARS-CoV-2- or mock-infected ST-CH$^{ACE\text{-}2}$ using TriLS (TriLS, Sigma) reagent protocol previously described in detail in Sanchez David et al. (2016). All experiments with SARS-CoV-2 were conducted under strict BSL3 conditions.

Raw data collection, preprocessing and normalization scheme

NGS libraries were built using a TruSeq mRNA-Seq library preparation kit (Cat#20020594 Illumina), according to the manufacturer's recommendations. Quality control was performed on an Agilent Bioanalyzer. Sequencing was performed on the Illumina NextSeq500 platform to generate single-end 75 bp reads bearing strand specificity. Reads were cleaned of adapter sequences and low-quality sequences using cutadapt version 2.9. Only sequences at least 25 nt in length were considered for further analysis. Bowtie version 2.1.0, with default parameters, was used for alignment on the reference genome (hCoV-19/France/GES-1973/2020, GISAID accession ID: EPI_ISL_414631). SARS-CoV-2 genome coverages were computed with bedtools genomecov for each strand.

For several analyses performed in this work, we needed to use data from the three biological replicates of our experimental setup. To compare in a more robust way the data from independent experiments, we normalized all the counts as follows: For each biological replicate, we took the number of reads mapped to the SARS-CoV-2 genome and divided this by the value obtained for Replicate 1 (the raw number of reads are reported in Supplemental Table 1). The three values that we obtained in this way (1 for Replicate 1, and other values for other replicates) are used to rescale all the number of reads before taking any

average over biological replicates. In particular, averages of normalized data have been used for Tables 3 and 1, and for Figures 2 and 4.

RT-qPCR validation of noncanonical U3dc4 ORF

RT-qPCR was performed on RNA samples from SARS-CoV-2- or mock-infected cells performed in three biological replicates and prepared as described for NGS. First-strand complementary DNA (cDNA) synthesis was performed on 2500 ng of total RNA in a final volume of 20 µL with the Superscript IV VILO (Thermo Scientific #11756050) according to the manufacturer's protocol. RT-qPCR analysis was performed using Applied Biosystems StepOnePlus technology. Reactions were performed on an equivalent of 50 ng of total RNA using the SYBER Green Kit (Thermo Fisher Scientific #4309155) for qPCR analysis according to the manufacturer's protocol. Reactions were performed in a final volume of 20 µL in the presence of 60 nM U3dc4-specific forward (5′-CCTTCCCAGGTAACAAAC) and reverse (5′-GTCTCAGTCC AACATTTTG) primers; or N-specific forward (5′-TAAAGGTTTA TACCTTCCCA) or reverse (5′-CGTTCTCCATTCTGGTTA) primers; or GAPDH forward (5′-CACATGGCCTCCAAGGAGTAA) and reverse (5′-TGAGGGTCTCTCTCTTCCTCTTGT) primers.

sgDI-tector pipeline: from NGS data to sgRNA detection

When sgDI-tector is run, it first calls DI-tector (Beauclair et al. 2018) (here used in version 0.6) with default parameters (using bwa v0.7.17, bedtools v2.17.0, and samtools v1.9) to detect SARS-CoV-2 DVGs. DI-tector outputs four different types of DVG, namely deletions, insertions, 3′- and 5′-copy backs/snapbacks. For each deletion, the BP (breakpoint) and RI (reinitiation) sites are specified, consisting in the two sites that, despite being separated in the full-length genome, are brought together in the junction read. The pipeline for sgRNA detection starts by finding the window of a user-modifiable length (the default value is 20) with the largest number of BP. Under the hypothesis that the virus is replicating, and that replication needs sgRNAs, we expect (and verify in each in vitro sample analyzed here) that this window coincides with the end of the leader sequence. Then, sgDI-tector filter deletion DVGs by requiring the BP to lay into this window. The resulting deletions are then associated with an ORF, by finding the first ATG subsequence after the RI. The families of junctions obtained in this way are then sorted by the number of reads belonging to the family, and given as an output. Optionally the user can provide a list of reference subgenomic ORFs (in fasta format), that will be used by sgDI-tector to name the sgRNAs found. In particular, this is done by aligning the putative expressed protein to the list of known proteins, and comparing the alignment score with a fixed threshold. When more than one hit is obtained for the same viral protein (for instance, because two different sgRNA produce two proteins with few amino acids of difference), the name of the top hit is taken from the user-provided file, and the name of successive hits are obtained from it by adding increasing numbers (e.g., ORF 3A-1). In addition to the sgRNA list, sgDI-tector outputs the leader sequence used. The full sgDI-tector pipeline described here, which takes as input the DI-tector results and gives as output a list of putative sgRNAs, is presented in Figure 8. In all samples analyzed, all the observed canonical subgenomic ORFs (S, 3A, E, M, 6, 7a, 7b, 8, N) were within the 13 families with the largest number of reads (see Figs. 3 and 4).

TRS detection through junction analysis

DI-tector itself allows several graphical outputs, and among them there is the sequence logo of nucleotides just before and after the RI position. However, this functionality cannot be used as it is for sgRNA junctions. Indeed, it is well known (Sola et al. 2015) that the leader-body junction is regulated by a core subsequence that is identical in the two sides of the junction, and this creates ambiguity in precisely defining BP and RI, and makes a further alignment step necessary to correctly compare short subsequences spanning the RI position. However, the alignment step is nontrivial, as these subsequences are typically not alignable but for some small parts. Therefore, we used a different approach: Firstly, sgDI-tector computes the probability of a random subsequence of the viral genome to have a subsequence of length L (putative TRS), which appears also in the final part of the leader sequence. This allows sgDI-tector to fix L^*, the length for which this probability is lower than 0.05. Then, for all sequences spanning the RI position in junctions, putative TRS of length larger than L^* are collected, and saved in an output file. Once the putative TRSs have been obtained, they can be aligned to the leader sequence. The resulting logo, which has been directly obtained from the sequences in Table 4, is presented in Supplemental Figure 3. To have even more complete information about the body side of the junctions, once the putative TRSs are aligned, the remaining nucleotides of the body side of the junctions can be used together with TRSs to produce the logo presented in Figure 7 (and Supplemental Figure 2 for the other two biological replicates). Equivalently, the logos are obtained from the sequences around the body part of the junctions listed in Table 4, aligned to the leader sequence so that the TRSs obtained in Table 4 overlap with the corresponding identical sequence in the leader part. Finally, we decided not to include any information

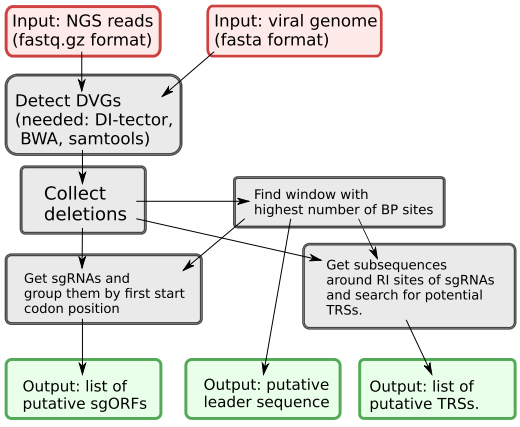

FIGURE 8. Scheme of the sgDI-tector pipeline introduced here to find the putative position of the leader sequence, sgRNAs, and a list of putative transcription-regulatory sequences (TRSs). Red boxes denote necessary inputs for the sgDI-tector tool, and green boxes denote outputs.

about abundances of the junctions in the logos. If such information were included, of course, the full canonical TRS 5′-ACGAAC-3′ (highlighted by a green box in Fig. 7) would become almost perfectly conserved, because it is present in all the junctions with a high number of counts.

DATA DEPOSITION

The data collected and used for this work have been deposited in NCBI's Gene Expression Omnibus (Edgar et al. 2002) and are accessible through GEO Series accession number GSE180632, at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE180632. sgDI-tector code is publicly available on Github, at https://github.com/adigioacchino/sgDI-tector.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

COMPETING INTEREST STATEMENT

B.D.G. is a consultant or received honoraria for Darwin Health, Merck, PMV Pharma, ROME Therapeutics (of which he is a cofounder), Bristol-Meyers Squibb, and Chugai Pharmaceuticals and has research funding from Bristol-Meyers Squibb and Merck. The other authors declare that they have no competing interests.

ACKNOWLEDGMENTS

We would like to thank J. Pipoli da Fonseca, L. Lemée from Biomics Platform, C2RT, Institut Pasteur, Paris, France for RNA NGS, IBISA and the Illumina COVID-19 Projects' offer. The authors would like to thank members of the Tangy, van der Werf laboratories and the National Reference Center (CNR) for Respiratory Viruses at the Institut Pasteur for support and valuable discussions. We acknowledge the ANR (Agence Nationale de la Recherche) and FRM (Fondation de la Recherche Médicale) for funding this work through the AAP Flash-Covid 19 project SARS-Cov-2immunRNAs. A.V.K. received support from ANR through the grant ANR-LBX-62 IBEID CoV-2SENSING/COVID 19. RNA NGS has been supported by France Génomique (ANR-10-INBS-09-09).

Received August 30, 2021; accepted December 7, 2021.

REFERENCES

Alexandersen S, Chamings A, Bhatta TR. 2020. SARS-CoV-2 genomic and subgenomic RNAs in diagnostic samples are not an indicator of active replication. *Nat Commun* **11:** 6059. doi:10.1038/s41467-020-19883-7

Beauclair G, Mura M, Combredet C, Tangy F, Jouvenet N, Komarova AV. 2018. DI-tector: defective interfering viral genomes' detector for next-generation sequencing data. *RNA* **24:** 1285–1296. doi:10.1261/rna.066910.118

Davidson AD, Williamson MK, Lewis S, Shoemark D, Carroll MW, Heesom KJ, Zambon M, Ellis J, Lewis PA, Hiscox JA, et al. 2020. Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. *Genome Med* **12:** 1–15. doi:10.1186/s13073-020-00763-0

Dimmock NJ, Easton AJ, Goff SP. 2014. Defective interfering influenza virus RNAs: time to reevaluate their clinical potential as broad-spectrum antivirals? *J Virol* **88:** 5217–5227. doi:10.1128/JVI.03193-13

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2012. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30:** 207–210. doi:10.1093/nar/30.1.207

Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, Tamir H, Achdout H, Stein D, Israeli O, et al. 2021. The coding capacity of SARS-CoV-2. *Nature* **589:** 125–130. doi:10.1038/s41586-020-2739-1

Irigoyen N, Firth AE, Jones JD, Chung BY-W, Siddell SG, Brierley I. 2016. High-resolution analysis of coronavirus gene expression by RNA sequencing and ribosome profiling. *PLoS Pathog* **12:** 1–44. doi:10.1371/journal.ppat.1005473

Kearse MG, Wilusz JE. 2017. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev* **31:** 1717–1731. doi:10.1101/gad.305250.117

Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. 2020. The architecture of SARS-CoV-2 transcriptome. *Cell* **181:** 914–921.e10. doi:10.1016/j.cell.2020.04.011

Lazzarini RA, Keene JD, Schubert M. 1981. The origins of defective interfering particles of the negative-strand RNA viruses. *Cell* **26:** 145–154. doi:10.1016/0092-8674(81)90298-1

Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, et al. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395:** 565–574. doi:10.1016/S0140-6736(20)30251-8

Matrosovich M, Matrosovich T, Garten W, Klenk H-D. 2006. New low-viscosity overlay medium for viral plaque assays. *Virol J* **3:** 1–7. doi:10.1186/1743-422X-3-63

Miao Z, Tidu A, Eriani G, Martin F. 2021. Secondary structure of the SARS-CoV-2 5′-UTR. *RNA Biol* **18:** 447–456. doi:10.1080/15476286.2020.1814556

Parker MD, Lindsey BB, Leary S, Gaudieri S, Chopra A, Wyles M, Angyal A, Green LR, Parsons P, Tucker RM, et al. 2021. Subgenomic RNA identification in SARS-CoV-2 genomic sequencing data. *Genome Res* **31:** 645–658. doi:10.1101/gr.268110.120

Pasternak AO, Spaan WJM, Snijder EJ. 2006. Nidovirus transcription: how to make sense …? *J Gen Virol* **87:** 1403–1421. doi:10.1099/vir.0.81611-0

Pathak KB, Nagy PD. 2009. Defective interfering RNAs: foes of viruses and friends of virologists. *Viruses* **1:** 895–919. doi:10.3390/v1030895

Sanchez David RY, Combredet C, Sismeiro O, Dillies M-A, Jagla B, Coppée J-Y, Mura M, Guerbois Galla M, Despres P, Tangy F, et al. 2016. Comparative analysis of viral RNA signatures on different RIG-I-like receptors. *Elife* **5:** e11275. doi:10.7554/eLife.11275

Schaecher SR, Mackenzie JM, Pekosz A. 2007. The ORF7b protein of severe acute respiratory syndrome coronavirus (SARS-CoV) is expressed in virus-infected cells and incorporated into SARS-CoV particles. *J Virol* **81:** 718–731. doi:10.1128/JVI.01691-06

Simon-Loriere E, Holmes EC. 2011. Why do RNA viruses recombine? *Nat Rev Microbiol* **9:** 617–626. doi:10.1038/nrmicro2614

Sola I, Almazán F, Zúñiga S, Enjuanes L. 2015. Continuous and discontinuous RNA synthesis in coronaviruses. *Annu Rev Virol* **2:** 265–288. doi:10.1146/annurev-virology-100114-055218

Strauss EG, Strauss JH. 2007. *Viruses and human disease, 2nd ed.* Elsevier, Amsterdam.

Wang D, Jiang A, Feng J, Li G, Guo D, Sajid M, Wu K, Zhang Q, Ponty Y, Will S, et al. 2021. The SARS-CoV-2 subgenome landscape and its novel regulatory features. *Mol Cell* **81:** 2135–2147.e5. doi:10.1016/j.molcel.2021.02.036

Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, et al. 2020. A new coronavirus associated with human respiratory disease in China. *Nature* **579:** 265–269. doi:10.1038/s41586-020-2008-3

Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579:** 270–273. doi:10.1038/s41586-020-2012-7

MEET THE FIRST AUTHOR

Andrea Di Gioacchino

***Meet the First Author(s)* is a new editorial feature within *RNA*, in which the first author(s) of research-based papers in each issue have the opportunity to introduce themselves and their work to readers of *RNA* and the RNA research community. Andrea Di Gioacchino is the first author of this paper, "sgDI-tector: defective interfering viral genome bioinformatics for detection of coronavirus subgenomic RNAs." Andrea is currently a postdoctoral student at the physics department of Ecole Normale Superieure (in Paris, France), on the "statistical physics and inference for biology" team. Andrea trained as a theoretical physicist, specializing in statistical mechanics and disordered systems, and is focusing on using tools and ideas from statistical physics and machine learning to address biological problems, ranging from inferring the host-induced pressures that shape the viral genome to machine-learning-assisted design of RNA and DNA aptamers that can bind targets in a strong and specific fashion.**

What are the major results described in your paper and how do they impact this branch of the field?

In our paper we introduce a new algorithm to quantify and analyze subgenomic RNA produced during coronavirus infections, starting from next-generation sequencing (NGS) data. Our tool, which we named sgDI-tector, is the first specifically designed for this aim. In our paper, we collected NGS data from cells infected with SARS-CoV-2 and compared the results obtained through sgDI-tector with other approaches, showing that sgDI-tector is extremely effective. Finally, our software has been designed to be as user-friendly as possible, which I think is very important!

What led you to study RNA or this aspect of RNA science?

I have always been fascinated by the ability of viruses to compress the information in their genomes, while dealing at the same time with the immune pressure induced by the host: It is amazing (and also a bit scary!) how viruses evolved to exploit very complex mechanisms to code for multiple proteins in their genetic code, for instance through inducing ribosomal frameshifting or through subgenomic RNA production.

In 2020, at the beginning of the COVID-19 pandemic, my group leaders Dr. Simona Cocco and Dr. Rémi Monasson asked me to work, together with them and several international collaborators, on a project to apply statistical physics methods to study different aspects of SARS-CoV-2's genome. Soon we realized how complex the replication mechanisms of this virus are, and how we should always take this into account in all our analyses.

Therefore, together with Dr. Komarova at Institut Pasteur, who has extensive experience in studying RNA viruses and the roles of virus-host RNA–protein interactions in innate immunity response to RNA virus infections, we decided to investigate in more detail these mechanisms. Our interdisciplinary collaboration is the beginning of the story that led to the development of sgDI-tector.

Our work on sgDI-tector is actually only the tip of the iceberg: We are continuing our work on these topics, and we hope we will be able to present new results soon!

What are some of the landmark moments that provoked your interest in science or your development as a scientist?

As far as I remember, I have always been interested in science (especially in physics and mathematics at the beginning!).

But I think the most important role in shaping a scientist is actually played by the people one meets and works with.

For instance, I can remember clearly, during my first years at the Physics department of the University of Milan (Italy), the exam given by the professor who, a few years later, became my PhD supervisor. And I distinctly remember also a number of other moments when I met someone who was very passionate about her/his research and passed to me a bit of that passion. For instance, the amazing research group I am working with right now at Ecole Normale Superieure (together with our collaborators in France and in the US) taught me a lot about viruses, bioinformatics, RNA, DNA, proteins, human immune system …, and my research right now is focused on all these topics!

Continued

What are your subsequent near- or long-term career plans?

I just started looking "under the hood" of biology problems from the very peculiar point of view of a theoretical physicist. For sure I can say right now that I am very happy about the topics I am dealing with, and that they will be more and more relevant in my future career!

sgDI-tector: defective interfering viral genome bioinformatics for detection of coronavirus subgenomic RNAs

Andrea Di Gioacchino, Rachel Legendre, Yannis Rahou, et al.

RNA 2022 28: 277-289 originally published online December 22, 2021
Access the most recent version at doi:10.1261/rna.078969.121