The puzzling RNA world of SARS-CoV-2

Barbara Illi

Institute of Molecular Biology and Pathology, National Research Council (IBPM-CNR), c/o

Department of Biology and Biotechnology “Charles Darwin”, Sapienza University, Rome, Italy.

Note: highlighted in italic are information for non-biologysts

Introduction

The emergence of the 2019 novel Cororonavirus outbreak (2019-nCoV), actually renamed Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2), during the past late December, still represents a big challenge for the scientific community. An enormous effort is ongoing to provide the highest number of information in the smallest fraction of time and, indeed, since January 2020, about 10000 papers have been published to date.

Other Coronavirus outbreaks have spread in the last 17 years, SARS and Middle East Respiratory Syndrome (MERS) in 2003 and 2012, respectively, but they have been so efficiently controlled that the produced vaccines have been useless. SARS-CoV-2 (hereafter simply Cov-2) is about 80% identical to its brother SARS-CoV. They also share several biological features, including the mechanism and proteins they use to entry host cells. CoV-2 replication and transcription processes seem common to all Coronaviruses. Therefore, why CoV-2 has striking differences in epidemiological and clinical manifestations with respect to other human CoVs? To answer these questions, we have to step back to CoVs replication and transcriptional mechanisms.

CoVs RNA synthesis.

Coronaviruses allocate the largest RNA genome – about 30 kilobases (KB) - among all RNA viruses*. As well as DNA, RNA is constituted by “bricks” called nucleotides (nt). The genetic information in both DNA and RNA is included in a sequence of “letters”, which are represented by the bases within the nucleotides and which for RNA are: adenine (A), guanine (G), cytosine (C) and uracyl (U). Different combinations of these four letters constitute different instructions for the synthesis of all protein products. This is true for higher organisms, as well as for microorganisms, such as bacteria and viruses.* As soon as the virus enter the host cell, its genome is immediately translated– by using ribosomes and

specific proteins of the host cell – into a major polyprotein (pp1ab), encoded by the replicase gene (figure 1), starting from two Open Reading Frames (ORF1a and ORF 1b), *which are genomic regions capable to be translated into proteins.* Pp1ab is cleaved into 16 non-structural proteins (nps 1-16; figure 1)[1,2] . Of note, two of the first translated nps – nps 3 and 5 – possess the cleavage activity to "cut" pp1ab. These 16 nps are employed both for the replication of the entire viral genome (gRNA) and for the production of each viral mRNA (subgenomic RNAs, sgRNAs). Downstream the replicase gene, are the genes encoding the structural proteins (Spike (S), Envelope (E), Membrane (M), Nucleocapsid (N)), which pack the copied gRNA, leading to the assembly of progeny virions (figure 1), and accessory proteins, whose role is still largely unknown (figure 1)[3].

Figure 1

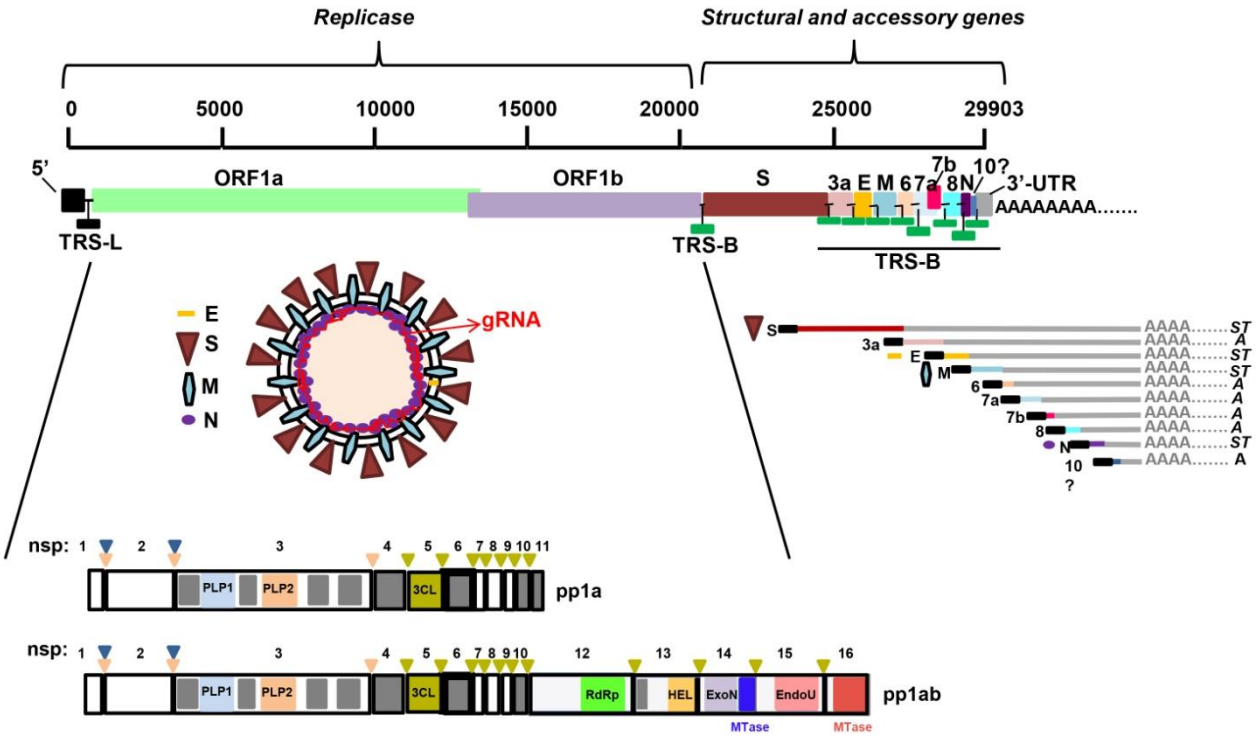

Figure 1. Structure of SARS-CoV-2 genome. Most of the genome is occupied by the replicase gene, encoding for 16 nps by the polyprotein pp1ab. Downstream, structural and accessory genes are depicted. Abbreviations: L= leader sequence; TRS-L=transcription regulatory sequence – leader; TRS-B= transcription regulatory sequence-body; ORF1a = open reading frame 1a; ORF1b=open reading frame 1b; S=spike gene; E=envelope gene; M=membrane gene; N=nucleocapsid gene. 3a, 6, 7a, 7b, 8, 10=accessory genes. Nsp=non structural protein; PLP=papain-like protease; 3CL=3C-like protease; RdRp=RNA-dependent RNA polymerase; HEL=helicase; ExoN=exonuclease; EndoU=endonuclease; MTase=methyltransferase; UTR=untranslated region. ST= structural protein; A=accessory protein. (Adapted from *Kim et al., Cell, 2020* and *Sola et al., Ann Rev Virol, 2015*)

Coronavirus RNA is basically a long messenger RNA (mRNA). Therefore, it is a "positive"strand, with a "so called" 5'→ 3' sense (figure 2a). RNA replication, that is the

Figure 2

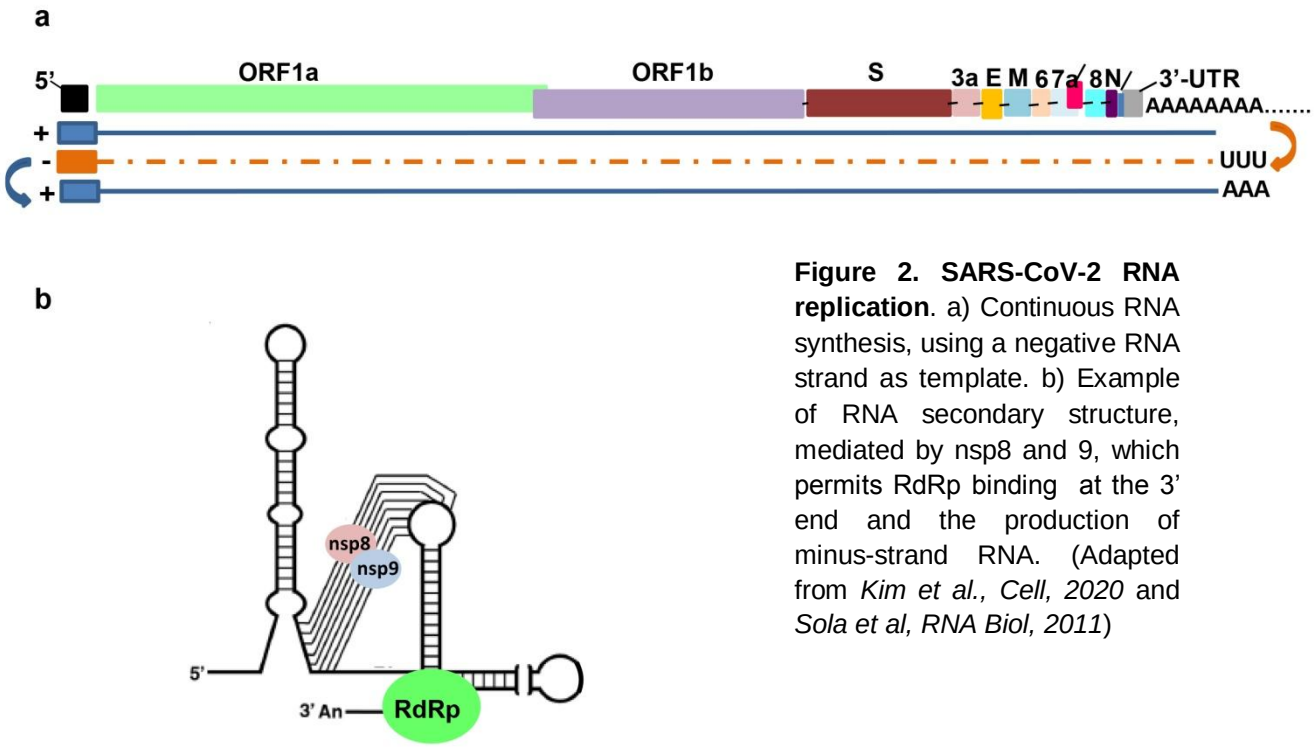

Figure 2. SARS-CoV-2 RNA replication. a) Continuous RNA synthesis, using a negative RNA strand as template. b) Example of RNA secondary structure, mediated by nsp8 and 9, which permits RdRp binding at the 3' end and the production of minus-strand RNA. (Adapted from *Kim et al., Cell, 2020* and *Sola et al, RNA Biol, 2011*)

production of several copies of the viral genome, is a "continuous synthesis". The new gRNA is produced by a negative intermediate, which represent the template for the positive gRNA (figure 2a). This process involves mainly nsp12, which encode the RNA-dependent RNA polymerase (RdRp), which binds to the 3' end of gRNA, across specific RNA sequences and structures, which may "flex" gRNA to promote replication (figure 2b). On the contrary, transcription, that is the synthesis of sgRNA, is a discontinuous process. Each sgRNA has a common sequence at the 5' end of 70 nucleotides, which is present only once at the 5' end of gRNA, called "leader" (L). The discontinuous sgRNA synthesis relies on the presence of small stretch of nt, named transcription regulatory sequences (TRS), which are present downstream to the leader at the 5' end (TRS-L) and preceding each sgRNA in the body of the genome (TRS-B, figure 1). TRSs are characterized by a conserved sequence (CS) which is identical for the Leader (CS-L) and for each TRS-B (CS-B) and which permits the pairing of CS-L with the nascent negative RNA intermediate which is complementary to the CS-B when the RNA bends over to move the complementary CS-B (cCS-B) to the CS-L (figure 3a), aligning TRS-L and each TRS-B[1,2].

This process allows the discontinuous transcription of each sgRNA, leading to Leader-body fusions and requiring long distance RNA-RNA interactions, mediated by

Figure 3

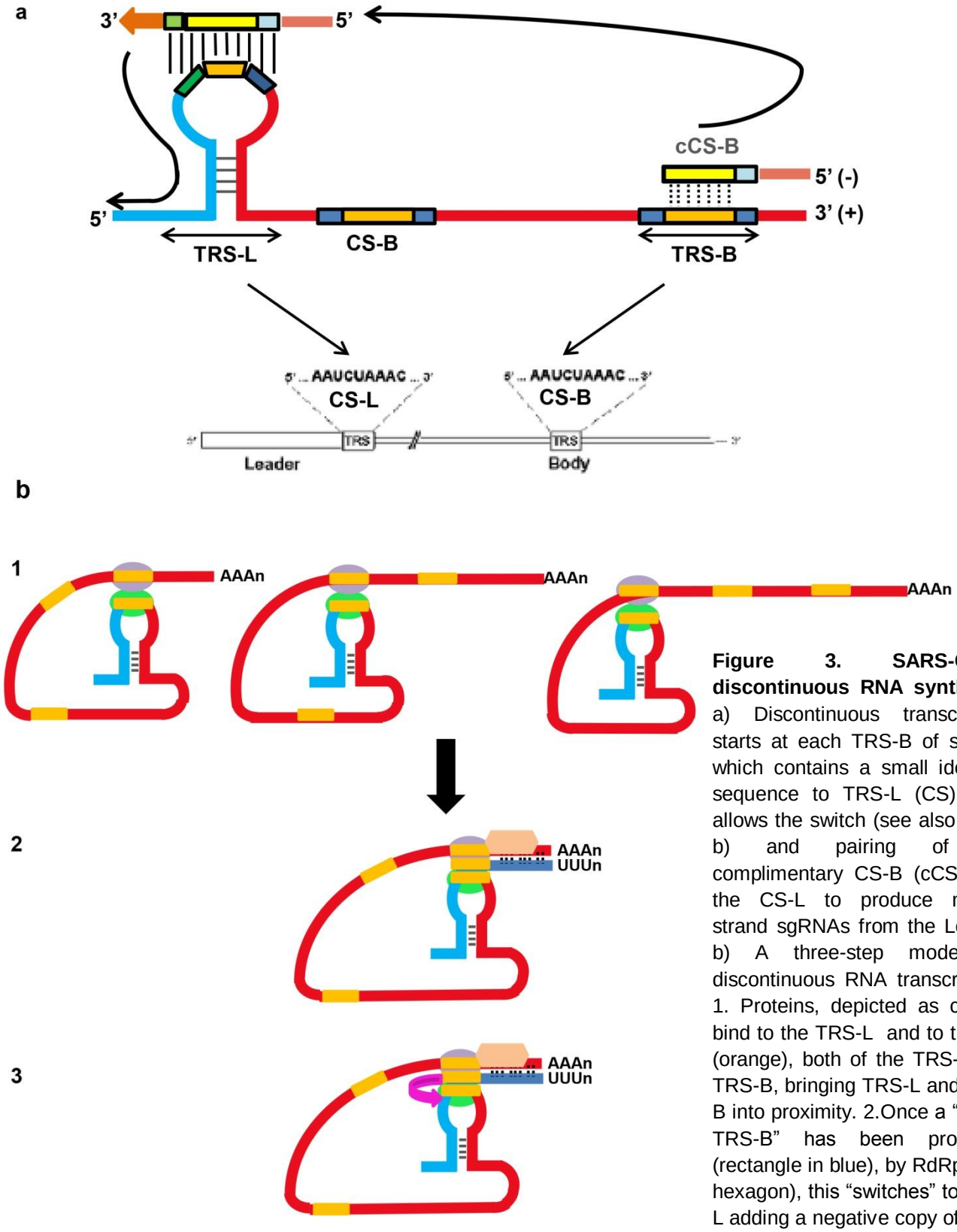

Figure 3. SARS-CoV-2 discontinuous RNA synthesis. a) Discontinuous transcription starts at each TRS-B of sgRNA which contains a small identical sequence to TRS-L (CS). This allows the switch (see also panel b) and pairing of the complimentary CS-B (cCS-B) to the CS-L to produce minus-strand sgRNAs from the Leader. b) A three-step model for discontinuous RNA transcription. 1. Proteins, depicted as circles, bind to the TRS-L and to the CS (orange), both of the TRS-L and TRS-B, bringing TRS-L and TRS-B into proximity. 2.Once a "minus TRS-B" has been produced (rectangle in blue), by RdRp (pink hexagon), this "switches" to TRS-L adding a negative copy of TRS-L completing the negative strand of sgRNA, which will serve as template for the positive, encoding, sgRNA. (Adapted from *Sola et al., Ann Rev Virol, 2015*)

protein-RNA complexes. An example of this mechanism is shown in figure 3b.

SARS-CoV2 RNA-dependent-RNA polymerase (RdRp): a "Jumping Jack Flash".

CoV-2 RNA synthesis is almost identical to that of other CoVs. However, some features of some of its sgRNAs suggest alternate mechanisms. Interestingly, CoV-2 whole panel of sgRNAs (i.e. the "transcriptome"), is expressed to a percentage higher than the transcriptome of the host cell, as if this latter is turned-off (figure 4a)[4]. 92% of viral sgRNAs are produced according to the mechanism described in the previous paragraph, and are mainly represented by 9 sgRNAs: N, S, M, E, 3a, 6, 7a, 7b and 8 (figure 4b), but a number of them are unusual (figure 4b, red box). Short sequences (3-4 nt) common to the 5' and 3' sites of the junction of these fusions suggests RdRp "jumping" from a donor to an acceptor RNA (figure 4c), a mechanism already described in other systems. This produces a number of sgRNAs with unknown functions in the viral life cycle, but some of them have the potential to produce protein products at a comparable quantity levels of accessory proteins. Therefore, it will be important to verify whether these proteins are effectively produced and which is their role.

Matters are even more complicated: RNA modifications.

Epigenetics is a scientific field which studies the reactions of a genome to a variety of stimuli, which may be environemental, intracellular, mechanical etc. Epigenetic modifications are hallmarks of these responses and, to make things as clear as possible, are chemical modifications of genomes and even proteins. These modifications are carried by multiproteins called "epigenetic machineries", which are activated by the above mentioned stimuli. Epigenetic modifications change the characteristics of a genome, for example, indicating which genes have to be turned on or off to make the cell able to respond rapidly and efficiently to external or internal inputs. Well known epigenetic modifications affect DNA and associated histone proteins (constituting the chromatin, then chromosomes within the cells). Also RNA may be susceptible of epigenetic modifications and, indeed, some chemical RNA modifications have been detected in viral RNAs. These modifications may be due to the addition of chemical groups (as a methyl group on cytosine and adenine), addition of nucleotides (such as uracyl containing nt), loss of amino groups from bases. In CoV-2 RNA modifications have been found, but their nature has not been yet identified nor their function (figure 4d). However, it may be supposed that these modifications may regulate RNA stability or may contribute to evade host immune response.

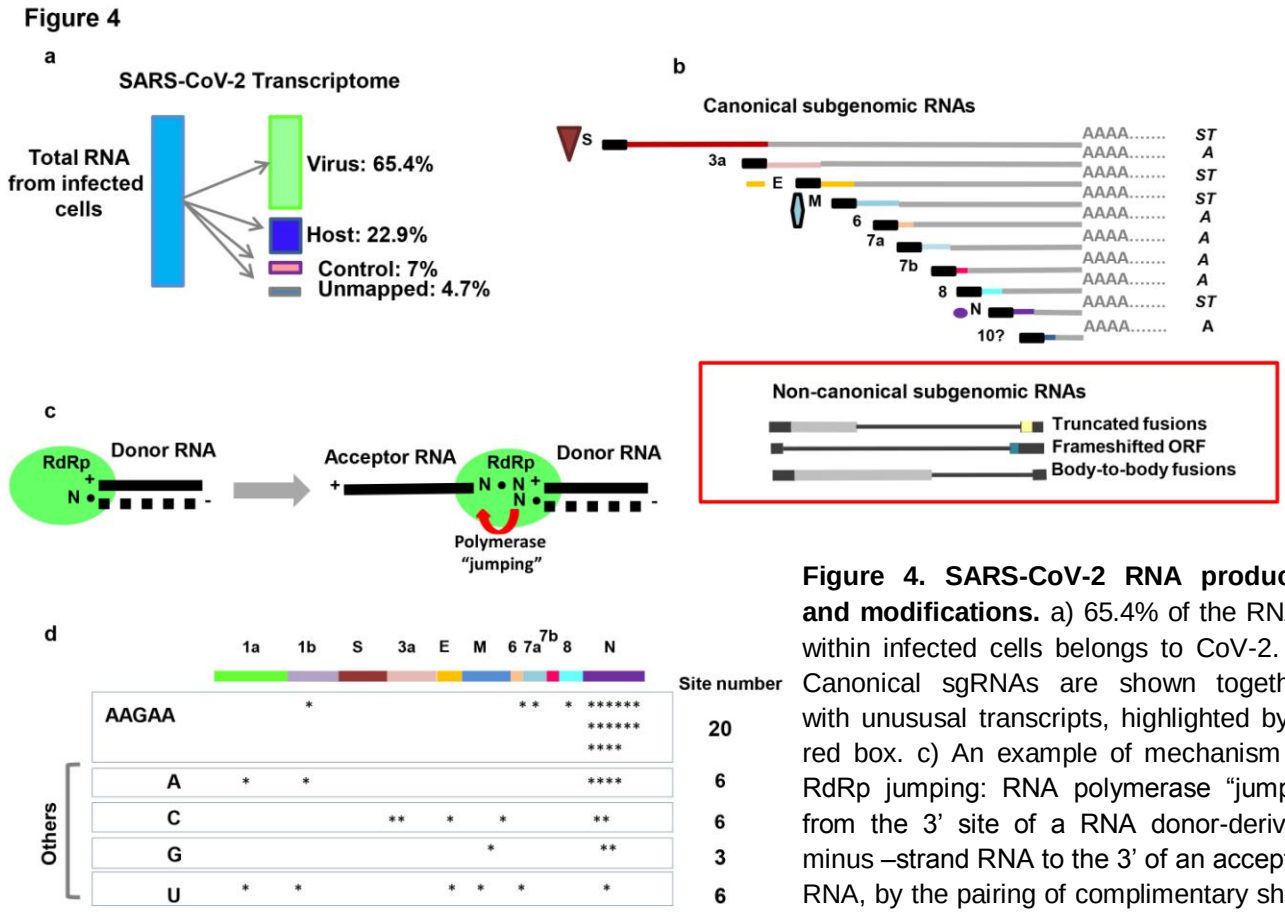

Figure 4. SARS-CoV-2 RNA products and modifications. a) 65.4% of the RNAs within infected cells belongs to CoV-2. b) Canonical sgRNAs are shown together with unusual transcripts, highlighted by a red box. c) An example of mechanism of RdRp jumping: RNA polymerase "jumps" from the 3' site of a RNA donor-derived minus –strand RNA to the 3' of an acceptor RNA, by the pairing of complimentary short sequences. d) Several RNA modifications are shown as asterisks. The highest number occur at the sequence similar to AAGAA. The number of modifications across the gRNA of CoV-2 is shown at right. (Adapted from *Kim et al., Cell, 2020*).

Conclusions

SARS-CoV-2 RNA biology is very similar to that of other CoVs. Nevertheless its RNA synthesis process is highly complex and, in some cases, escapes canonical mechanisms. Specifically, the high frequency of unusual fusions may produce viral variants, an issue which deserves a special attention, in light of drugs resistance or immune response evasion mechanisms CoV-2 may adopt to survive and propagate. RNA modifications may also play a role. These features have to be studied in animal tissues, where an immune system is active and responding. Furthermore, it will be interesting to evaluate whether CoV-2 RNA modifications are present in other CoVs, to better understand CoV-2 biology.

References

1. Snijder EJ, Decroly E, Ziebuhr J. The nonstructural proteins directing Coronavirus RNA synthesis and processing. Adv Virus Res. 2016;96:59-126. doi:10.1016/bs.avir.2016.08.008.
2. Sola I, Almazán F, Zúñiga S, Enjuanes L. Continuous and Discontinuous RNA Synthesis in Coronaviruses Annu Rev Virol. 2015;2: 265-88. doi: 10.1146/annurev-virology-100114-055218.
3. Masters PS. The molecular biology of Coronaviruses. Adv Virus Res. 2006;66:193-292.
4. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 transcriptome. Cell. 2020 Apr 18. pii: S0092-8674(20)30406-2. doi: 10.1016/j.cell.2020.04.011.