Molecular Biology of SARS-COV-2

A quick overview of coronavirus molecular biology

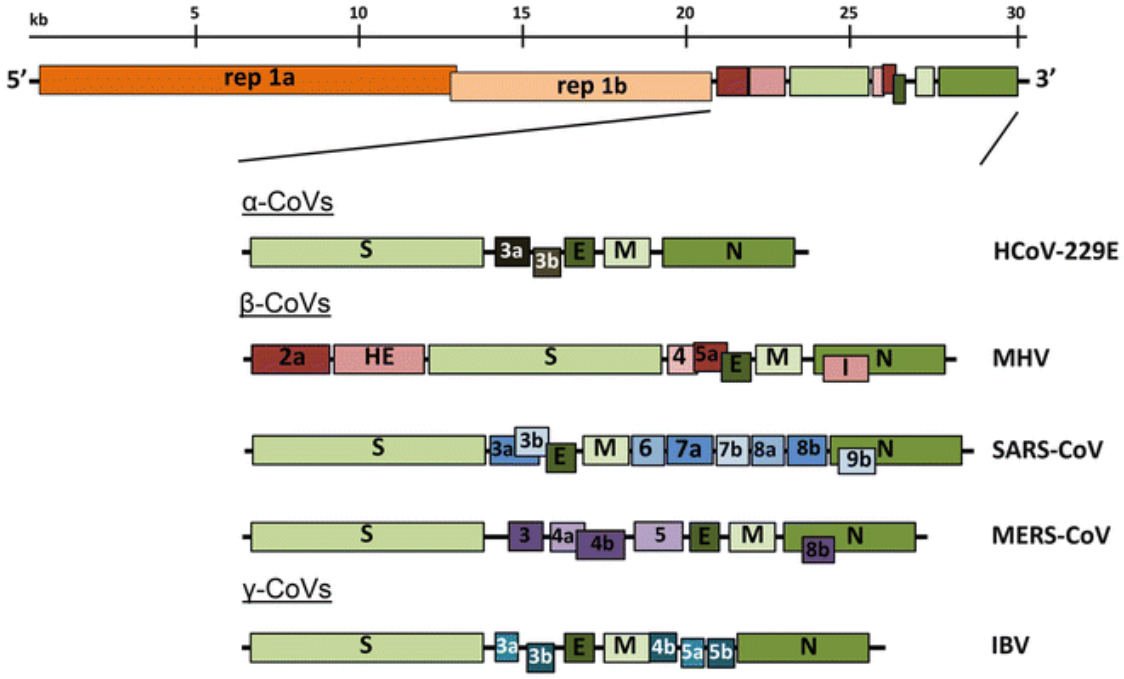
Given the current situation we will continue this class using SARS-COV-2 as example for all further analyses. So far we have covered variant calling and genome assembly. Next we will be looking into RNAseq. Since SARS-COV-2 is a positive strand RNA virus that pretends to be a typical human transcript, it is a great system to learn about RNAseq. But first we need to learn about the coronavirus molecular biology. This is what we will do in this lecture.

The following summary is based on these publications:

- Masters:2006
- Fehr and Perlman:2015
- Sola:2015
- Kirchdoerfer:2016
- Walls:2020

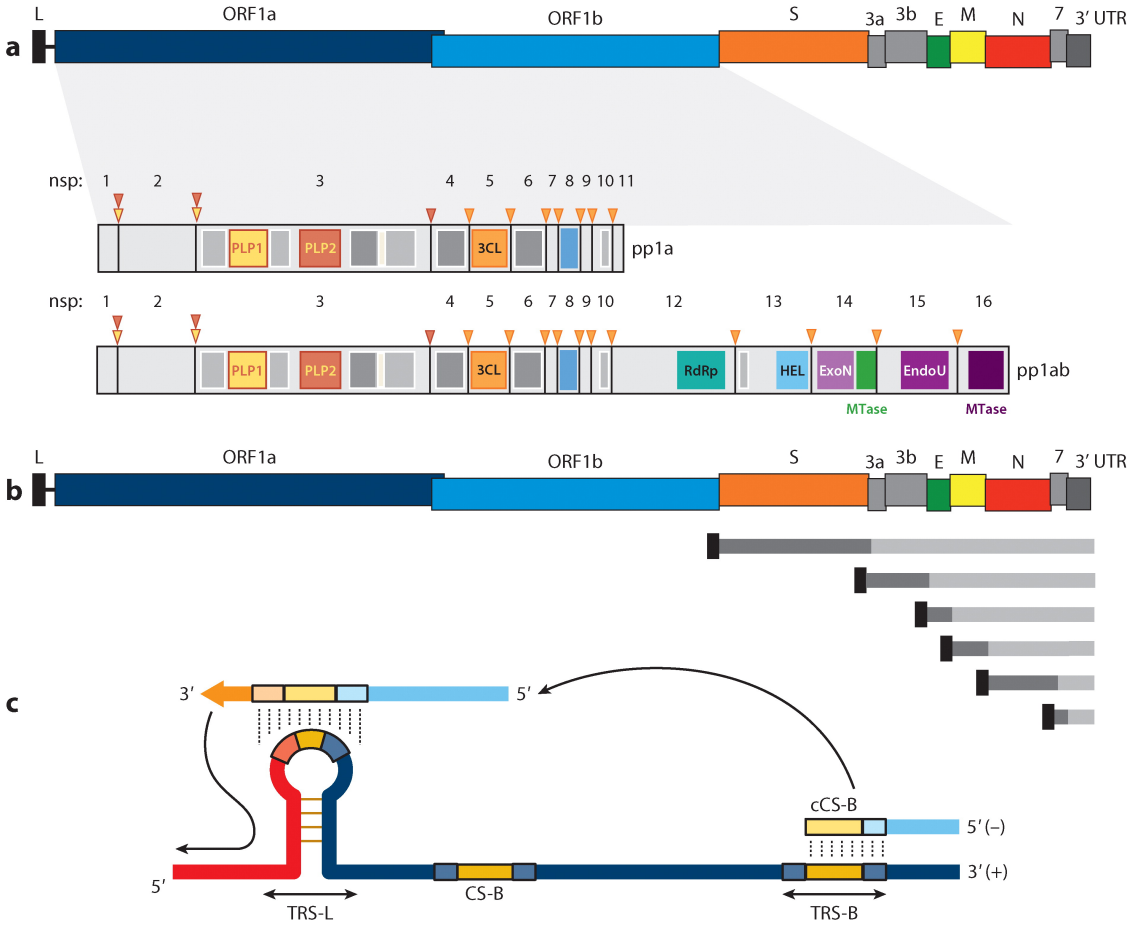
Genome organization

All coronaviruses contain non-segmented positive strand RNA genome approx. 30 kb in length. It is invariably 5'-leader-UTR-replicase-S-E-M-N-3'UTR-poly(A). In addition, it contains a variety of accessory proteins interspersed throughout the genome.



Genomic organization of representative α, β, and γ CoVs. An illustration of the MHV genome is depicted at the top. The expanded regions below show the structural and accessory proteins in the 3′ regions of the HCoV-229E, MHV, SARS-CoV, MERS-CoV and IBV. Size of the genome and individual genes are approximated using the legend at the top of the diagram but are not drawn to scale. HCoV-229E human coronavirus 229E, MHV mouse hepatitis virus, SARS-CoV severe acute respiratory syndrome coronavirus, MERS-CoV Middle East respiratory syndrome coronavirus, IBV infectious bronchitis virus. (From Fehr and Perlman:2015)

The long replicase gene encodes a number of non-structural proteins (nsps) and occupies 2/3 or the genome. Because of -1 ribosomal frameshifting within ORFab approx. in 25% of the cases polyprotein pp1ab is produced in place of pp1a. pp1a encodes 11 nsps while pp1ab encodes 15.

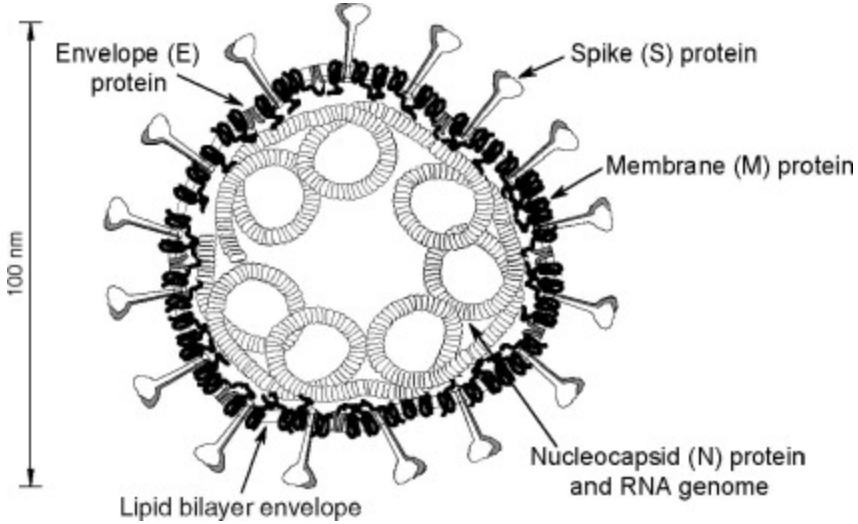


Coronavirus genome structure and gene expression. (a) Coronavirus genome structure. The upper scheme represents the TGEV genome. Labels indicate gene names; L corresponds to the leader sequence. Also represented are the nsps derived from processing of the pp1a and pp1ab polyproteins. PLP1, PLP2, and 3CL protease sites are depicted as inverted triangles with the corresponding color code of each protease. Dark gray rectangles represent transmembrane domains, and light gray rectangles indicate other functional domains. (b) Coronavirus genome strategy of sgmRNA expression. The upper scheme represents the TGEV genome. Short lines represent the nested set of sgmRNAs, each containing a common leader sequence (black) and a specific gene to be translated (dark gray). (c) Key elements in coronavirus transcription. A TRS precedes each gene (TRS-B) and includes the core sequence (CS-B) and variable 5′ and 3′ flanking sequences. The TRS of the leader (TRS-L), containing the core sequence (CS-L), is present at the 5′ end of the genome, in an exposed location (orange box in the TRS-L loop). Discontinuous transcription occurs during the synthesis of the negative-strand RNA (light blue), when the copy of the TRS-B hybridizes with the TRS-L. Dotted lines indicate the complementarity between positive-strand and negative-strand RNA sequences. Abbreviations: EndoU, endonuclease; ExoN, exonuclease; HEL, helicase; MTase, methyltransferase (green, N7-methyltransferase; dark purple, 2′-O-methyltransferase); nsp, nonstructural protein; PLP, papain-like protease; RdRp, RNA-dependent RNA polymerase; sgmRNA, subgenomic RNA; TGEV, transmissible gastroenteritis virus; TRS, transcription-regulating sequence; UTR, untranslated region. (From Sola:2015).

Sola I, et al. 2015.
Annu. Rev. Virol. 2:265–88

Virion structure

Coronovirus is a spherical particle approx. 125 nm in diameter. It is covered with S-protein projections giving it an appearance of solar corona - hence the term coronavirus. Inside the envelope is nucleocapsid that has helical symmetry far more typical of (-)-strand RNA viruses. There are four main structure proteins: spike (S), membrane (M), envelope (E), and nucleocapsid (N).



Schematic of the coronavirus virion, with the minimal set of structural proteins (From Masters:2006).

The mature S protein is trimer of two subunits: S1 and S2. The two subunits are produced from a single S-precursor by host proteases (see Kirchdoerfer:2016; this however is not the case for all coronaviruses such as SARS-CoV). S1 forms the receptor-binding domain, while S2 forms the stalk.

M protein is the most abundant structural components of the virion and determines its shape. It possibly exists as a dimer.

E protein is least abundant protein of the capsid and possess ion channel activity. In facilitates assembly and release of the virus. In SARS-CoV it is not required for replication but is essential for pathogenesis.

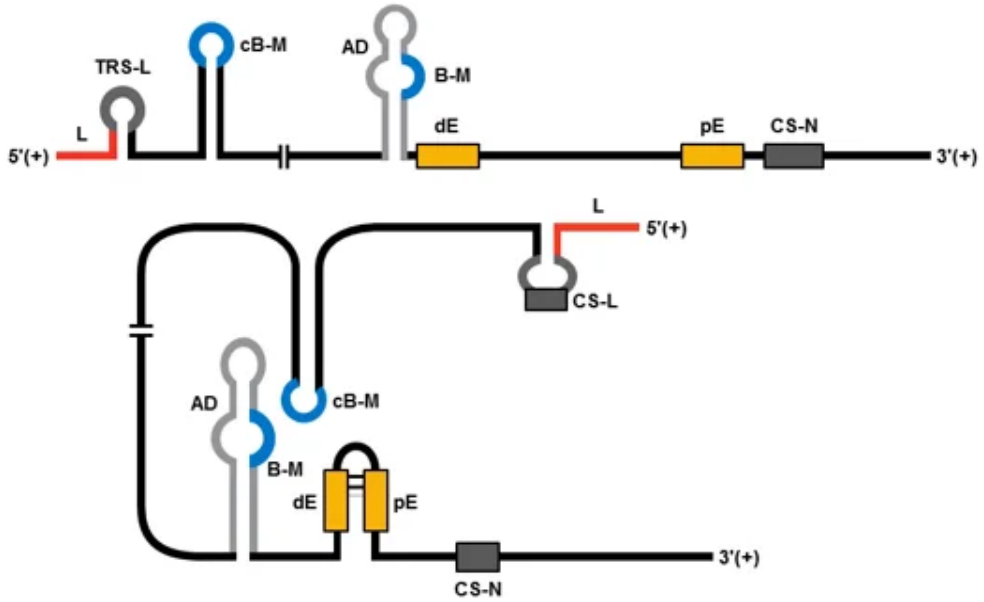
N proteins forms the nucleocapsid. It N- and C-terminal domains are capable of RNA binding. Specifically it binds to transcription regulatory sequences and the genomic packaging signal. It also binds to `nsp3` and M protein possible tethering viral genome and replicase-transcriptase complex.

Entering the cell

Virion attaches to the cells as a result of interaction between S-protein and a cellular receptor. In case of SARS-CoV-2 (as is in SARS-CoV) angiotensin converting enzyme 2 (ACE2) serves as the receptor binding to C-terminal portion of S1 domain. After receptor binding S protein is cleaved at two sites within the S2 subdomain. The first cleavage separates receptor-binding domain of S from membrane fusion domain. The second cleavage (at S2') exposes the fusion peptide. The SARS-CoV-2 is different from SARS-CoV in that it contains a furin cleavage site that is processed during viral maturation within endoplasmatic reticulum (ER; see Walls:2020). The fusion peptide inserts into the cellular membrane and triggers joining on the two heptad regions with S2 forming antiparallel six-helix bundle resulting in fusion and release of the genome into the cytoplasm.

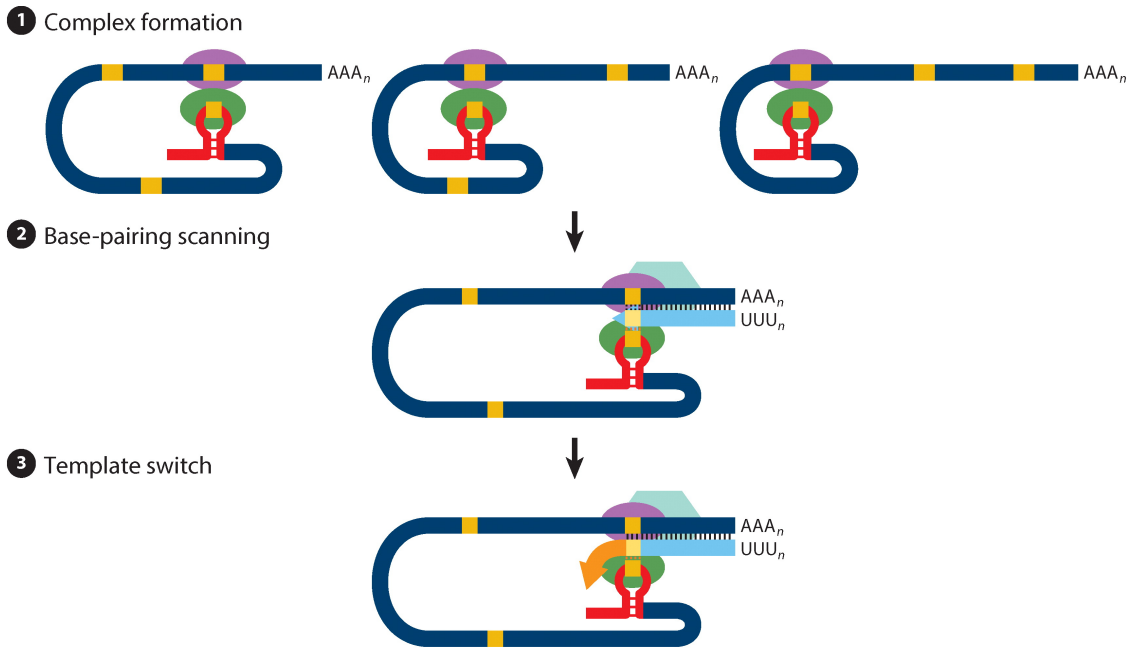
Replication

The above figure shows that in addition to the full length (+)-strand genome there is a number of (+)-strand subgenomic RNAs corresponding to 3'-end of the complete viral sequence. All of these *subgenomic* RNAs (sgRNAs) share the same leader sequence that is present only once at the extreme 5'-end of the viral genome. These RNAs are produced via discontinuous RNA synthesis when the RNA-dependent RNA-polymerase (RdRp) switches template.



Model for the formation of genome high-order structures regulating N gene transcription. The upper linear scheme represents the coronavirus genome. The red line indicates the leader sequence in the 5′ end of the genome. The hairpin indicates the TRS-L. The gray line with arrowheads represents the nascent negative-sense RNA. The curved blue arrow indicates the template switch to the leader sequence during discontinuous transcription. The orange line represents the copy of the leader added to the nascent RNA after the template switch. The RNA-RNA interactions between the pE (nucleotides 26894 to 26903) and dE (nucleotides 26454 to 26463) and between the B-M in the active domain (nucleotides 26412 to 26421) and the cB-M in the 5′ end of the genome (nucleotides 477 to 486) are represented by solid lines. Dotted lines indicate the complementarity between positive-strand and negative-strand RNA sequences. Abbreviations: AD, active domain secondary structure prediction; B-M, B motif; cB-M, complementary copy of the B-M; cCS-N, complementary copy of the CS-N; CS-L, conserved core sequence of the leader; CS-N, conserved core sequence of the N gene; dE, distal element; pE, proximal element; TRS-L, transcription-regulating sequence of the leader (From Sola:2015).

Furthermore Sola:2015) suggest propose the coronavirus transcription model in which transcription initiation complex forms at the 3'-end of (+)-strand genomic RNA.



Sola I, et al. 2015.
Annu. Rev. Virol. 2:265–88

Three-step model of coronavirus transcription. (1) Complex formation. Proteins binding transcription-regulating sequences are represented by ellipsoids, the leader sequence is indicated with a red bar, and core sequences are indicated with orange boxes. (2) Base-pairing scanning. Negative-strand RNA is shown in light blue; the transcription complex is represented by a hexagon. Vertical lines indicate complementarity between the genomic RNA and the nascent negative strand. (3) Template switch. Due to the complementarity between the newly synthesized negative-strand RNA and the transcription-regulating sequence of the leader, template switch to the leader is made by the transcription complex to complete the copy of the leader sequence (From Sola:2015).

So what can we do?

Given this information we ask the following questions:

1. Can we reconstruct individual transcripts given the SARS-CoV data?
2. Can we observe and confirm the ribosomal frameshifting?
3. Can we identify secondary structures within SARS-CoV RNAs?
4. .... Suggest your own question ....

Yes, we can do all these things. The next lecture will start with question I: transcript reconstruction.