Turkish Journal of Immunology - Open Access
www.turkishimmunology.org

Turk J Immunol 2020; 8(2):73–88

Molecular Biology of SARS-CoV-2

SARS-CoV-2'nin Moleküler Biyolojisi

Sofia PIEPOLI[1], Bahar SHAMLOO[2], Aylin BİRCAN[1], Ogun ADEBALİ[1], Batu ERMAN[1]

Abstract

The Severe Acute Respiratory Syndrome Corona Virus-2 (SARS-CoV-2) has resulted in the COVID-19 pandemic, which is currently wreaking havoc in human societies. To understand how this virus causes disease, the molecular biology of the virus needs to be studied in more detail. There is a large body of work on the molecular strategies of previous coronaviruses that infected humans, which can be directly applied to SARS-CoV-2. In the current review we highlight the novel aspects of the SARS-CoV-2 coronavirus life cycle, and how this and other viruses interact with the biochemistry of the host organism. We provide a discussion of different types of viruses as a background to understand coronaviruses. Specifically, we compare the life cycles of coronaviruses with that of a model retrovirus, the Human Immunodeficiency Virus (HIV). We describe the genomic, transcription and translation control features of coronaviruses with a focus on protein structures and activities that can be selected as molecular targets of therapy.

Keywords: SARS-CoV-2, COVID-19, coronavirus, HIV, Baltimore Classification, ORF10

Öz

Severe Acute Respiratory Syndrome Corona Virus-2 (SARS-CoV-2) virüsünün neden olduğu Şiddetli Akut Solunum Sendromu ve COVID-19 pandemisi halen sınır tanımadan büyük zararlar vermektedir. Virüsün hastalığa nasıl neden olduğunu anlayabilmek için moleküler biyolojisini daha detaylı çalışmak gerekmektedir. Daha önceki koronavirüs salgınlarına neden olan farklı virüslerin kullandığı moleküler stratejiler hakkında pek çok bilgi bulunmakta ve bu bilgiler doğrudan SARS-CoV-2'yi anlamamıza yardım etmektedir. Bu derlemede SARS-CoV-2 yaşam döngüsüne özel mekanizmaları özetleyerek bu ve benzer virüslerin enfekte ettikleri hücrelerdeki biyokimya ile nasıl etkileşime girdiklerini açıkladık. Değişik virüslerin kullandıkları mekanizmaları da özetleyerek koronavirüslerin daha iyi anlaşılmasına neden olacak bilgileri derledik. Özellikle çok çalışılmış bir retrovirüs model sistemi olan İnsan Bağışıklık Yetmezliği Virüsünün (İBYV) yaşam döngüsü ile koronavirüslerinkini karşılaştırmalı olarak analiz ettik. Bu derlemede koronavirüslerin genomik, transkripsiyon ve translasyon mekanizmalarını açıkladık ve tedavi hedefi olarak kullanılabilecek protein yapıları ve aktivitelerini detaylı bir şekilde irdeledik.

Anahtar Sözcükler: SARS-CoV-2, COVID-19, koronavirüs, HIV, Baltimore Klasifikasyonu, ORF10

[1]Sabanci University Faculty of Engineering and Natural Sciences, Molecular Biology, Genetics and Bioengineering Program, Istanbul, Turkey

[2]Huntsman Cancer Institute, University of Utah, Department of Oncological Sciences, Salt Lake City, Utah, USA

Correspondence:
Batu ERMAN
Molecular Biology, Genetics and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabanci University Orhanli, Tuzla, Istanbul 34956, Turkey
E-mail: batu@sabanciuniv.edu

Received: Jul 10, 2020
Accepted: Aug 04, 2020

https://doi.org/10.25002/tji.2020.1293

Available online at
http://www.turkishimmunology.org

Introduction

The Severe Acute Respiratory Syndrome Corona Virus-2 (SARS-CoV-2) pandemic has rapidly made virology, immunology and epidemiology very popular fields of research. There are many excellent reviews on the molecular biology and the cellular interactions of coronaviruses.[1–5] The aim of the current review is to highlight some novel aspects of the SARS-CoV-2 coronavirus life cycle, and how this and other viruses interact with the biochemistry of the host organism. Herein, we provide a special focus on viral molecular strategies and structural features of coronaviral proteins, which should serve as a starting point for studying SARS-CoV-2.

Baltimore Classification of Enveloped Viruses

Different conventions exist to classify viruses. Historically, shared properties such as morphology of the virion, or sequence alignment were used to classify viruses into

Turk J Immunol 2020; 8(2):73–88

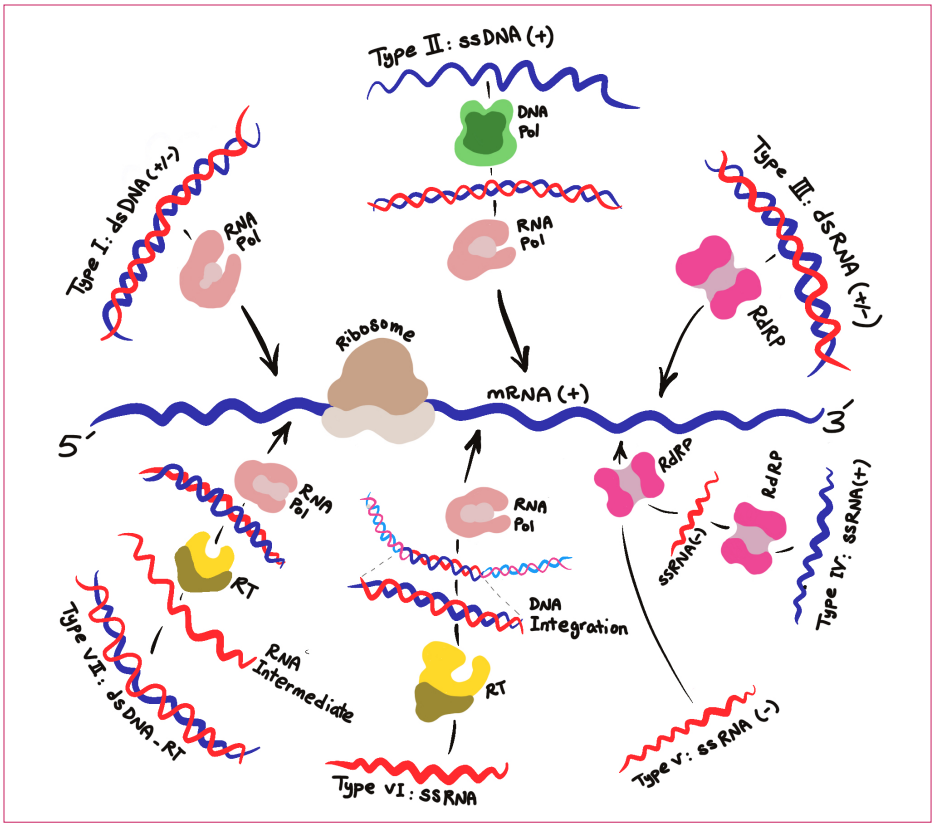

Figure 1. The Baltimore Classification of Viruses. Nobel Prize Laureate, virologist David Baltimore's classification system is commonly used to classify all viruses. Because viruses are obligate parasites dependent on the cell that they infect, they need to convert their genetic material into positive strand mRNA that can be read by the cellular ribosome *(center)*. Clockwise from top left, class I-VII, with their nucleic acid genomes shown either as positive strands *(blue)* or negative strand *(red)* that are either single stranded *(ss)* or double stranded *(ds)*. Some well-known members of the six classes are listed in Table 1. Belonging to the same class does not imply sequence or functional similarity between viruses (RNA Pol, RNA polymerase; DNA Pol, DNA polymerase; RdRP, RNA dependent RNA polymerase; RT, Reverse transcriptase).

different taxonomic ranks.[6] Because the viral life cycle switches between an inert viral particle and a replicative state inside the host cell, a mechanistic, rather than a structural classification seems more suitable.[7] The Baltimore Classification system[8] does exactly that, sorting viruses according to the type of nucleic acid that gets packaged into the virion (Figure 1). The two main limiting factors for viruses are genome size and physical size of the viral particles. These factors limit the number of molecules that can be packaged into the virion. Because of these limitations, all viruses are obligate parasites that depend on the biochemistry of the cells that they infect. Converting the viral genotype (nucleic acid) into a phenotype (protein), according to the Central Dogma of Molecular Biology, requires translation by cellular ribosomes.[9] While some ribosomal proteins and tRNAs may be encoded by some viruses, the typical virion is too small to pack a functional ribosome.[10] Thus, regardless of the type of nucleic acid molecule the viral genome is encoded in, it has to be converted into a positive strand, "sense" RNA molecule (mRNA) that cellular ribosomes translate into protein.

All viruses use different strategies to convert their genetic material to RNA, regardless of the type of nucleic acid in which their genetic material is stored.[11] In so doing, different classes require the activity of either cellular RNA polymerases (viral class I and II) or viral RNA dependent RNA polymerases (RdRP) (viral class III, IV and V) or in the case of retroviruses, a viral Reverse Transcriptase (RT) followed by cellular RNA polymerases (viral class VI and VII) (Table 1). For packaging new viral particles, the original nucleic acid genome has to be copied and amplified.[11] The molecular problem that needs to be solved here is that positive or negative strand nucleic acid genomes cannot be directly copied to generate exact replicas. First, they have to be turned into a daughter strand template and then converted back to the original strand. This is because all polymerases (DNA, RNA, RT, RdRP) read their templates in the 3' to 5' direction and synthesize the new daughter strands in the 5' to 3' direction.[12] Regardless of the identity of the template, the newly synthesized copy polymerizes by forming a new phosphodiester bond between the 3' OH group of the last base of the growing strand and the 5' phosphate group of the nucleotide that is added.[9] Because of this limitation, a positive strand RNA virus cannot directly copy its genome by duplication, it needs to first synthesize a complementary template copy of its genome which in turn is used to synthesize duplicate genomes. Whereas class IV positive

Table 1. The Baltimore Classification system compares viruses by the type of nucleic acid packaged into the virion structure. Different viruses use different strategies and enzymes to convert these genomes into mRNA molecules that can be converted to proteins by cellular ribosomes.

Class	Viral genome	Intermediates	Replication enzymes	Example species
I	Double strand DNA		DNA Dependent RNA Polymerase	Herpesvirus, Polyomavirus (SV40), Papillomavirus, Cytomegalovirus (CMV), Pox viruses (vaccinia) T4, T7 bacteriophages
II	Single strand DNA	Double strand DNA	DNA Dependent DNA Polymerase	Adeno associated virus (AAV), bacteriophage M13
III	Double strand RNA		RNA Dependent RNA Polymerase	Rotavirus
IV	Positive strand RNA	Negative strand RNA	RNA Dependent RNA Polymerase	Coronavirus, Hepatitis A/C/E viruses, Poliovirus, Rubella virus, Tobacco mosaic virus
V	Negative strand RNA		RNA Dependent RNA Polymerase	Ebola, Marburg, Lyssavirus (Rabies), Morbillivirus (Measles), Rubulavirus (Mumps), Influenza virus A/B/C, Lymphocytic choriomeningitis (LCMV)
VI	Positive strand RNA	Negative strand DNA Double strand DNA	Reverse Transcriptase + DNA Dependent RNA Polymerase	Mouse mammary tumor virus (MMTV), Murine leukemia virus (MLV), Human immunodeficiency virus (HIV)
VII	Double strand DNA	Positive strand RNA Double strand DNA	Reverse Transcriptase + DNA Dependent RNA Polymerase	Hepatitis B virus

strand RNA viral genomes such as that of SARS-CoV-2, can be directly accessed by the host ribosomes, all other classes have to first convert their genome emanating from the virion, into a positive strand RNA molecule that can be translated.

Viruses are Small and Obey the Rules of Their Host's Molecular Mechanisms

Viruses are small in genome size and in physical size. Virus diameter varies between 20 and 400 nanometers.[11] A typical human cell, for comparison, measures on average 20 micrometers. This simplicity presumably gives viruses a speed advantage when replicating their genomes and assembling their virions. The challenge for a virus is to pack as many proteins as possible into a small virion that gives it a replicative advantage immediately after fusion into the target cell. But more proteins in a virion means larger genomes that encode them and larger virions that have to assemble. The size of the SARS-CoV-2 is ~100 nm, typical for a coronavirus and average when all viruses are considered.[13] However, with a genome size of about 30 kilo bases, Coronaviruses contain the largest RNA genomes.[14] A large genome may encode much more complex biochemistry but also increases the number of mutations that accumulate, clearly a selective advantage and disadvantage respectively. While overall viral mutation rate depends on the frequency of replication per cell division and environmental pressure, the base mutation rate depends on the type of enzyme that replicates the viral genome. For DNA and RNA viruses, the mutation rate is around 10^{-7} and 10^{-4} substitutions per nucleotide per cell infection respectively.[15] For large genomes such as coronaviruses, these mutation rates could be prohibitive in sustaining genome integrity. Unlike retroviruses with small genome sizes that rely on high mutation rate reverse transcriptases, coronaviruses, with larger size genomes, use an RNA dependent RNA polymerase (RdRP) that has proofreading activity.[11] The SARS-CoV-2 RdRP is encoded by the viral non-structural protein Nsp12 gene and the protein product of the Nsp14 gene is an exonuclease responsible for the proofreading mechanism.[16]

Eukaryotic ribosomes can only translate RNA molecules that are 5'capped and can only initiate translation at a start codon near the 5'end of the RNA.[9] At first sight, positive strand RNA viruses (class IV) have an advantage, as immediately after fusion, their genome can be translated by ribosomes without the need for further amplification or copying. However, eukaryotic ribosomes end translation and dissociate from the mRNA when they reach a termination codon. This is a problem for this class of virus, because without alternative mechanisms of translation, regardless of the number of genes encoded in the genome, only a single polypeptide encoded at the 5'end of the viral genome can be directly translated. In the case of coronavirus, this is open reading frame one (ORF1). But there are other ORFs with their individual start codons in the coronavirus genome. How can these be translated from the viral RNA if eukaryotic ribosomes

Table 2. Numerical comparison of coronavirus outbreaks in recent years

Corona Virus induced disease	Outbreak	Confirmed cases	Deaths	Average mortality rate	Source
SARS	2003	8098	774	9.5%	Ref. [17]
MERS	2012-	2494	858	34.4%	Ref. [18]
COVID19 30/07/2020	2019-	>17 million	>670000	3.9%	Ref. [19]

cannot assemble onto their internal start codons? Some viruses solve this problem by having multiple RNA molecules that are packed into the virion that are translated by independent ribosomes, others contain internal ribosome entry site (IRES) sequences.[20,21] But not all viral ORFs have IRES elements in front of them and there must be an evolutionary price for containing such a long non-coding IRES sequence in the relatively small genome of a virus. We discuss how coronaviruses specifcally solve this problem in the pursuing sections focusing on the life cycle and genomic structure of the SARS-CoV-2.

SARS-CoV-2 is not a New Virus

The early days of the COVID-19 (Coronavirus Disease 2019) pandemic brought with it many discussions about the origin of the SARS-CoV-2 virus.[22] Human coronaviruses (HCoVs), can be separated into two families, alpha (HCoV-229E, HCoV-NL63), and betacoronavirus (HCoV-OC43, HCoV-HKU1, SARS-CoV and MERS-CoV). These viruses are known to be responsible for about 10% of seasonal colds in humans. The latter two were responsible for severe acute respiratory syndrome epidemics in 2002 (SARS) and 2012 (MERS) respectively.[23,24] These human viruses are related to bat coronaviruses. The Virus Pathogen Resource[25] lists more than 300 species of coronavirus, many of them in bats.[26-28] More information about SARS-CoV, MERS-CoV and SARS-CoV-2 and the diseases they cause are listed in Table 2. An easily detectable difference between these viruses is the insertion of a putative polybasic cleavage site in the spike protein that is surrounded by O-linked glycosylation sites and mutations in the receptor binding domain (RBD) that potentially explain the different infectivity of these viruses.[29] This same polybasic insertion may be important in the immune responses against the different viruses, potentially encoding a superantigenic peptide.[30] As superantigens are typically associated with a polyclonal T lymphocyte response to HLA type II, how such a viral peptide results in this immune activation is not yet clear. With a total of 380 nucleotide substitutions,[22] SARS-CoV, which was responsible for the 2002–2003 outbreak, is the most similar human virus to SARS-CoV-2. A unique addition to the SARS-CoV-2 is the last ORF in its genome all the way at the 3'end, named ORF10 which potentially encodes a very short protein of no known function. Our analysis of this ORF is detailed in the last section of the manuscript.

The Life Cycles of HIV vs Coronavirus

The AIDS pandemic which started in the early 1980's has infected approximately 75 million humans[31] and has resulted in a tremendous amount of knowledge accumulated on the HIV-1 virus. As of July 2020, the current SARS-CoV-2 pandemic has infected more than 17 million people in less than one year and also resulted in an explosion of research on coronaviruses.[32] While the HIV and coronavirus life cycles have drastic differences, it is helpful to compare and contrast the strategies of these viruses and the therapeutic options targeting them (Figure 2). Both HIV-1 and coronavirus are enveloped viruses that use lipidic components of the host cell to encapsulate the core of the virion. As such, both of these viruses are dependent on the eukaryotic post-translational modifications necessary for membrane budding and transmembrane insertion of proteins. The endoplasmic reticulum (ER) and the ER-Golgi intermediate compartment (ERGIC) is a critical organelle for the coronavirus. The name coronavirus comes from the presentation of viral particles in electron microscopic pictures that was compared to the solar corona during an eclipse.[33] This picture is dependent on the viral Spike transmembrane protein. Coronaviruses are morphologically characterized as large (~85 nm diameter), homogeneously spherical virions, containing one of the largest known viral genomes encapsulated in a lipidic envelope almost twice as thick as normal biological membranes.[34] A model coronavirus is the mouse hepatitis virus (MHV).[34] The internal layer of the virion membrane is filled by the M proteins (Membrane glycoproteins) that contact both the internal ribonucleoproteins (RNP) with the C-terminus and the external membrane in which they are embedded with 3 transmembrane domains.[35] Cryo

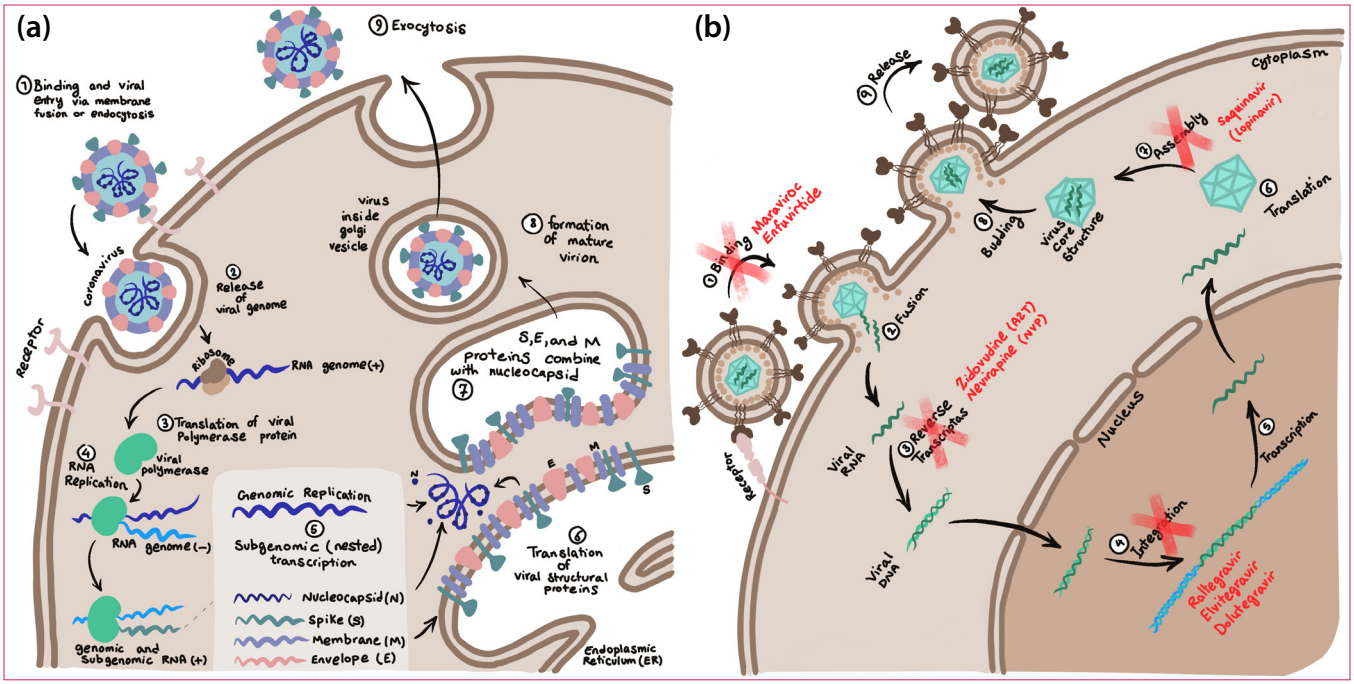

Figure 2. The Life cycles of lentiviruses compared to coronaviruses: Coronavirus **(a)**. 1) The virion attaches to the ACE-2 receptor (pink membrane protein) though its surface Spike protein (blue). The surface of the virion contains other exposed proteins such as Membrane (pink). 2) Endocytosis and viral membrane fusion results in the release of the positive strand RNA genome into the host cell cytoplasm and ribosomes have access to the start codon of ORF1. 3) Translation of ORF1 results in the synthesis of the viral RdRP among other proteins. 4) RdRP synthesizes a negative strand template copy of the viral genome and more positive strand genomes in addition to 5) subgenomic transcripts that encode Nucleocapsid, Spike, Membrane and Envelope proteins. 6) The translation of viral proteins that have signal sequences or start transfer sequences occurs in the rough endoplasmic reticulum. Viral proteins such as nucleocapsid (N) are synthesized in the cytoplasm. 7) Viral transmembrane proteins cause membrane bending similar to the formation of autophagosome multivesicular bodies. 8) Mature virions are packed within these bodies and 9) exocytosis results in the release of new virions to the extracellular milieu. HIV **(b)**. 1) Binding of HIV to the CD4 positive helper T lymphocyte is through the gp120 viral envelope glycoprotein (brown trimer) targets the CD4 protein (pink single pass transmembrane protein) and CCRV (not shown). 2) Viral membrane fusion releases the capsid (green crystalline structure) and two copies of the viral genome and two molecules of reverse transcriptase enzyme into the cellular cytoplasm. 3) This viral genome is reverse transcribed into first a first strand cDNA copy and then into a double stranded DNA molecule, which is imported through the nuclear pore structure in the nucleus and 4) is integrated randomly to accessible sites in the human genome. Viral long terminal repeat (LTR) sequences on either end of the integrated viral genome act as strong promoters that synthesize 5) an RNA copy of the lentiviral genome which translocates into the cytoplasm using the same pathway as cellular mRNAs. The viral genome is translated 6) and is packaged 7) in the cytoplasm and buds 8) from the plasma membrane using the exocytosis pathway resulting in 9) the release of new virions. Similar to coronavirus membrane proteins the gp120 viral envelope glycoprotein is synthesized in the ER and is transported to the plasma membrane using the exocytic pathway (not shown). Components of the highly active anti-retroviral therapy and the lifecycle steps they inhibit are shown in red.

electron microscopy reveals that a typical coronavirus virion contains between 50–100 Spike proteins.[36] But these are not essential for the packing of the virion and only a few Spike proteins (~10) are sufficient for efficient infection of target cells when the ligand binding domain of this Spike binds the Angiotensin-converting enzyme-2 (ACE-2) target.[34]

The Spike protein is responsible for the attachment of the viral particle to target cells that express its ligand, the ACE-2 receptor on their surface.[37] A second target cell membrane component is Transmembrane protease, serine 2 (TMPRSS-2), a type II transmembrane protease that cleaves and activates the Spike protein.[38] The fact that TMPRSS-2 is an androgen controlled protease, previously linked to translocations in prostate cancer, could offer an explanation for the male bias of COVID-19 disease.[39-41] The Spike protein is a type I single-pass transmembrane protein with its N-terminus facing outside and its C-terminus facing inside of the virion. Like the HIV gp120, the Spike forms a trimeric structure whose ectodomain is highly glycosylated in the Golgi compartment.[42] This glycosylated Spike protein is cleaved in this intracellular compartment by Furin like proteases that generate the form that can bind to the ACE-2 receptor when it is packaged into the membrane of the virion.[43] Which organ systems express the ACE-2 and TMPRSS-2 ligands of Spike are

subjects of intense scrutiny.[44-46] Tissue specific and age and gender dependent differences in the expression of these ligands are likely causes for the differential intensity of disease in different populations.

When the Spike trimer on the virion is cleaved by the TMPRSS-2 protease, it undergoes a conformational shift that releases an alpha helical fusion peptide that is targeted to the host cell membrane.[47,48] This strategy is shared by many viruses that result in an ordered series of events that culminate in the fusion of the viral membrane with that of the host cell.[47] When this happens, for the coronaviruses, the positive strand RNA gains access to the cellular translation machinery. The cellular encoded Golgi Furin proteases and the TMPRSS-2 should not be confused with the virally encoded Papain-like Protease (PLP) and 3CL Main Protease. PLP and 3CL are cytoplasmic proteases encoded by the viral genome ORF1a and are necessary for releasing the ten polypeptide chains from the protein product of this ORF1a.[49] On the other hand, Furin like proteases and TMPRSS-2 are membrane associated proteases encoded by the target cellular genome that process the viral proteins during virion assembly and cell fusion steps of the life cycle.[37]

An immediate early gene that is translated from the positive strand coronavirus RNA genome is the Nsp12 encoded RNA dependent RNA polymerase (RdRP) which is a subcomponent of ORF1.[50] The concurrent translation (by the cellular ribosome) and transcription (by the viral RdRP) is thought to occur in the ERGIC. The RdRP in complex with co-factors NSP7 and NSP8[51-53], is charged with first transcribing a complementary negative strand copy of the viral genome and then amplifying this negative strand copy into many new positive strand copies that end up being packaged into new viral particles. The RdRP is also responsible for the generation of shorter, subgenomic transcription products, which are necessary to solve the aforementioned problem of having a multicistronic viral RNA that can only be translated by a ribosome seeking a single 5' start codon. The mechanism and significance of these subgenomic transcripts are detailed in the section focussing on the genomic structure of SARS-CoV-2.

The lifecycle of the coronavirus can be broken down into nine steps, 1) binding and viral entry by endocytosis or membrane fusion, 2) cytoplasmic release of the viral genome, 3) translation of viral genes from positive strand viral genome, 4) replication of viral genome in the cytoplasm by RdRP, 5) synthesis of sub-genomic transcripts by recombination and RdRP mediated transcription, 6) synthesis and membrane insertion of viral membrane proteins (spike, membrane glycoprotein, envelope protein, nucleocapsid protein and some non-virion nonstructural membrane proteins necessary for replication) on the ER, 7) assembly of the nucleocapsid bound viral RNA genome into vesicles by membrane folding mediated by viral membrane proteins, 8) formation of mature virion inside of double membrane vesicles, 9) shedding of virus particles by exocytosis. Our current knowledge of this viral life cycle does not indicate that the host cell nucleus is involved in any way.[54]

The HIV-1 lifecycle follows a similar path with several major differences.[55] The target specificity of the HIV-1 virus is against CD4 positive helper T lymphocytes. The HIV-1 envelope glycoprotein, gp120 binds to CD4 co-receptor which normally functions in T lymphocyte receptor signaling and the chemokine receptor CCR5.[56] Like the haploid genome coronavirus, the HIV-1 genome is diploid, having two copies of a single stranded RNA molecule in the virion.[11] Unlike coronavirus however, these HIV RNAs are not directly transcribed by cellular ribosomes upon entry into the cytoplasm.[57] Instead, two molecules of reverse transcriptase (RT) protein, packaged into the virion converts the genome first into a single stranded DNA and next into double stranded DNA in the cytoplasm.[11] The DNA copies translocate to the nucleus and with the help of the retroviral integrase protein form a pre-integration complex (PIC) that is randomly inserted into the human cell genome.[58] Long terminal repeats (LTR) on either side of the viral genome function as strong promoter/enhancers when the DNA copy of the viral genome is inserted into the T lymphocyte genome.[11] After genomic integration, LTRs usurp the cellular transcription machinery to synthesize mRNA copies of the genomic contents of the retrovirus. These mRNAs are indistinguishable from cellular mRNAs in terms of 5'G capping and 3'polyadenylation and are translocated into the cytoplasm where cellular ribosomes translate ORFs into viral proteins. Assembly of viral membrane proteins again occurs in the ER, but virion assembly occurs in the cytoplasm and virions are shed by exocytosis, in the process picking up lipids from the plasma membrane that form the envelope of the virion.

Molecular Targets of Therapy

Soon after the beginning of the HIV-1 pandemic, a wonder drug, AZT, a nucleotide analog reverse

transcriptase inhibitor exploded into the market.[59] Dramatic improvements in AIDS disease were recorded but as in other therapies involving long-term drug use, mutant, drug resistant viruses evolved.[60] Currently HIV positive patients are treated with a triple therapy labelled highly active antiretroviral therapy (HAART) that is composed of a cocktail of drugs including: 1) a fusion inhibitor (Maraviroc or Enfuvirtide), 2) a reverse transcriptase inhibitor (Zidovudine-AZT or Nevirapine), 3) genomic integration inhibitors (Raltegravir, Elvitegravir or Dolutegravir) and optionally 4) a virion assembly inhibitor (Lopinavir or Saquinavir).[61] HAART therapy is hugely successful in HIV treatment because of the cocktail of compounds it contains that target different stages of the viral replication cycle (Figure 2b).[61] The current frenzied search for drugs targeting SARS-CoV-2 likely will be fully successful only when a similar combination therapy that targets multiple steps of the coronaviral life cycle is developed.

The logical initial target for vaccines and drugs targeting SARS-CoV-2 is the Spike glycoprotein because it acts in the first step of the viral infection using its ACE-2 receptor binding domain.[37] Another therapeutic target could be the enzymes directly involved in the initial steps of the viral life cycle.[62] PLP (NSP3) and 3CL-Protease (NSP5) are responsible for polyprotein processing; RdRP (NSP12) is the main virus polymerase whose function is absent in human cells.[52] Because human cells also do not have reverse transcriptase activity, drugs (AZT) targeting the HIV RT enzyme were the initial candidates in that pandemic. Remdesivir is a nucleoside analog drug designed to inhibit RdRP enzymatic activity.[63] Differently from HIV, the SARS-CoV-2 genome encodes for its own Helicase (Nsp13), an evolutionary feature that may render such a large viral genome possible and another potential drug target.[64,65] In addition to Spike, virion structural proteins E and N are also known at the molecular structural level.[37,66-73] These potential targets have been computationally screened against the main drug databases.[74] As in the case of HAART therapy of HIV, preferred molecular targets of small molecule inhibitors will likely be enzymes involved in the intracellular life of the virus.

A different therapeutic approach is the treatment of symptoms of the infection, in particular the effects of the acute immune response to the viral infection. In this respect, immunosuppressive therapies such as Anakinra targeting the IL-1R signaling pathway and tocilizumab and siltuximab targeting the IL-6R signaling pathway seem like front runners in the race.[75] While the Spike may be the initial candidate for vaccines or neutralizing antibodies, other membrane proteins of the virion may also be attractive targets. These studies must bear in mind that not all the viral proteins are packaged into the virion and that immune responses against these proteins may damage the infected cells, exacerbating the auto-immune destruction that is so evident in the macrophage activation syndrome and cytokine storm seen in many patients. An extensive catalog of potential peptides encoded by the viral genome that can be presented by major histocompatibility complex (MHC)/human leukocyte antigen (HLA)-I or -II is a good starting point for finding T cell epitopes.[76] Another approach may be to nonspecifically activate the immune system against alternative antigens, especially for health care workers.[77]

The Genomic Structure of SARS-CoV-2

About 70% of the entire coronavirus genome is dedicated to non-structural proteins (NSP) that are expressed in the infected cell but are not packaged into the virion. These are expressed as a single polyprotein and cleaved into their mature versions by the activity of two proteases which are also encoded by this polyprotein. The specificity of the Papain like protease (PLP)-NSP3 and the 3CL main protease-NSP5 are different.[78] By sequence conservation, the former has three targets in ORF1a, releasing NSP1, NSP2 and NSP3 from the polyprotein.[76,77] The second protease, the 3CL main protease can release the NSP4, NSP5 (encoding itself), NSP6, NSP7, NSP8 and NSP9 from the polyprotein and NSP12-15 from ORF1b. The last protein encoded by this ORF1b polyprotein is in fact the only one that is not cleaved, but instead terminates because of the stop codon located in the transcript just past the slippery sequence.[81] A list of the genes and proteins encoded by the SARS-CoV-2 genome are detailed in Table 3.

SARS-CoV-2 belongs to class IV in the Baltimore classification.[8] The positive strand coronavirus RNA genome, similar to cellular mRNA, is post-transcriptionally processed to contain structures that mimic 5'capping and 3'polyadenylation which allows direct translation of the viral genome by the host cell translation machinery. A single ribosome engagement event produces two polyproteins from the first open reading frame. ORF1a-b contains all the genetic material for the 15 non-structural proteins that are then post-translationally cleaved by two viral proteases (PLP and 3CL-P).[78] The two polyproteins (a-b) are separated by a slippery RNA sequence (UUUAAAC) and

Table 3. The sizes of the genes and protein encoded by the genomes of beta coronaviruses. For other coronavirus gene structures see reference: Tang XC, et al.[82]

	SARS-CoV			MERS-CoV			SARS-CoV-2		
Protein(s) Encoded	Start (bp)	Stop (bp)	aa	Start (bp)	Stop (bp)	aa	Start (bp)	Stop (bp)	aa
ORF1a (NSP1-10)	265	13398	4377	279	13433	4384	266	13468	4400
ORF1b (NSP12-16)	13398	21485	2695	13433	21514	2693	13468	21555	2695
S (Spike)	21492	25259	1255	21456	25517	1353	21563	25384	1273
ORF3a	25268	26092	274	25532	25843	103	25393	26220	275
ORF3b	25689	26153	154						
ORF4a				25852	26181	109			
ORF4b				26093	26833	246			
ORF5				26840	27514	224			
E (Envelope)	26117	26347	76	27590	27838	82	26245	26472	75
M (Membrane)	26398	27063	221	27853	28512	219	26523	27191	222
ORF6	27074	27265	63				27202	27387	61
ORF7a	27273	27641	122				27394	27759	121
ORF7b	27638	27772	44						
ORF8a	27779	27898	39				27894	28259	121
ORF8b	27864	28118	84	28762	29100	112			
N (Nucleocapsid)	28120	29388	422	28566	29807	413	28274	29533	419
ORF10							29558	29674	38

a pseudoknot structure located between the Nsp10 and Nsp12 genes that causes the translating ribosome to pause on the RNA and to recover the read after a -1 frameshift.[81] Viral slippery sequences followed by pseudoknot structures are not unique to coronaviruses; they are also present in the genomes of retroviruses and the herpes simplex virus.[83]

The last gene in SARS-CoV-2 ORF1a is Nsp10 which contains the slippery sequence at its 3'end followed by a stop codon.[81] Ribosomes that translate all of the ORF1a polyprotein pause at this pseudoknot RNA structure immediately following this slippery sequence. Paused ribosomes take a single step back and frameshift (-1) one nucleotide and continue to translate ORF1b, bypassing this Nsp10 stop codon only if they frameshift.[84] Thus the five proteins encoded by ORF1b are actually translated by the same ribosome molecule as the ten proteins in ORF1a. Assuming that the viral RNA is translated by a polysome, it would be interesting to find out potential interactions between different ribosomes translating the molecule, and how many stop at Nsp10 and how many continue on to translate ORF1b by frameshifting. Why such a mechanism is necessary to translate ORF1 is not known.[85] As the ribosome will stall at the pseudoknot the five polypeptide chains encoded after this sequence will be translated less efficiently than the ten preceding it. This imbalance may result in a relative over concentration of the two proteases encoded by Nsp3 and Nsp5, which are located before the frameshift, focusing the lifecycle into chopping the polypeptide chains rather than synthesizing more copies of the RNA catalyzed by the RdRP encoded by Nsp12, which comes after the frameshift. An important step of the viral lifecycle may be the function of Nsp1, an RNAse encoded by ORF1a, before the frameshift, that is responsible for degrading host mRNAs and focusing the infection to escape cellular responses at the early stages of the infection.[86]

Once the Nsp12 encoded RdRP is expressed from the second polyprotein (ORF1b), it takes control of virus genome replication. The original positive strand genome from the virion acts as the template for the RdRP to synthesize a negative complementary strand.[87] The full genome is either replicated into a secondary positive strand RNA that gets packaged into the next generation viruses, or it is replicated into multiple subgenomic mRNAs that become substrates for translation of the nine ORFs in the 3' of the genome. This part of the genome encodes all the structural and accessory proteins necessary for the assembly of the virion.[50] The four main structural

proteins encoded by the viral genome are spike, envelope, membrane and nucleocapsid.

The generation of the subgenomic transcripts is necessary for the translation of coronaviral structural proteins.[50] As mentioned before, eukaryotic ribosomes can only translate new proteins if they assemble onto the first start codon (AUG) after the 5' end of the mRNA molecule. If this relatively complex mechanism did not exist, and all nine ORFs existed in a single mRNA molecule as a polycistronic array, only the first ORF encoding the spike protein in this region of the genome would be translated by a single ribosome assembling on the Spike start codon. This is because the ORFs in this region of the genome all have individual start and stop codons (Figure 3). This is in contrast to the genes encoded by ORF1a and ORF1b which only have a single start codon at the 5' end of the Nsp1 gene and a single stop codon at the 3' end of their last genes, Nsp10 for ORF1a and Nsp16 for ORF1b. The discontinuous transcription activity of the RdRP generating the subgenomic RNA molecules ensures that each transcript has one ORF that is translated.[50] The subgenomic transcripts are generated by an intramolecular recombination event generated by the RdRP which jumps from transcription regulatory sequences (TRS) in the 5' leader sequences upstream of each ORF in this 3' region of the genome (TRS-B) to a homologous sequence upstream of ORF1 (TRS-L).[88] Thus, there is the potential for generating nine different subgenomic RNA molecules all of which share the same 5' end but have different sizes and encode different proteins.[89] The longest subgenomic RNA molecules encode all of the ORFs but effectively only translate the first gene in the array, which happens to be the Spike gene. The second longest subgenomic RNA molecule has ORF3a in its 5' end and can only translate this gene into protein, even though it has the rest of the ORFs in the molecule. As such, each of the progressively smaller subgenomic RNA molecule translate into protein, only its most 5' ORF.

Insertion of Eukaryotic Membrane Proteins into the ER Membrane

Because enveloped viruses use cellular lipids and pack viral proteins with transmembrane domains into lipid vesicles that turn into extracellular virion membranes, they have to usurp the eukaryotic cell's membrane protein insertion machinery. For all eukaryotic cells, membrane proteins are synthesized in the endoplasmic reticulum. Type I single-pass transmembrane proteins have their N-termini facing the extracellular space and their C-termini facing the cytoplasm. Conversely, type II single-pass transmembrane proteins have their C-termini facing the extracellular space and their N-termini facing the cytoplasm.[90] Transmembrane domains can be identified as stretches of 10–15 amino acids with hydrophobic properties and a propensity to form alpha helical structures.[91] Typically type I proteins contain an additional hydrophobic region in their N-terminus which serves as a leader or signal peptide. During the synthesis of the nascent protein, this N-terminal signal peptide is the first thing that comes out of the ribosome and is bound by a cytoplasmic signal recognition particle (SRP), that targets the nascent protein-ribosome complex to the transport channels in the rough endoplasmic reticulum (RER). The nascent peptide is transferred into this translocon channel and the growing peptide is pumped into the RER. An ER luminal signal peptidase enzyme recognizes the signal peptide and cleaves it from the growing polypeptide. This cleavage results in the amino terminal residue of the mature protein being not the methionine encoded by the ATG start codon of the cDNA encoding it. The synthesis of type I transmembrane proteins continues, when the transmembrane domain is synthesized, it is transferred from the channel into the ER membrane, anchoring the protein and serving as a stop transfer sequence.[90]

The SARS-CoV-2 Spike protein is a type I protein with a signal sequence and a C-terminal transmembrane domain. The ORF7a protein belongs to the same class (type I), with its N-terminal 14 amino acids encoding a signal sequence and a transmembrane domain in its C-terminus followed by a positively charged, lysine and arginine rich ER retrieval sequence. As such, these proteins are both synthesized in the RER, but their final destination may differ because of their sorting signals. While ORF7a has an ER retrieval motif, structural similarity to intercellular adhesion molecule 1 and 2 (ICAM-1 and ICAM-2) indicates that this protein may also have a role in binding to Leukocyte Function Associated Molecule 1 alpha (LFA-1), suggesting a cell surface function.[92]

Type II membrane proteins also begin translation in the cytoplasm, but they do not have N-terminal signal peptides. Rather, they are synthesized until their first hydrophobic transmembrane domain associates with the SRP. This nascent protein-SRP complex is recruited to the cytoplasmic face of the ER, through associations between the SRP and SRP receptor. The nascent protein

is transferred into the assembled translocon channel and synthesis ensues, keeping the N-terminus of the protein in the cytoplasm and pumping the C-terminus into the ER lumen. Multipass proteins that have more than one transmembrane domain, follow a similar synthetic pathway to type II proteins. The first hydrophobic transmembrane domain coming out of the ribosome serves as a signal anchor sequence, associating with the SRP and mediating ER recruitment through the SRP receptor. The second hydrophobic transmembrane domain serves as a stop transfer sequence which allows transfer and insertion of both domains into the hydrophobic membrane from the lumen of the channel. Synthesis ensues, and if there are alternative transmembrane regions, they progressively serve as signal anchor and stop transfer sequences, keeping the N- and C-terminus of the protein in the cytoplasm.[90]

Membrane and Secreted Proteins Encoded by the Coronavirus Genome

Hydropathy analysis indicates that the coronavirus viral genome encodes eight proteins with predicted transmembrane domains. These are NSP3, NSP4, NSP6, Spike, ORF3a, Envelope, Membrane, and ORF7a (Figure 3). All of these proteins are synthesized on the cytoplasmic face of the RER and inserted into the RER membrane.

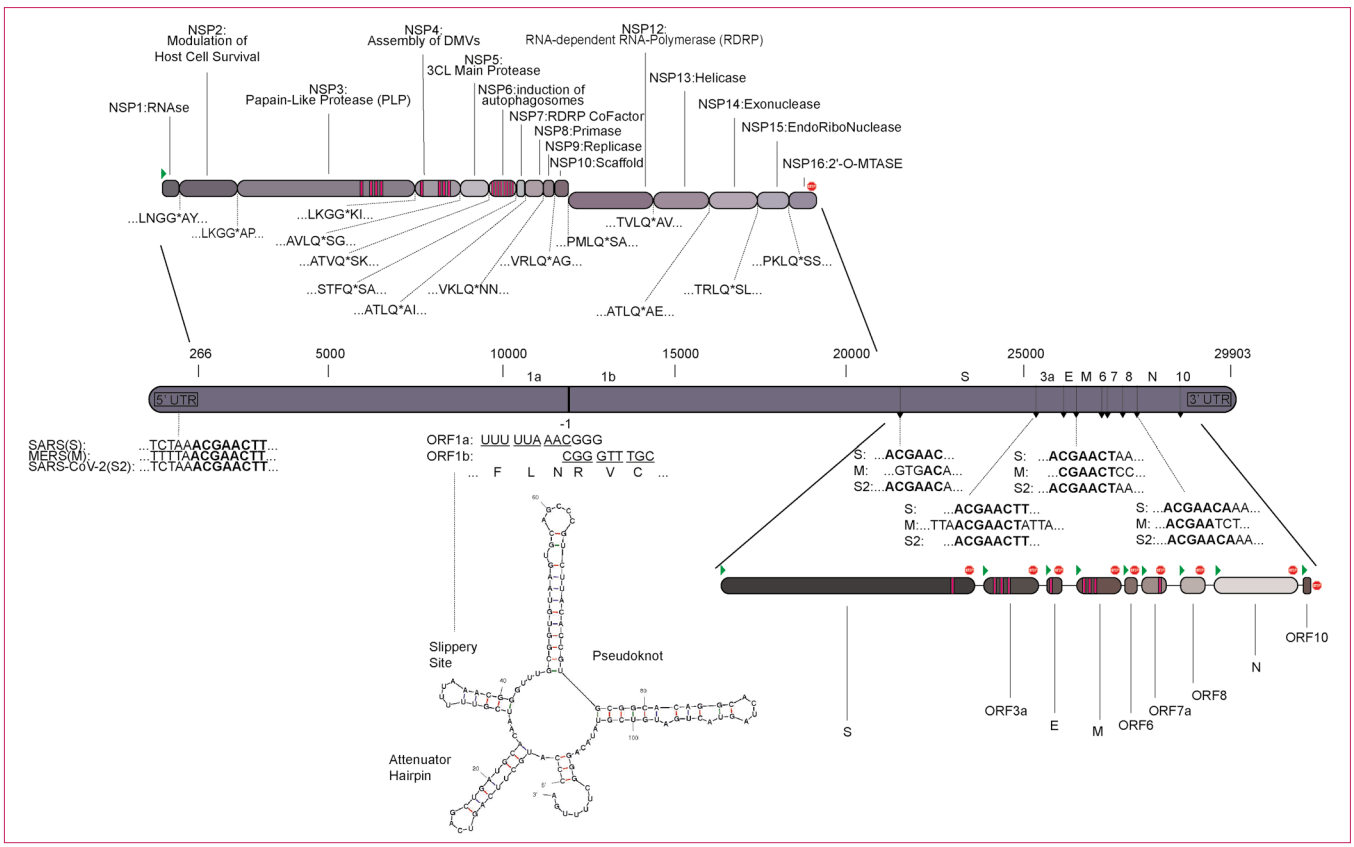

Figure 3. The genome structure of the SARS-CoV-2 coronavirus. The genome is a 29903 nucleotide long positive strand RNA. On either end of this RNA molecule, are the 5'UTR and 3'UTR, which do not code for protein but make structures necessary for translation initiation and polyadenylation. The genome encodes 15 nonstructural proteins from a single ORF at the 5' half and contains 9 ORFs at the 3' end. ORF1a is 13203 nucleotides long and encodes a polyprotein that gets cleaved by at least two viral proteases. In between the ORF1a and ORF1b genes, is a structural pseudoknot structure and slippery sequence that causes a single base pair frameshift. The 3' end of the genome undergoes a process of subgenomic transcription that is a product of recombination between the 5' UTR that behaves as a 5' G cap and the short intergenic sequences between the 3' genes that is homologous to the 5' leader sequence. The homologous sequences between selected ORFs at the 3' of the genome and the 5' leader sequence are indicated by small black arrowheads. For each region of homology, the conserved sequences of the SARS-CoV-1 (S), MERS-CoV (M) and SARS-CoV-2 (S2) are indicated. Each ORF of the genome is indicated by a green arrowhead start codon and a red STOP sign termination codon. The proteins encoded by the ORFs are shown above and below the genomic structure. For the non-structural proteins generated from ORF1a and ORF1b, the peptide sequence at the junction is shown and the protease cleavage site is indicated by an asterisk. The transmembrane domains of the relevant proteins are indicated by different numbers of red lines. The predicted structure of the SARS-CoV-2 slippery sequence, attenuator hairpin, pseudoknot structure and the shifted reading frames of ORF1a and ORF1b are shown below their junction (DMV, double membrane vesicles; OMTASE, O-methyl transferase; ORF, open reading frame).[93]

Spike and ORF7a are type I proteins whose N-terminal signal sequences are presumably cleaved off and have their N-termini facing the inside of the ER and their C-termini exposed to the cytoplasm. The Envelope protein is a type II protein with a single transmembrane domain and an ER lumenal C-terminal domain. NSP3, NSP4, NSP6, ORF3a, and Membrane protein have multiple transmembrane domains.[94-96]

The ER derived double membrane vesicles (DMV) observed in cells infected by various members of the coronavirus family are thought to be induced by the nonstructural proteins NSP3, NSP4 and NSP6.[96,97] Electron micrographic pictures indicate that these multivesicular structures are the site of viral transcription and replication could be functionally analogous to MIIC compartments where MHC II molecules exchange invariant chain peptides with antigenic peptides.[35,98,99] NSP3 is a pleiotropic protein. Its N-terminus encodes the papain like protease (PLP) which is cleaved and remains in the cytoplasm to cleave the different proteins encoded by the ORF1a polyprotein. The cytoplasmic facing protease domain contains an YTGNY motif, previously associated with superantigen activity in other viruses.[100,101] Crystal structures of this motif indicate it to be solvent exposed (PDB entries: 6W9C and 6YVA). The C-terminus of NSP3 has three transmembrane domains and interacts with NSP4 and NSP6 to induce folding of the membranes of the ERGIC vesicles, analogous to the formation of autophagosome vesicles.[97] Both NSP3 and NSP4 have short luminal loops between two transmembrane domains containing residues likely involved in glycosylation and disulfide bond formation.

The Spike protein of the SARS-CoV-2 virus is the focus of numerous studies, as its critical function in viral membrane fusion is obviously a targetable step in viral infection. However, one must also consider the other lives of this protein after the viral RNA is injected into the cytoplasm of infected cells, in the form of a nascent protein being translated in the rough endoplasmic reticulum of the infected cell. The SARS-CoV-2 Spike protein is a type I transmembrane protein with a leader peptide encoded by the N-terminal 22 amino acids and a single transmembrane domain encoded by residues 1223–1245 and a short cytoplasmic tail. Structural studies indicate that this protein trimerizes in the ER after synthesis.[42]

The Curious Case of SARS-CoV-2 ORF10

Evolutionary analyses with available sequences of bat, pangolin and human coronaviruses show that the genome of SARS-CoV-2 has an additional ORF at the genomic 3' end; ORF10.[102] While SARS-CoV lineages have an early stop codon in the reading frame of ORF10, SARS-CoV-2 lineages have a longer ORF10. The presence of potentially protein-coding ORF10 in SARS-CoV-2 is one of its major genomic differences from SARS-CoV. Although it is still unknown whether ORF10 encodes a protein, its sequence is predicted to encode putative protein-like structures. Sequence homology searches using Blast pairwise alignment or PFAM domain matching using *HMMer* did not reveal any conservation.[103] A more sensitive method of profile-profile comparison using *CDvist* with the *hhsearch* option revealed a single domain (YvrJ) that partially matched the putative ORF10 polypeptide sequence, but the statistical significance of this finding is questionable.[104,105] Transmembrane domain prediction failed for the *TMHMM* algorithm but predicted a primary transmembrane domain in the SOSUI algorithm.[106] While transcriptomic studies show no evidence of subgenomic RNA expression from this ORF, its novelty and specificity to SARS CoV-2 warrants further study.[89,107,108] In support of a possible gene product, ribosome mapping (Ribo-seq) analysis revealed a non-negligible translation level of ORF10 based on ribosome footprint densities.[109]

Molecular evolution analysis revealed that the putative ORF10 protein is under positive selection based on the high nonsynonymous over synonymous substitution rate while there was no selection on the truncated ORF10.[102] Thus, positive selection on ORF10 in the SARS-CoV-2 lineage, but not in the SARS-CoV lineage, suggests a potentially functional protein. Another computational analysis suggested that the ORF10 of SARS-CoV-2 evolved because of the mutation of a stop codon at nucleotide 76 and the addition of a new 15 nucleotide long motif in the genomic 3' end.[110] Comparative genomic analysis of closely related coronavirus genomes showed that ORF10 might perform an important function because it has nucleotide-level conservation which extends beyond both sides of the putative ORF.[111]

If expressed, this 38-amino acid putative protein could contain an N-terminal hydrophobic region possibly encoding a signal peptide. If indeed this signal sequence is cleaved, the remaining 15 amino acid long C-terminal

peptide encoded by this ORF could be secreted into the ER lumen of infected cells and indeed could be secreted outside of the infected cell. Curiously the peptide binding groove of MHC-II optimally fits 15 amino acids and whether this peptide is somehow loaded onto the MHC-II is an open question. An MHC II associated CD4 epitope (MGYINVFAFPFTIYS) from ORF10 protein was predicted using the *Tepitool* resource in the immune epitope database (IEDB).[76] Which parts of SARS-CoV-2 sequence is recognized by the human immune system is not entirely known, but these predictions point to a possible immune-modulating function for ORF10. Furthermore, to predict possible CD8 T cell epitopes, a set of the 12 most prominent HLA class I alleles which cover the general population were used to predict binding of SARS-CoV-2 encoded peptides to HLA class I.[76] The IEDB net MHC pan 4.0 EL algorithm was used and for each HLA I allele, the top 1% scoring peptides were selected. In this analysis, ORF10 had three putative CD8 T cell epitopes that could be presented by HLA class I. Whether these epitopes are actually expressed on infected cells is not yet clear.

To experimentally identify $CD4^+$ and $CD8^+$ T cell targets encoded by the ORFs of SARS-CoV-2, sets of overlapping peptides spanning the entire sequence of SARS-CoV-2 were synthesized, and pools of antigenic peptides were tested for lymphocyte reactivity. In this experiment, even though ORF10 had 6 peptides in the pool, none were recognized by CD4 T cells of COVID-19 patients.[112] Also arguing against the presence of a polypeptide encoded by ORF10, two very recent studies tested the antibody responses from COVID-19 patients against SARS-COV-2 proteins by proteome microarray and the luciferase immunoprecipitation system. These studies could not detect significant antibody levels against the ORF10 protein, while antibodies against many other SARS-COV-2 encoded proteins were readily present in convalescent sera.[113,114]

Many viruses have evasion strategies targeting the major histocompatibility complex I or II, often by degradation through targeted ubiquitination.[115] ORF10 could be one such strategy for SARS-CoV-2. Experimentally expressed ORF10 protein was shown to interact with the CUL2 ligase complex and suggest to hijack it to ubiquitinate and degrade restriction factors for replication and pathogenesis.[116] Other possible outcomes for ORF10 are that it may not be cleaved and remain as a membrane protein with putative functions either in the ER, ERGIC, Golgi or the plasma membrane. If so, the topology of this protein in the membrane is not immediately evident by sequence analysis. If the C-terminal extension of the protein is extended into extracellular space, it may easily be used as a biomarker for identifying infected cells if an antibody can be generated against this region.

Conclusion

SARS-CoV-2, is responsible for the ongoing COVID-19 pandemic. In modern history, relatively few diseases diffused fast and geographically widely enough to earn the title of a pandemic. However, it seems very likely that new diseases will emerge and spread worldwide among humans again in the future. Recent viral epidemics often involve a cross-species jump from other animals. Bats are the largest animal reservoir for Coronaviruses.[117,118] With advancing environmental change and increasing rates of human mobility, the ecosystems in which viruses live in equilibrium with host species can more easily be disrupted and cross-species jumps will likely become more frequent. The current pandemic demonstrated that the level of preparedness was insufficient. Yet, at the same time, governments, scientists and healthcare professionals could join forces to ensure immediate data availability. Most of our knowledge about SARS-CoV-2, is modelled on the numerous studies already available on the previous SARS-CoV and MERS-CoV outbreaks (Table 2). Rapid publication of the structural and molecular details of SARS-CoV-2 proteins allowed screening of drug databases for drug repurposing.[116,119,120] Understanding the molecular and cellular biology of coronaviruses will increase the speed at which therapies can be generated against COVID-19 and other viral diseases. It took many years and a global effort to generate HAART therapy against HIV-1 infections. Another global effort to engineer COVID-19 therapies will surely benefit from the faster dissemination of information. An understanding of basic virus molecular biology is fundamental to the design of new therapies, and to be prepared against the next pandemic.

Peer-review: Externally peer-reviewed.

Conflicts of Interest: None

Funding Information: None

Contribution of authors: Concept: BE and OA; Design: BS; Data Collection or Processing: SP and AB; Analysis or Interpretation: SP, AB, BE, and OA; Literature Search: SP and AB; Writing: SP, BE, OA and AB; Critical Review: everyone.

Acknowledgements: We thank all past and present members of the Erman Lab, especially Serkan Belkaya, Emre Deniz, Nazlı Keskin, Ronay Çetin and Minel Arınel for critical comments on the manuscript.

References

1. Masters PS. The molecular biology of coronaviruses. Adv Virus Res 2006;66:193–292. [Crossref]
2. Satija N, Lal SK. The molecular biology of SARS coronavirus. Ann N Y Acad Sci 2007;1102:26–38. [Crossref]
3. Lal SK. Molecular Biology of the SARS-Coronavirus, 1st ed. Berlin Heidelberg: Springer-Verlag; 2010. p.328.
4. Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP. The trinity of COVID-19: immunity, inflammation and intervention. Nat Rev Immunol 2020;20:363–74. [Crossref]
5. Feng W, Zong W, Wang F, Ju S. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2): a review. Mol Cancer 2020;19:100. https://molecular-cancer.biomedcentral.com/articles/10.1186/s12943-020-01218-1
6. International Committee on Taxonomy of Viruses. Virus Taxonomy: 2019 Release, https://ictv.global/taxonomy/ (accessed 25 April 2020).
7. Mahmoudabadi G, Phillips R. A comprehensive and quantitative exploration of thousands of viral genomes. eLife 2018;7:e31955. [Crossref]
8. Baltimore D. Expression of animal virus genomes. Bacteriol Rev 1971;35:235–41. [Crossref]
9. Alberts B. Molecular Biology of the Cell. 6th ed.: W.W. Norton, 2017.
10. Mizuno CM, Guyomar C, Roux S, Lavigne R, Rodriguez-Valera F, Sullivan MB, et al. Numerous cultivated and uncultivated viruses encode ribosomal proteins. Nat Commun 2019;10:752. [Crossref]
11. Flint SJ, Racaniello VR, Rall GF, et al. Principles of Virology, 2. 4th ed.: Wiley, 2015.
12. Steitz TA. DNA polymerases: structural diversity and common mechanisms. J Biol Chem 1999;274:17395–8. [Crossref]
13. Bar-On YM, Flamholz A, Phillips R, Milo R. SARS-CoV-2 (COVID-19) by the numbers. eLife 2020;9. [Crossref]
14. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. Nidovirales: evolving the largest RNA virus genome. Virus Res 2006;117:17–37. [Crossref]
15. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R,. Viral mutation rates. J Virol 2010;84:9733–48. [Crossref]
16. Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y, et al. Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. Proc Natl Acad Sci U S A 2015;112:9436–41. [Crossref]
17. Centers for Disease Control and Prevention. SARS Basics Fact Sheet, https://www.cdc.gov/sars/about/fs-sars.html (accessed 30 July 2020).
18. World Health Organization. Middle East respiratory syndrome coronavirus (MERS-CoV), https://www.who.int/emergencies/mers-cov/en/ (accessed 30 July 2020).
19. Johns Hopkins University & Medicine, Coronavirus Resource Center. Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU), https://coronavirus.jhu.edu/map.html (accessed 30 July 2020).
20. Zhao J, Li Y, Wang C, Zhang H, Zhang H, Jiang B, et al. IRESbase: A Comprehensive Database of Experimentally Validated Internal Ribosome Entry Sites. Genomics Proteomics Bioinformatics 2020;S1672-0229(20)30057-7. [Crossref]
21. Mokrejs M, Vopálenský V, Kolenatý O, Mašek T, Feketová Z, Sekyrová P, et al. IRESite: the database of experimentally verified IRES structures (www.iresite.org). Nucleic Acids Res 2006;34:D125–30. [Crossref]
22. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe 2020;27:325–8. [Crossref]
23. Peiris JS, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet 2003;361:1319–25. [Crossref]
24. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med 2012;367:1814–20. [Crossref]
25. Virus Pathogen Resource. Latest data, tools, and analysis results for Coronaviridae and SARS-CoV-2, https://www.viprbrc.org (accessed June 2020).
26. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res 2012;40:D593–8. [Crossref]
27. Woo PC, Lau SKP, Li KSM, Poon RWS, Wong BHL, Tsoi H, et al. Molecular diversity of coronaviruses in bats. Virology 2006;351:180–7. [Crossref]
28. Han HJ, Wen HL, Zhou CM, Chen FF, Luo LM, Liu JW, Yu XJ. Bats as reservoirs of severe emerging infectious diseases. Virus Res 2015;205:1–6. [Crossref]
29. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med 2020;26:450–2. [Crossref]
30. Cheng MH, Zhang S, Porritt RA, Arditi M, Bahar I. An insertion unique to SARS-CoV-2 exhibits superantigenic character strengthened by recent mutations. bioRxiv 2020. [Crossref]
31. World Health Organization. Global Health Observatory (GHO) data, HIV/AIDS, https://www.who.int/gho/hiv/en/ (accessed 07 July 2020).
32. World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard, https://covid19.who.int (accessed 30 July 2020).
33. NIAID. SARS-CoV-2, https://flic.kr/s/aHsmL6MyYF (accessed June 2020).
34. Barcena M, Oostergetel GT, Bartelink W, Faas FGA, Verkleij A, Rottier PJM. Cryo-electron tomography of mouse hepatitis virus: Insights into the structure of the coronavirion. Proc Natl Acad Sci U S A 2009;106:582–7. [Crossref]
35. Angelini MM, Akhlaghpour M, Neuman BW, et al. Severe acute respiratory syndrome coronavirus nonstructural proteins 3, 4, and 6 induce double-membrane vesicles. mBio 2013; 4 2013/08/15. [Crossref]
36. Neuman BW, Adair BD, Yoshioka C, Quispe JD, Orca G, Kuhn P, et al. Supramolecular architecture of severe acute respiratory syndrome coronavirus revealed by electron cryomicroscopy. J Virol 2006;80:7918–28. [Crossref]
37. Walls AC,Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 2020;181:281–92.e6. [Crossref]

38. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell 2020;181:271–80.e8. [Crossref]

39. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310:644–8. [Crossref]

40. Goren A, McCoy J, Wambier CG, Vano-Galvan S, Shapiro J, Dhurat R, et al. What does androgenetic alopecia have to do with COVID-19? An insight into a potential new therapy. Dermatol Ther 2020:e13365. [Crossref]

41. Grasselli G, Zangrillo A, Zanella A, Antonelli M, Cabrini L, Castelli A, et al. Baseline Characteristics and Outcomes of 1591 Patients Infected with SARS-CoV-2 Admitted to ICUs of the Lombardy Region, Italy. JAMA, 2020;323:1574. [Crossref]

42. Delmas B, Laude H. Assembly of coronavirus spike protein into trimers and its role in epitope expression. J Virol 1990;64:5367–75. [Crossref]

43. Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M. Site-specific glycan analysis of the SARS-CoV-2 spike. Science 2020;33:eabb9983. [Crossref]

44. Hamming I, Timens W, Bulthuis MLC, Lely AT, Navis GJ, van Goor H. Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. J Pathol 2004;203:631–7. [Crossref]

45. Qi F, Qian S, Zhang S, Zhang Z. Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses. Biochem Biophys Res Commun 2020;526:135–40. [Crossref]

46. Ziegler CGK, Allon SJ, Nyquist SK, Mbano IM, Miao VN, Tzouanas CN, et al. SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. Cell 2020;181:1016–35.e19. [Crossref]

47. Kim IS, Jenni S, Stanifer ML, Roth E, Whelan SPJ, van Oijen AM, Harrison SC. Mechanism of membrane fusion induced by vesicular stomatitis virus G protein. Proc Natl Acad Sci U S A 2017;114:E28–36. [Crossref]

48. Zimmerman MI, Porter JR, Ward MD, Singh S, Vithani N, Meller A, et al. Citizen Scientists Create an Exascale Computer to Combat COVID-19. bioRxiv 2020. [Crossref]

49. Ziebuhr J, Schelle B, Karl N, et al. Human coronavirus 229E papain-like proteases have overlapping specificities but distinct functions in viral replication. J Virol 2007; 81: 3922-3932. 2007/01/26. [Crossref]

50. Sola I, Almazan F, Zuniga S, et al. Continuous and Discontinuous RNA Synthesis in Coronaviruses. Annu Rev Virol 2015; 2: 265-288. 2016/03/10. [Crossref]

51. Peng Q, Peng R, Yuan B, Zhao J, Wang M, Wang X, et al. Structural and Biochemical Characterization of the nsp12-nsp7-nsp8 Core Polymerase Complex from SARS-CoV-2. Cell Rep 2020;31:107774. [Crossref]

52. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science 2020;368:779–82. [Crossref]

53. Sutton G, Fry E, Carter L, Sainsbury S, Walter T, Nettleship J, et al. The nsp9 replicase protein of SARS-coronavirus, structure and functional insights. Structure 2004;12:341–53. [Crossref]

54. Expasy. SARS-CoV-2, COVID-19 Coronavirus Resource, https:// viralzone.expasy.org/9056 (accessed June 2020).

55. U.S. Department of Health and Human Services. AIDSinfo, https://aidsinfo.nih.gov/ (accessed June 2020).

56. Hutter G, Nowak D, Mossner M, et al. Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. N Engl J Med 2009; 360: 692-698. 2009/02/14. [Crossref]

57. Burke DS. Recombination in HIV: an important viral evolutionary strategy. Emerg Infect Dis 1997;3:253–9. [Crossref]

58. Craigie R. The molecular biology of HIV integrase. Future Virol 2012;7:679–86. [Crossref]

59. Cohen MS, Smith MK, Muessig KE, Hallett TB, Powers KA, Kashuba AD. Antiretroviral treatment of HIV-1 prevents transmission of HIV-1: where do we go from here? Lancet 2013;382:1515–24. [Crossref]

60. Menendez-Arias L. Mechanisms of resistance to nucleoside analogue inhibitors of HIV-1 reverse transcriptase. Virus Res 2008; 134: 124-146. 2008/02/15. [Crossref]

61. Shafer RW, Vuitton DA. Highly active antiretroviral therapy (HAART) for the treatment of infection with human immunodeficiency virus type 1. Biomed Pharmacother 1999;53:73–86. [Crossref]

62. Zhirnov OP. Molecular Targets in the Chemotherapy of Coronavirus Infection. Biochemistry (Moscow) 2020;85:523–30. [Crossref]

63. Eastman RT, Roth JS, Brimacombe KR, Simeonov A, Shen M, Patnaik S, Hall MD. Remdesivir: A Review of Its Discovery and Development Leading to Emergency Use Authorization for Treatment of COVID-19. ACS Cent Sci 2020;6:672–83. [Crossref]

64. Koonin EV, Dolja VV, Morris TJ. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit Rev Biochem Mol Biol 1993;28:375–430. [Crossref]

65. Adedeji AO, Singh K, Calcaterra NE, DeDiego ML, Enjuanes L, Weiss S, Sarafianos SG. Severe acute respiratory syndrome coronavirus replication inhibitor that interferes with the nucleic acid unwinding of the viral helicase. Antimicrob Agents Chemother 2012;56:4718–28. [Crossref]

66. Cao D, Gao Y, Roesler C, Rice S, D'Cunha P, Zhuang L, et al. Cryo-EM structure of the respiratory syncytial virus RNA polymerase. Nat Commun 2020;11:368. [Crossref]

67. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020;367:1260–3. [Crossref]

68. Pinto D, Park YJ, Beltramello M, Walls AC, Tortorici MA, Bianchi S, et al. Structural and functional analysis of a potent sarbecovirus neutralizing antibody. bioRxiv 2020. [Crossref]

69. Henderson R, Edwards RJ, Mansouri K, Janowska K, Stalls V, Gobeil SMC, et al. Controlling the SARS-CoV-2 Spike Glycoprotein Conformation. Nat Struct Mol Biol 2020. [Crossref]

70. Barnes CO, West AP, Huey-Tubman KE, Robbiani DF, Nussenzweig MC, Bjorkman PJ. Structures of human antibodies bound to SARS-CoV-2 spike reveal common epitopes and recurrent features of antibodies. Cell 2020;182:1–15. [Crossref]
71. Surya W, Li Y, Torres J. Structural model of the SARS coronavirus E channel in LMPG micelles. Biochim Biophys Acta Biomembr 2018;1860:1309–17. [Crossref]
72. Kang S, Yang M, Hong Z, Zhang L, Huang Z, Chen X, et al. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. Acta Pharm Sin B 2020. [Crossref]
73. Ye Q, West AMV, Silletti S, Corbett KD. Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. Protein Sci 2020. [Crossref]
74. Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm Sin B, 2020;10:766–88. [Crossref]
75. Maes B, Bosteels C, De Leeuw E, Declercq J, Van Damme K, Delporte A, et al. Treatment of severely ill COVID-19 patients with anti-interleukin drugs (COV-AID): A structured summary of a study protocol for a randomised controlled trial. Trials 2020;21:468. [Crossref]
76. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe 2020;27:671–80.e2. [Crossref]
77. Ayanoğlu İC, İpekoglu EM, Yazar V, Yılmaz İC, Gürsel İ, Gürsel M. Could Individuals From Countries Using BCG Vaccination Be Resistant to SARS-COV-2 Induced Infections? Turk J Immunol 2020;8:29–36. [Crossref]
78. Tang T, Bidon M, Jaimes JA, et al. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. Antiviral Res 2020; 178: 104792. 2020/04/10. [Crossref]
79. Harcourt BH, Jukneliene D, Kanjanahaluethai A, Bechill J, Severson KM, Smith CM, et al. Identification of severe acute respiratory syndrome coronavirus replicase products and characterization of papain-like protease activity. J Virol 2004;78:13600–12. [Crossref]
80. Baez-Santos YM, St John SE, Mesecar AD. The SARS-coronavirus papain-like protease: structure, function and inhibition by designed antiviral compounds. Antiviral Res 2015;115:21–38. [Crossref]
81. Brian DA, Baric RS. Coronavirus genome structure and replication. Curr Top Microbiol Immunol 2005;287:1–30. [Crossref]
82. Tang XC, Zhang JX, Zhang SY, et al. Prevalence and genetic diversity of coronaviruses in bats from China. J Virol 2006; 80: 7481-7490. 2006/07/15. [Crossref]
83. Farabaugh PJ. Programmed translational frameshifting. Microbiol Rev 1996;60:103–34. [Crossref]
84. Baranov PV, Henderson CM, Anderson CB, Gesteland RF, Atkins JF, Howard MT. Programmed ribosomal frameshifting in decoding the SARS-CoV genome. Virology 2005;332:498–510. [Crossref]
85. Buchan JR, Stansfield I. Halting a cellular production line: responses to ribosomal pausing during translation. Biol Cell 2007;99:475–87. [Crossref]
86. Lokugamage KG, Narayanan K, Huang C, Makino S. Severe acute respiratory syndrome coronavirus protein nsp1 is a novel eukaryotic translation inhibitor that represses multiple steps of translation initiation. J Virol 2012;86:13598–608. [Crossref]
87. Wang Q, Wu J, Wang H, Gao Y, Liu Q, Mu A, et al. Structural Basis for RNA Replication by the SARS-CoV-2 Polymerase. Cell 2020;182:417–28.e13. [Crossref]
88. Irigoyen N, Firth AE, Jones JD, Chung BY, Siddell SG, Brierley I. High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling. PLoS Pathog 2016;12:e1005473. [Crossref]
89. Kim D, Lee JY, Yang JS, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell 2020;181:914–21.e10. [Crossref]
90. Lodish HF. Insertion of Membrane Proteins into the ER. In: Lodish H, Berk A, Kaiser CA, Krieger M, Bretscher A, Ploegh H. Molecular Cell Biology, 8th ed. New York: Freeman-Macmillan Learning; 2016.
91. Mitaku Group. Classification and Secondary Structure Prediction of Membrane Proteins, http://harrier.nagahama-i-bio.ac.jp/sosui/ (accessed July 2020).
92. Hanel K, Stangler T, Stoldt M, Willbold D. Solution structure of the X4 protein coded by the SARS related coronavirus reveals an immunoglobulin like fold and suggests a binding activity to integrin I domains. J Biomed Sci 2006;13:281–93. [Crossref]
93. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 2003; 31: 3406-3415. 2003/06/26. [Crossref]
94. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567–80. [Crossref]
95. Angeletti S, Benvenuto D, Bianchi M, Giovanetti M, Pascarella S, Ciccozzi M. COVID-2019: The role of the nsp2 and nsp3 in its pathogenesis. J Med Virol, 2020;92:584–8. [Crossref]
96. Benvenuto D, Angeletti S, Giovanetti M, Bianchi M, Pascarella S, Cauda R, et al. Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy. J Infect 2020;81:e24–7. [Crossref]
97. Hagemeijer MC, Monastyrska I, Griffith J, van der Sluijs P, Voortman J, van Bergen en Henegouwen PM, et al. Membrane rearrangements mediated by coronavirus nonstructural proteins 3 and 4. Virology 2014;458-459:125–35. [Crossref]
98. Vogt AB, Stern LJ, Amshoff C, Dobberstein B, Hämmerling GJ, Kropshofer H. Interference of distinct invariant chain regions with superantigen contact area and antigenic peptide binding groove of HLA-DR. J Immunol 1995;155:4757–65. https://www.jimmunol.org/content/155/10/4757
99. Blanchard E, Roingeard P. Virus-induced double-membrane vesicles. Cell Microbiol 2015;17:45–50. [Crossref]
100. Fortin JS, Genève L, Gauthier C, Shoukry NH, Azar GA, Younes S, et al. MMTV superantigens coerce an unconventional topology between the TCR and MHC class II. J Immunol 2014;192:1896–906. [Crossref]
101. Aswad A, Katzourakis A. Convergent capture of retroviral superantigens by mammalian herpesviruses. Nat Commun 2015;6:8299. [Crossref]

102. Cagliani R, Forni D, Clerici M, Sironi M. Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses. Infect Genet Evol 2020;83:104353. [Crossref]

103. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res 2018;46:W200–4. [Crossref]

104. Adebali O, Ortega DR, Zhulin IB. CDvist: a webserver for identification and visualization of conserved domains in protein sequences. Bioinformatics 2015;31:1475–7. [Crossref]

105. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33:W244–8. [Crossref]

106. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 1998;14:378–9. [Crossref]

107. Davidson AD, Williamson MK, Lewis S, Shoemark D, Carroll MW, Heesom KJ, et al. Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site. Genome Med 2020;12. [Crossref]

108. Taiaroa G, Rawlinson D, Featherstone L, Pitt M, Caly L, Druce J, et al. Direct RNA sequencing and early evolution of SARS-CoV-2. bioRxiv 2020. [Crossref]

109. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Yahalom-Ronen Y, Tamir H, et al. The coding capacity of SARS-CoV-2. bioRxiv 2020. [Crossref]

110. Michel CJ, Mayer C, Poch O, Thompson JD. Characterization of accessory genes in coronavirus genomes. bioRxiv 2020. [Crossref]

111. Jungreis I, Sealfon R, Kellis M. Sarbecovirus comparative genomics elucidates gene content of SARS-CoV-2 and functional impact of COVID-19 pandemic mutations. bioRxiv 2020. [Crossref]

112. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, et al. Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. Cell 2020;181:1489–501.e15. [Crossref]

113. Hachim A, Kavian N, Cohen CA, Chin AWH, Chu DKW, Mok CKP, et al. Beyond the Spike: identification of viral targets of the antibody response to SARS-CoV-2 in COVID-19 patients. medRxiv 2020. [Crossref]

114. Wang H, Hou X, Wu X, Liang T, Zhang X, Wang D, et al. SARS-CoV-2 proteome microarray for mapping COVID-19 antibody interactions at amino acid resolution. bioRxiv 2020. [Crossref]

115. Hansen TH, Bouvier M. MHC class I antigen presentation: learning from viral evasion strategies. Nat Rev Immunol 2009;9:503–13. [Crossref]

116. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 2020;583:459–68. [Crossref]

117. Anthony SJ, Johnson CK, Greig DJ, Kramer S, Che X, Wells H, et al. Global patterns in coronavirus diversity. Virus Evol 2017;3:vex012. [Crossref]

118. Banerjee A, Kulcsar K, Misra V, Frieman M, Mossman K. Bats and Coronaviruses. Viruses 2019;11:41. [Crossref]

119. Pandey A, Nikam AN, Shreya AB, Mutalik SP, Gopalan D, Kulkarni S, et al. Potential therapeutic targets for combating SARS-CoV-2: Drug repurposing, clinical trials and recent advancements. Life Sci 2020;256:117883. [Crossref]

120. Guy RK, DiPaola RS, Romanelli F, Dutch RE. Rapid repurposing of drugs for COVID-19. Science 2020;368:829–30. [Crossref]