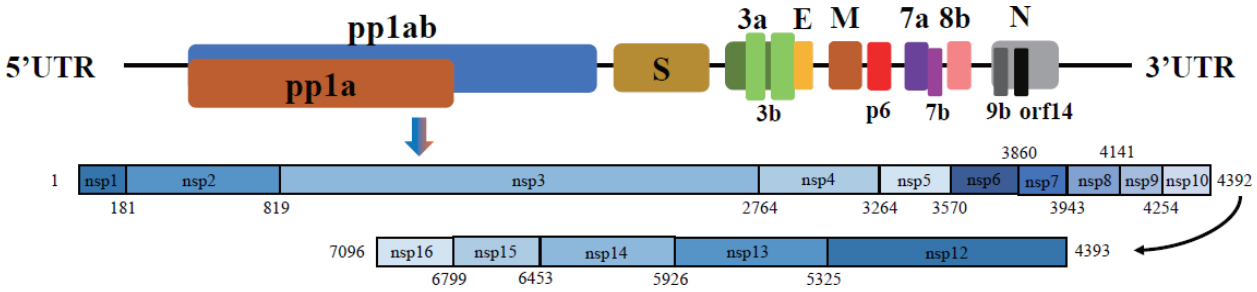


How is orf1ab cleaved into polypeptides? Can we predict this from the sequence?

➔ See R1AB_CVHSA function at https://www.uniprot.org/uniprot/P0C6X7

ORF1AB or "Open Reading Frame 1ab" contains the *rep or pp1ab* gene which codes for many different proteins in a single polypeptide chain *r1ab* or "Replicase Polyprotein 1ab" You can predict cleavage sites from the sequence except in the rare case of post-translational modifications (such as glycosylation) sterically hindering (physically blocking) the site. 3CL-PRO cleaves 11 sites on R1AB and recognizes the core sequence: [ILMVF]-Q-|-[SGACN].

How do the researchers know (guess?) where orf1ab cleaves?

This long polypeptide can be thought of as a payload package with several constituent components that are chained together. If we look at post-translational modifications there is some information such as, "enzymatic cleavages in vivo by its own proteinases" which means that the polypeptide is broken down inside the infected cell using viral machinery. 3CL-PRO and PL-PRO are autocatalytically processed, meaning they cleave themselves off of the polypeptide chain before subsequently cleaving the rest of the proteins in the payload. Here is a chart of the cleavage sites and proteinases:

Post-translational modification

Specific enzymatic cleavages in vivo by its own proteases yield mature proteins (PubMed:32083638). 3CL-PRO and PL-PRO proteinases are autocatalytically processed (By similarity). By similarity 1 Publication

Sites

Feature key	Position(s)	Description	Actions	Graphical view	Length
Site	180 – 181	Cleavage By similarity			2
Site	818 – 819	Cleavage; by PL-PRO By similarity			2
Site	2740 – 2741	Cleavage; by PL-PRO By similarity			2
Site	3240 – 3241	Cleavage; by 3CL-PRO By similarity			2
Site	3546 – 3547	Cleavage; by 3CL-PRO By similarity			2
Site	3836 – 3837	Cleavage; by 3CL-PRO 1 Publication			2
Site	3919 – 3920	Cleavage; by 3CL-PRO 1 Publication			2
Site	4117 – 4118	Cleavage; by 3CL-PRO 1 Publication			2
Site	4230 – 4231	Cleavage; by 3CL-PRO 1 Publication			2
Site	4369 – 4370	Cleavage; by 3CL-PRO By similarity			2
Site	5301 – 5302	Cleavage; by 3CL-PRO By similarity			2
Site	5902 – 5903	Cleavage; by 3CL-PRO By similarity			2
Site	6429 – 6430	Cleavage; by 3CL-PRO By similarity			2
Site	6775 – 6776	Cleavage; by 3CL-PRO By similarity			2

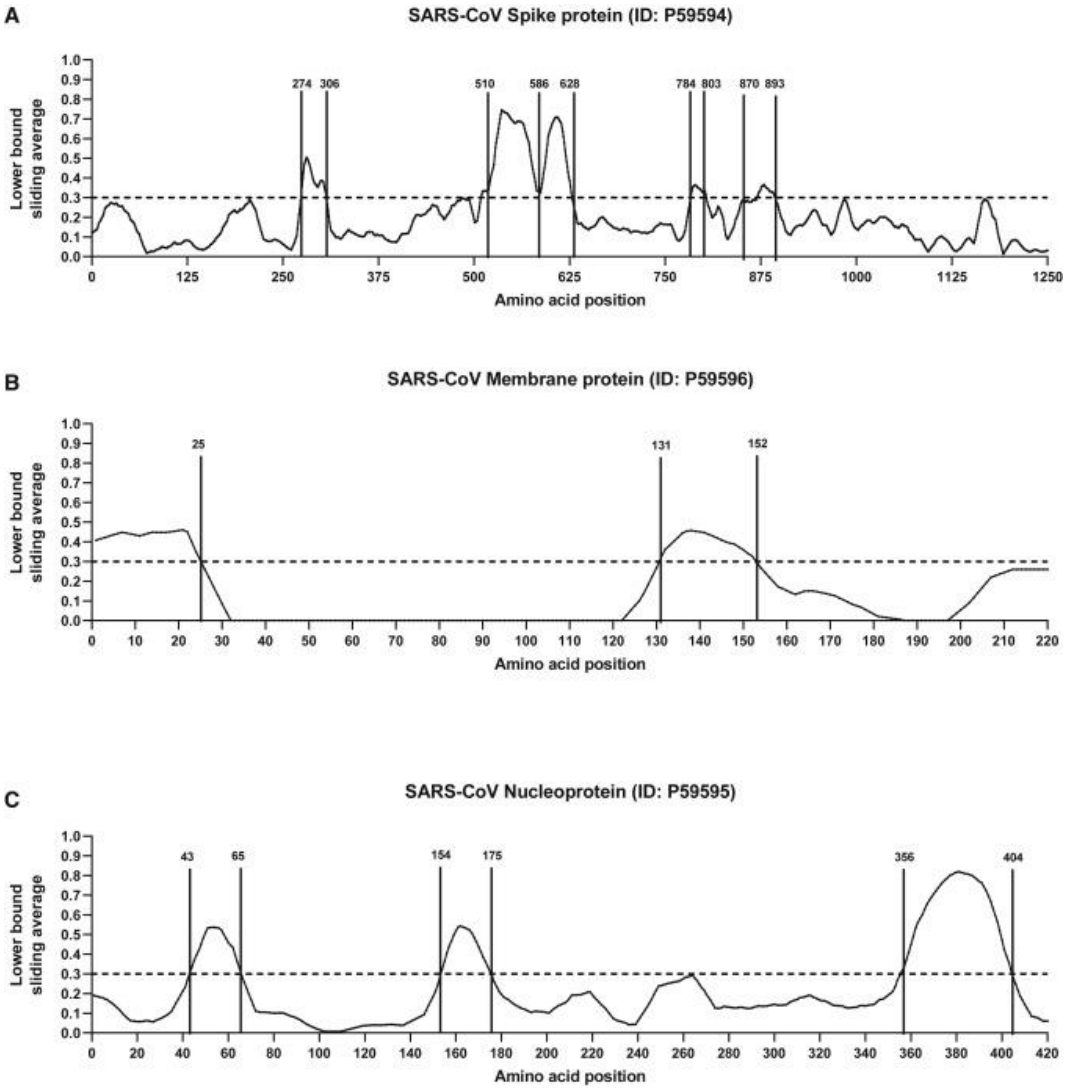
These sites are determined by a number of different methods such as homology (the active sites of these proteinases have a very similar structure to other proteinases for which we know the amino acid sequence that is recognized and cleaved), the cleavage sites can also be determined by analyzing the proteins that are extracted from an infected cell through various separation, mass spectrometric, and amino acid

sequence analysis schemes. For example, if we know the sequence of the CoV RNA we can find the START codon *Met* as well as the STOP codon and then match up the identified, known, and unknown proteins (from, for instance, a proteomics experiment ex: https://www.ebi.ac.uk/pride/archive/projects/PXD019119) from which we can infer that these various independent and functional proteins were initially derived from the same polypeptide chain. This phenomenon wherein inactive apoenzymes or proenzymes become activated upon cleavage by a peptidase is a common biochemical strategy in dynamically regulated cellular environments.

Which protein is the immune system responding to?

There are likely several antigenic determinants (epitopes / parts of a protein that antibodies can bind to) for regions of surface proteins (such as the capsid and spike proteins) that are both exposed to and accessible by the receptors on lymphocytes. This occurs due to the way that antibodies are produced, which involves phagocytosis (aka one cell cannibalizing another) of an infected cell, proteolytic degradation, and finally antigen presentation. A short polypeptide from the virus is presented on the surface of the cell and any antibodies possessing a hypervariable region that binds with very high affinity are chosen for mass production. Cross-reactivity between antibodies generated against one virus have been observed successfully challenging other viruses in the same family, likely due to homology overlap (ex: the same part of a protein is coded for in both viruses of that family so that antibody will bind to that region on both viruses). It has been hypothesized that this could be a potential reason for California having such a low infection rate but no conclusory evidence has been furnished as studies are ongoing.

Immune response maps can provide further insight into the most probable targets and bioinformatics approaches suggest the spike surface glycoprotein (S) and nucleocapsid protein (N). “Analyses of the spike glycoprotein, membrane protein, and nucleoproteins are shown in Figure 2. In the case of the spike glycoprotein (Figure 2A), we identify five regions of potential interest (residues 274–306, 510–586, 587–628, 784–803, and 870–893), all representing regions associated with high immune response rates.”



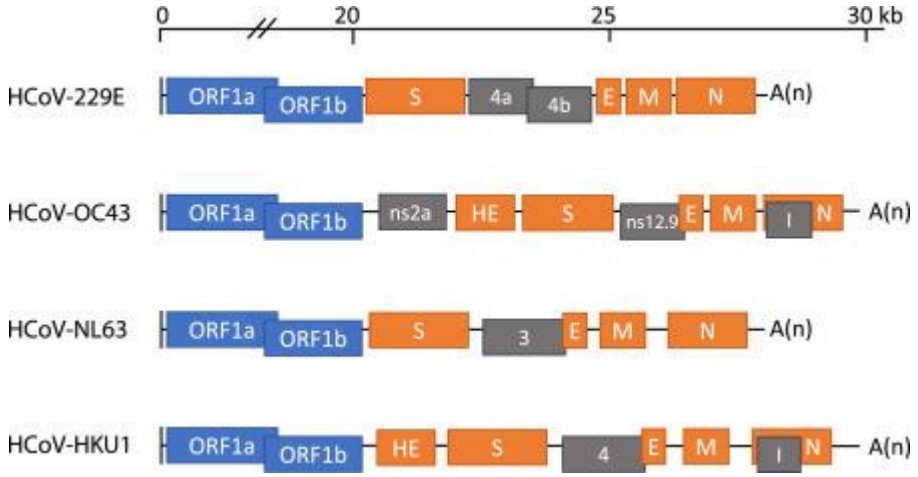
(*Source:* https://www.sciencedirect.com/science/article/pii/S1931312820301669)

Find the "furin cleavage site" in the "spike glycoprotein"

```
Alpha    HCoV-NL63   735 - GICADGSLI----PVRPRNSS - 751
Alpha    HCoV-229E   554 - GVCADGSII----AVQPRNVS - 570
Beta 2a  HCoV-OC43   753 - GYCVDYSK-----NRRSRGAI - 768
Beta 2a  HCoV-HKU1   742 - GFCVDYNSPSSSSSRRKRRSI - 762
Beta 2b  SARS-CoV    655 - GICASYHTVS-L----LRSTS - 670
Beta 2b  SARS-CoV-2  669 - GICASYQTQT-NSPRRARSVA - 688
Beta 2c  MERS-CoV    734 - SLCALPDTPSTLTPRSVRSVP - 754
                             . *.          *
```

How similar are the other coronaviruses?

There is high homology within and across the genomes of these pathogenic strains of coronaviridae, although a direct side-by-side comparison has not yet been published.



source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7204879/

It may be beneficial to have such a comparison and the best methodology with which to compare and contrast would most likely take the following form:

Take the amino acid chain for each of the 27 proteins encoded by SARS-CoV-2 (NC_045512.2**)** and align it against HCoV-229E (NC_002645.1), HCoV-NL63 (MG772808.1), HCoV-OC43 (NC_006213.1), HCoV-HKU1 (NC_006577.2) using the FFT-NS-2 algorithm in MAFFT v7 (https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html).

What adds the phosphate group to the N protein?

There are many different types of kinase enzymes that are responsible for phosphorylation and dephosphorylation of specific sites on specific proteins under a specific set of circumstances. The viral nucleocapsid phosphoprotein is synthesized in

the endoplasmic reticulum before traveling through the Golgi Apparatus, which means that it has full access to the host cell machinery for any number of complex post-translational modifications including phosphorylation and glycosylation. It appears as though the Serine/Threonine phosphorylation is carried out by casein kinase II (CK2) but experimental confirmation has not yet been published.

If you have any questions or comments you can contact me on github.io/bioanalytica or email me at bioanalytical@protonmail.com.