Functional pangenome analysis provides insights into the origin, function and pathways to therapy of SARS-CoV-2 coronavirus

Intikhab alam[1], Allan Kamau[1], Maxat Kulmanov[1], Stefan T. Arold[1,2], Arnab Pain[3,4], Takashi Gojobori[1] and Carlos M. Duarte[1*]

[1]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), [3]Biological and Environmental Science and Engineering (BESE), Thuwal, 23955-6900, Saudi Arabia
[2]Centre de Biochimie Structurale, CNRS, INSERM, Université de Montpellier, 34090 Montpellier, France
[3]Biological and Environmental Science and Engineering (BESE), Thuwal, 23955-6900, Saudi Arabia
[4]Research Center for Zoonosis Control, Hokkaido University, Sapporo, 001-0020 Japan

*corresponding authors: Intikhab.alam@kaust.edu.sa, Carlos.Duarte@kaust.edu.sa.

Abstract

Background. Recent epidemic of novel coronavirus (SARS-CoV-2) has triggered a rising global health emergency that demands urgent analysis of its genome and solutions for detection and therapy.

Methods. We used a comparative pangenomic analysis of *Betacoronavirus* sequenced thus far to detect the core and accessory gene clusters of this genus including SARS-CoV-2. We then annotate the functions, which are confirmed by structural analysis, and predict the potential location within the host cells of these proteins.

Findings. We found five accessory gene clusters common to the SARS clade, including SARS-CoV-2, that perform functions supporting their pathogenicity. Phylogenetic analysis showed one of the accessory gene clusters, the protein E, to be present across the inferred evolutionary pathway of the SARS clade, including that of the horseshoe bat virus (Hp-betacoronavirus/Zhejiang2013), inferred to be the parental member of the clade. The E protein is highly conserved in the clade, differing between SARS and SARS-Cov2 with a difference of single amino acid substitution and a single amino acid insertion present in SARS but absent from SARS-CoV-2. *Betacoronavirus* pangenomedb is available at http://pangenomedb.cbrc.kaust.edu.sa.

Interpretation. The characterization and functional assessment of SARS-CoV-2 envelope, E, protein in gene cluster 6, together with previous findings on this protein for SARS, leads us to recommend that detection of COVID-2019 be developed based on the SARS-CoV-2 E protein and that treatment using mutated SARS-Cov-2 lacking the E protein be explored as a promising candidate for a vaccine.

Funding. This research was funded by the King Abdullah University of Science and Technology (KAUST) through funding allocated to Computational Bioscience Research Center (CBRC) and another KAUST award under the award number FCC/1/1976-25-01.

Research in context

Evidence before this study

The response to the recent epidemic of novel coronavirus 2019-nCoV, now named SARS-CoV-2, and its disease named COVID-2019, require elucidating the possible origin, gene functions and potential treatment options. The genome sequence released for SARS-CoV-2 is a key resource to these efforts. We used all available whole-virus genomes released for the *Betacoronavirus* to conduct a functional pangenomic analysis of these viruses, including that of SARS-CoV-2. SARS-CoV-2 was previously reported in publications as 2019-nCoV. NCBI produced a dedicated resource for SARS-CoV-2 related publications and sequence data (https://www.ncbi.nlm.nih.gov/genbank/SARS-CoV-2-seqs/). We searched this resource for terms "2019-nCoV AND envelope" (1 hit), or "2019-nCoV AND orf10" (0 hits) or "2019-nCoV pangenome" (0 hits) to find any published work related to these genes in SARS-CoV-2. We retrieved all *Betacoronavirus* (taxon id 694002) whole genome sequences from NCBI (January 26, 2020 from web page https://www.ncbi.nlm.nih.gov/assembly) using the search term "txid694002[Organism:exp]". This search retrieved a total of 22 whole genomes, of which 18 were unique genomes, including 4 that for SARS-CoV-2, see supplementary table 1.

Added value of this study

We report and make available an interactive pangenome analysis resource (defining core and accessory genome components) for the genus *Betacoronavirus* (https://pangenomedb.cbrc.kaust.edu.sa), including SARS-CoV-2. We identify and define all core and accessory gene alignments and phylogenetic trees for this genus. We then explore these data to provide insights into potential functions for uncharacterized genes (ORFs 6,8,10) unique to SARS-like coronaviruses, and suggest potential options for detection and therapy of COVID-2019. This analysis points at protein E, highly conserved across all SARS-like coronaviruses, as a promising candidate for COVID-2019 detection and therapy.

Implications of all the available evidence

The *Betacoronavirus* pangenome data we made public allows exploration of core and accessory elements of these viruses to understand the pathogenicity of SARS-CoV-2, thereby expediting progress in responding to the pandemic. The characterization and functional assessment of SARS-CoV-2 envelope, E, protein, in gene cluster 6 together with previous findings on this protein for SARS, leads us to recommend that detection of COVID-2019 be developed based on the SARS-CoV-2 E protein and that treatment using mutated SARS-Cov-2 lacking the E protein be explored as a promising candidate for a vaccine.

.

Introduction

Recent epidemic of novel coronavirus (SARS-CoV-2) has triggered a rising global health emergency (WHO https://www.who.int/emergencies/diseases/novel-coronavirus-2019). Three major actions, confinement, detection and therapy, are required to contain this pandemic. Detection and therapy can be guided by virus sequence data, which was first made available in 17 December, 2019 at gsaid.org, soon after detection of the first cases. This SARS-CoV-2 genome sequence data is a pivotal resource on its own, but can gain greater value when embedded with those of other *Betacoronavirus*, allowing a comparative pangenomic analysis. This approach can help identify the core genome of *Betacoronaviruses* and extract accessory genomic features shared by a subset of these viruses or unique to SARS-CoV-2. Whereas core genomic features are required for the virus to be functional, accessory features are candidates to provide insights into the drivers of the unique capacities of SARS-CoV-2 explaining its spread and virulence. Genome annotation can then be used to assess the possible functions of these accessory features and guide approaches to detection and treatment.

Here we apply a comparative pangenomic approach of all *Betacoronavirus* genomes sequenced thus far to detect the core and accessory gene cluster of this genus, and then annotate the functions, further assessed through structural analysis and predict the potential location in the host cell of those accessory gene clusters present in SARS-CoV-2 that were uncharacterized to-date. The core gene clusters were then aligned to obtain SNPs allowing us to generate a phylogenetic tree of *Betacoronavirus*, providing insight into its likely evolutionary origin. Whole-genome sequence similarity analysis was then used to compare SARS-CoV-2 with phylogenetically close relatives from *Betacoronavirus* to further detect the regions of dissimilarity that may explain the distinct virulence of SARS-CoV-2, and to suggest targets for detection and therapy.

Results

The *Betacoronavirus* genomes available from NCBI (accessed 26 January, 2020) included 23 assemblies corresponding to 18 different strains, ranging in size from 29.9 to 31.5 Kb (Suppl. Table 1). About 10% of the gene clusters (4 out of 37) defined by panX [1] were present across all genomes, thereby defining the core genome for the genus *Betacoronavirus* (Fig. 1A). A species tree constructed based on SNPs in the genes related to core-genome gene clusters from *Betacoronaviruses* showed that the recently detected SARS-CoV-2 appears closest to the bat coronavirus, bat-SL-CoVZC45, followed by Severe acute respiratory syndrome (SARS) coronavirus (Fig. 1B).

From the accessory genome, we found 6 accessory gene clusters (18 % of the accessory genome of genus *Betacoronavirus*) common to SARS-like coronaviruses, including SARS-CoV-2, out of 33 total accessory gene clusters. Four of these gene clusters (ORF3, ORF6, ORF7, and ORF8, Fig. 1A), appear only in SARS-CoV-2, bat-SL-CoVZC45 and the SARS virus. A gene cluster, annotated as ORF10, appears to be unique to SARS-CoV-2 (Fig. 1). However, a blast-based search of this putative protein against the DNA of all sequences in NCBI's NT database, shows this sequence to match with 100% coverage in the DNA of other SARS-like genomes, but there is no open-reading frame predicted for those genomes in the matching (Fig 1A). This suggests that this

apparent uniqueness of ORF10 is a possible annotation artifact, for the SARS-CoV-2 genome or the annotation pipeline of GeneBank did not predict a gene for other matching genomes. An additional accessory gene cluster for envelope protein, E (cluster 6 with 7 strains) is found in the SARS, SARS-CoV-2, bat-SL-CoVZC45 as well as in Bat Hp-betacoronavirus/Zhejiang2013, and here Bat Hp-betacoronavirus/Zhejiang2013 appears to be a parental clade in the phylogenetic tree (Fig. 1B). There are two other and different envelope protein clusters (clusters 5 with 13 strains, cluster 22 with 2 strains) but none of these include SARS-like genomes that have been involved in human pandemics. SARS envelope protein (also present in our gene cluster 6) has been studied previously and shown to be involved in critical aspects of the viral life cycle (such as assembly, budding, envelope formation, and pathogenesis) and that CoVs lacking E make promising vaccine candidates[2].

Among the gene clusters unique to the SARS clade, including SARS-CoV-2, only cluster 6, related to E protein, is well annotated. Others gene clusters (ORF3, ORF6, ORF7, ORF8 and ORF10) appear to be uncharacterized according to existing GenBank annotations (https://www.ncbi.nlm.nih.gov/genome/proteins/86693?genome_assembly_id=760344). Re-annotation of proteins from all available *Betacoronavirus* genomes through KAUST metagenomic analysis platform (KMAP, www.cbrc.kaust.edu.sa/kmap, project 46) provided functional insights for five out of six accessory gene clusters present in SARS-CoV-2. These five clusters are now labelled with Protein Family (PFAM) domain IDs, as shown in Fig 1A. Two of these previously characterized gene clusters (ORF3a and ORF7a Fig 1A) appear to have a common function related to inducing apoptosis (cell death, see http://pfam.xfam.org/family/PF11289 and http://pfam.xfam.org/family/PF08779). Among other functions, ORF3a, also known as viroporin, modulates virus release and upregulates fibrinogen in host cells [3]. ORF7a seems to have immunoglobulin-like fold, which suggests ligand binding activity, possibly to integrin I domains [4]. The remaining three accessory gene clusters, ORF6, ORF8 and ORF10, present in SARS-CoV-2, SARS and bat-SL-CoVZC45 remain functionally uncharacterized. In the following section we provide results of function prediction using DeepGOPlus [5] and using several protein structural analyses.

Before using DeepGOPlus [5] to explore uncharacterized gene clusters further, we first tested its GO terms assignment for the well characterized envelope protein. As shown in Supplementary Fig 1, DeepGOPlus is able to assign functions such as interaction with host cell, virion assembly, budding, host cell disruption. Similar functions were described for SARS E protein previously[2] that differs from SARS-CoV-2 E protein for a single amino acid substitution and an insertion. DeepGOPlus also predicted localization of the E protein function in the host cell to be associated with the Golgi apparatus, vacuoles, endoplasmic reticulum, and membrane cell periphery of the host cell. DeepGOPlus predicts a role for ORF6 in the regulation of molecular function, cellular component biogenesis and organization with potential cellular localization in membrane-forming protein complexes (Suppl. Figure 2). GO for ORF8 predicts potential functions as response to stimulus along with biological regulation with potential cellular localization in vacuoles, cell periphery and extracellular regions (Suppl. Figure 3). For the smallest protein ORF10, DeepGOPlus predicts a role as regulator of molecular function and response to stimulus, similar to ORF8, but is predicted to be localized only in the extracellular region of the host cell (Suppl. Figure 4).

Our bioinformatic structural assessment showed that the structural architecture of protein E (a homopentameric single-pass transmembrane ion channel) is conserved compared to the 95% identical protein E from other SARS-CoV (Supplementary Figures 5,6). ORF3a displayed a multi-pass transmembrane domain and a cytoplasmic β-barrel or β-sandwich fold (Supplementary Figures 5,6). SARS-CoV-2 ORF7a also retains all functional features identified in other SARS ORF7a (~85% identical), including an N-terminal *sec*-pathway signal peptide, cleaved after residue 15, an immunoglobulin-like β-sandwich fold stabilized by two cysteine di-sulphide bonds, and a C-terminal single-pass transmembrane helix (Supplementary Figures 5,6). These analyses confirmed our and previous annotations of these proteins as accessory proteins mostly localized in the endoplasmic reticulum-Golgi intermediate compartment, but also occurring on the cell membrane with critical functions for enhancing viral pathogenicity and mortality[3,4,6].

The predicted functions of ORFs 6, 8 and 10 are also in agreement with their structural features. All are very short polypeptides (61, 121 and 38 residues for ORF6, 8 and 10, respectively) with a large percentage of hydrophobic residues (62%, 56% and 50%, respectively). None of the proteins have trans-membrane regions, but ORF8 has an N-terminal *sec*-pathway signal peptide with a cleavage site between residues 15 and 16, suggesting that it is secreted into the extracellular space (Supplementary Figures 5,6). Following signal peptide cleavage, the ORF8 protein core is predicted to consist largely of β-strands and features 7 cysteines. We predict that this protein adopts a cysteine disulphide-bond stabilized β-sandwich structure similar to the soluble domain of ORF7a (Supplementary Figures 5,6), inferring that ORF8 also functions as ligand binding module. ORF6 consists of a long amphipathic helical region, followed by an acidic tail, whereas ORF10 is predicted to harbour a long helix followed by a β-strand. Collectively, these structural characteristics confirm a regulatory role of ORF6, 8 and 10, most likely by binding to, and modulating, other proteins.

A whole-genome comparison at the nucleotide level among the four *Betacoronavirus* that are in closely related in the SARS clade (Fig. 2) shows the coronavirus from bat (bat-SL-CoVZC45) and a human coronavirus (SARS, AY274119) to be very similar (~88% and ~79% sequence identity, respectively) to Wuhan coronavirus (SARS-CoV-2). Genomes of SARS-CoV-2 and bat-SL-CoVZC45 differ mainly in the region where the spike protein, S, is encoded, suggesting the spike protein to be host specific. However, the SARS virus shows better similarity to SARS-CoV-2 than bat-SL-CoVZC45 in the spike protein region (Fig. 2), however the SARS virus shows more dissimilarities in the region of proteins ORF 1ab and ORF8.

Discussion

The comparative analysis of available *Betacoronavirus* genomes reported here helped us to identify the unique and shared features of SARS-CoV-2 virus and provided insights into its virulence while also providing guidance for their rapid detection and possible therapy. A comparative genomic analysis at a larger scale, including alpha, beta, gamma and delta coronaviruses, identified substitutions in specific amino acids among severe acute respiratory syndrome (SARS) or SARS-like coronaviruses [7], potentially leading to functional and pathogenic divergence in SARS-CoV-2. We complement this effort through a pan-genomic approach focused on functional similarities and annotation, rather than amino acid substitution. Our analysis

229 provides a unique, user-friendly and interactive resource (https://pangenomedb.cbrc.kaust.edu.sa)
230 of core and accessory gene clusters from the genus *Betacoronavirus* that sheds light on potential
231 function of uncharacterized genes unique to SARS-like coronaviruses. Another recent study [8]
232 described one of the SARS-CoV-2 isolate genome from Wuhan and discussed phylogenetic
233 analysis of its genes in context of comparison to *Betacoronavirus* genus, although protein E was
234 not described in detail. Furthermore, in our Markov Chain Clustering (MCL), SARS ORF7, ORF8
235 and ORF11 (gene symbols taken from GenBank annotation file for accession AY274119) are
236 clustering with ORF6, ORF7 and ORF8, respectively, for SARS-CoV-2. The conservation of
237 ORF11 from SARS with ORF8 of SARS-CoV-2 is however weak, all alignments are available at
238 our *Betacoronavirus* pangenome database.
239
240 Our analysis suggests no major function or gene cluster to be unique only to SARS-CoV-2, as
241 the unique gene cluster related to ORF10 identified is not robust and likely to be an annotation
242 artifact since TBLASTN of this protein finds matches with 100% coverage in all SARS-like
243 genomes. The spike protein present in SARS-CoV-2, a replicase that is an RNA-dependent RNA
244 polymerase (RdRp) and two proteases in coronavirus (namely proteinase (3CLpro) and the
245 papain-like protease (PLpro) have been proposed as therapeutic targets for possible treatments
246 for SARS-CoV-2 [9]. The similarity of the spike protein between SARS-CoV-2 and SARS shown
247 here implies that SARS antiviral drugs would also be effective in blocking the binding of SARS-
248 CoV-2 spike protein to human host cells [10]. However, a recent study notes that some of the most
249 potent SARS-CoV-specific neutralizing antibodies, targeting human receptor ACE2, failed to
250 bind SARS-CoV-2 spike protein (doi: https://doi.org/10.1101/2020.01.28.923011). These results
251 suggest that new antibodies that bind specifically to the SARS-CoV-2 Receptor Binding Domain
252 (RBD) need be developed. Our analysis provided potential functions for four previously
253 uncharacterized accessory gene clusters (ORF3, 6, 7 and 8), which provide valuable novel
254 insights into the function of the SARS clade, including SARS-CoV-2.
255
256 The assessment of gene clusters provided here suggests that one of the accessory gene clusters,
257 cluster 6 (Fig 1A), present in SARS-CoV-2, the E envelope protein, may be of potential interest
258 as a target for both SARS-CoV-2 detection and vaccine development. E-protein based detection
259 has already been used for the SARS-CoV-2 outbreak, applying RT-PCR for this envelope protein,
260 successfully reconfirming positive cases and allowing estimates of viral loads in carriers [11,12]. A
261 few studies have investigated the potential of vaccines based on either virus with mutated E protein
262 or lacking E protein altogether, specifically focusing on SARS- and MERS-CoV. This approach,
263 which resulted in robust immune responses in animal models for SARS- and MERS-CoV [2],
264 together with our results, show the E protein of SARS-CoV-2 to be critical for both detection and
265 therapy. Specifically, we suggest that a vaccine based on SARS-CoV-2 virus lacking the E protein
266 could be effective in inducing robust immune responses against this novel virus.
267
268 **Methods**
269
270 Collection of *Betacoronavirus* strains.
271 The NCBI assemblies page https://www.ncbi.nlm.nih.gov/assembly was searched on January 26,
272 2020 using the search term "txid694002[Organism:exp]" to find all complete viral genomes for
273 genus *Betacoronavirus*. As a result, 22 different available assemblies were downloaded in addition
274 to the assembly for bat virus bat-SL-CoVZC45, see Supplementary Table 1. Multiple strains with

the same taxon id are marked in yellow, these marked assemblies are from multiple isolate genomes including four isolates from the Wuhan coronavirus, SARS-CoV-2. These assemblies in GenBank format were updated for the locus_tag identifiers containing so that all genes can be traced with a locus id and the gene index number in order to comply with our pangenome analysis.

Creating a pangenome for *Betacoronavirus*

For pangenome analysis we used panX (Ding et al., 2018) that takes GenBank files as input and extracts gene coordinates, annotations and sequences from these genomes on both nucleotide and protein levels. These sequences are compared to each other and gene clustering is performed. Phylogenetic analysis is carried out on the core gene clusters to produce a species tree and similarly for each gene cluster to produce gene trees. These alignments and phylogenetic trees are then visualized using panX visualization module obtain from github, https://github.com/neherlab/pan-genome-visualization/. These gene clusters and their alignments at the nucleotide and amino acid levels as well as phylogenetic tree analysis at the species and gene levels, are available for further inspection (https://pangenomedb.cbrc.kaust.edu.sa). Other databases showing greater detail about many viruses, such as Virus Pathogen Resource (VIPR, https://www.viprbrc.org) and GSAID (https://www.gisaid.org) are very useful but they do not provide gene clustering or pan-genome analysis.

Re-annotation of proteins from *Betacoronavirus* genomes using KAUST Metagenomic Analysis Platform (KMAP) annotation pipeline.

All proteins clustered into groups using pan-genome analysis were re-annotated using KMAP annotation system that employs sequence-based BLAST to UniProtKB, KEGG and also Protein Family (PFam) domain detection using InterPro. Resulting annotations were indexed into KMAP database under project 46, http://www.cbrc.kaust.edu.sa/kmap/ , available after clicking on the public access button and the projects tab.

Prediction of function using DeepGOPlus.

We used all genes from *Betacoronavirus* dataset in order to explore Gene Ontology predictions through DeepGOPlus [5] tool that combines deep convolutional neural network (CNN) model with sequence similarity-based predictions to derive relevant functional classes from Gene Ontology alongside a confidence score. Results were included with a score 0.1 and filtered for class specificity. Predicted Gene Ontology (GO) terms with a confidence score from DeepGOPlus were visualized using QuickGO (https://www.ebi.ac.uk/QuickGO/) for SARS-CoV-2 specific gene clusters. DeepGOPlus is available on github, https://github.com/bio-ontology-research-group/DeepGOPlus

Structure based prediction of function

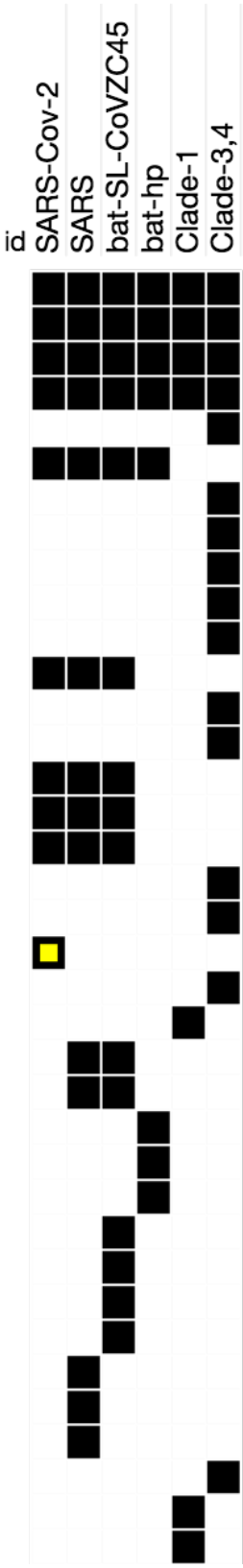
ProtParam (https://web.expasy.org/cgi-bin/protparam/protparam) was used for calculating protein features. Phobius (http://phobius.sbc.su.se/) and SignalP-5.0 were used for prediction of transmembrane regions and signal peptides [13]. Jpred4 was used to calculate secondary structure features (http://www.compbio.dundee.ac.uk/jpred4/). 3D modelling was carried out using SwissModel (homology modelling)[14] or QUARK (ab initio structure predictions)[15].

Author contributions

IA, TG and CMD conceived and designed the research, IA led the study and conducted the data analysis, A.K. helped developed the web-based resource and led computational components of the study, MK and SA contributed to the functional structural analysis, and AP contributed with inferences on virulence. IA and CMD developed the first draft of the manuscript and all authors contributed to writing and improving the manuscript, and approved the submission.

Acknowledgements

We are thankful to KAUST Supercomputing Laboratory (KSL), KAUST Information Technology (IT) and Muhammad Saif from CBRC, KAUST for providing support in computing resources.



#	Name	Description	Strains
1	M	membrane glycoprotein	23.00
2	N	nucleocapsid protein	23.00
3	orf1ab	orf1ab polyprotein	23.00
4	S	spike glycoprotein	23.00
5	E	envelope protein	13.00
6	E	envelope protein|PF02723	7.00
7	HE	hemagglutinin-esterase	7.00
8	ns12.9	ns12.9	6.00
9	NS2a3	32 kDa non-structural protein	6.00
10	NS3a	ORF 3a	6.00
11	NS3d	orf5	6.00
12	orf3a	orf3a protein |PF11289	6.00
13	orf4a	orf4a	6.00
14	orf4b	orf4b	6.00
15	ORF6	ORF6 protein|PF12133	6.00
16	orf7a	orf7a protein |PF08779	6.00
17	ORF8	ORF8 protein|PF12093	6.00
18	N2	internal protein	4.00
19	NS4	4.8 kDa non-structural protein	4.00
20	orf10	orf10 protein	3.00
21	orf8b	orf8b	3.00
22	E	small membrane protein	2.00
23	None	Orf13	2.00
24	None	Orf14	2.00
25	NS3	NS3 protein	2.00
26	NS7a	putative NS7a protein	2.00
27	NS7b	putative NS7c protein	2.00
28	None	hypothetical protein	1.00
29	None	hypothetical protein	1.00
30	None	hypothetical protein	1.00
31	None	hypothetical protein	1.00
32	None	Orf9	1.00
33	None	Orf10	1.00
34	None	Orf4	1.00
35	None	4.9 kDa non-structural protein	1.00
36	NS7b	putative NS7b protein	1.00
37	p10	putative p10 protein	1.00

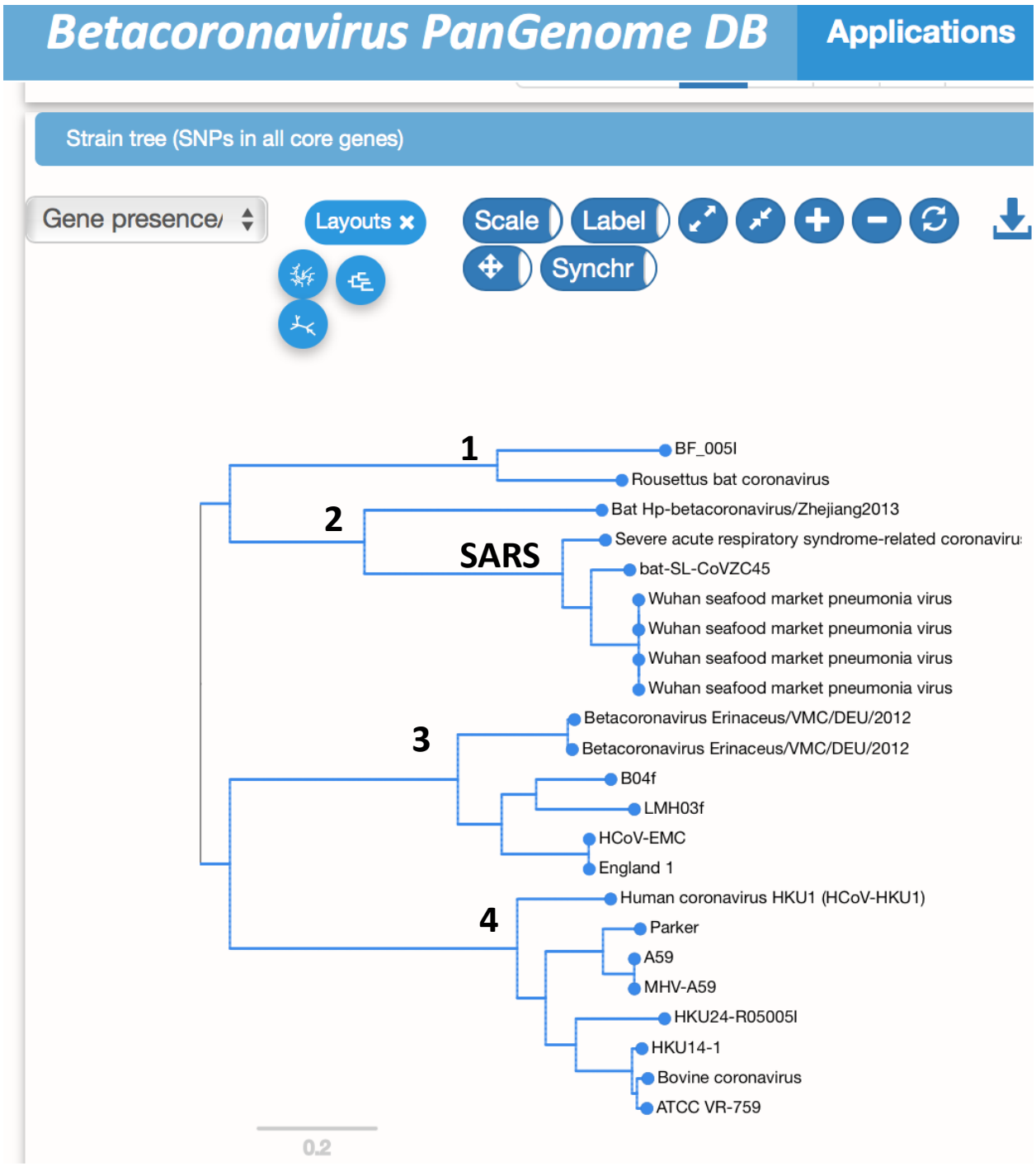
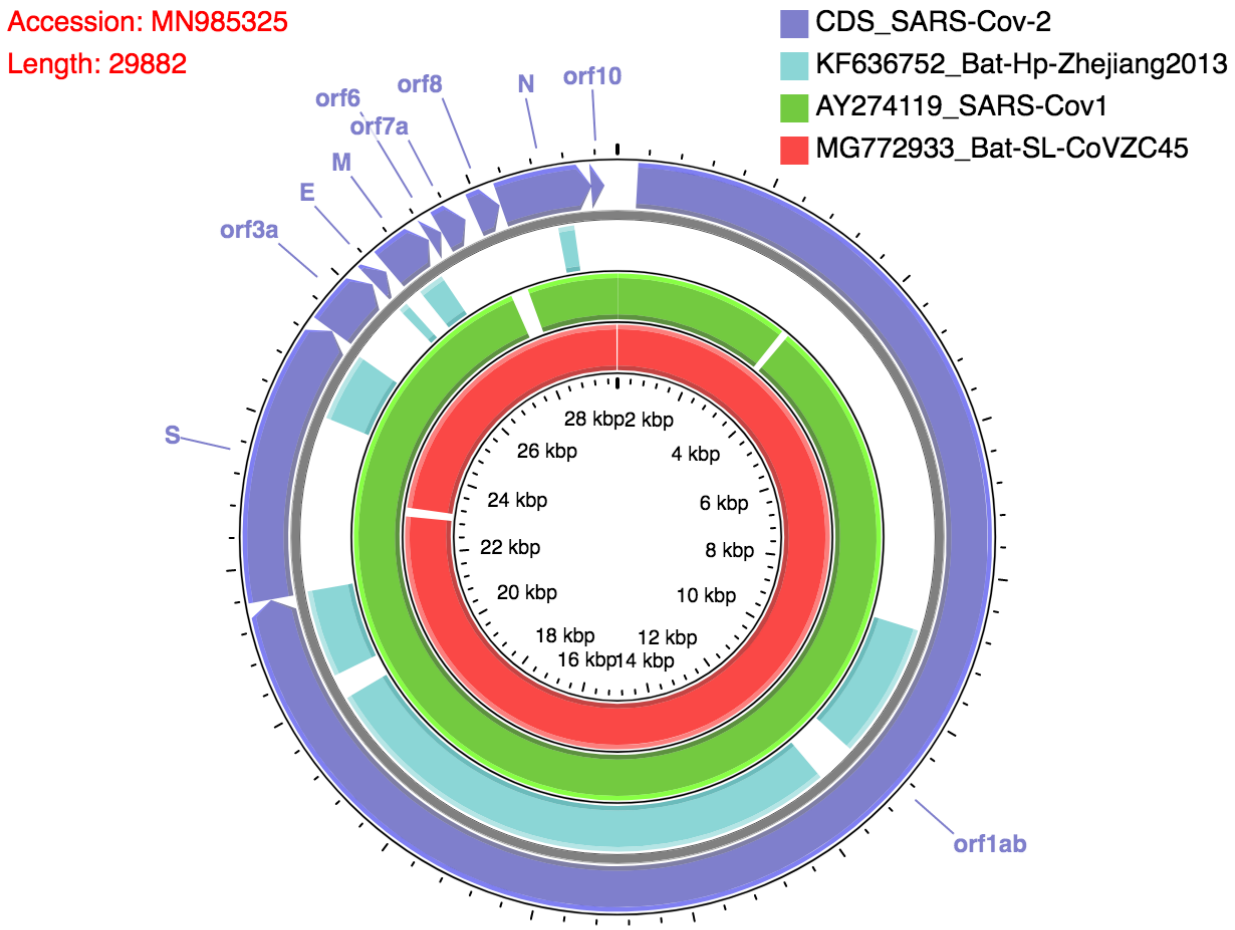
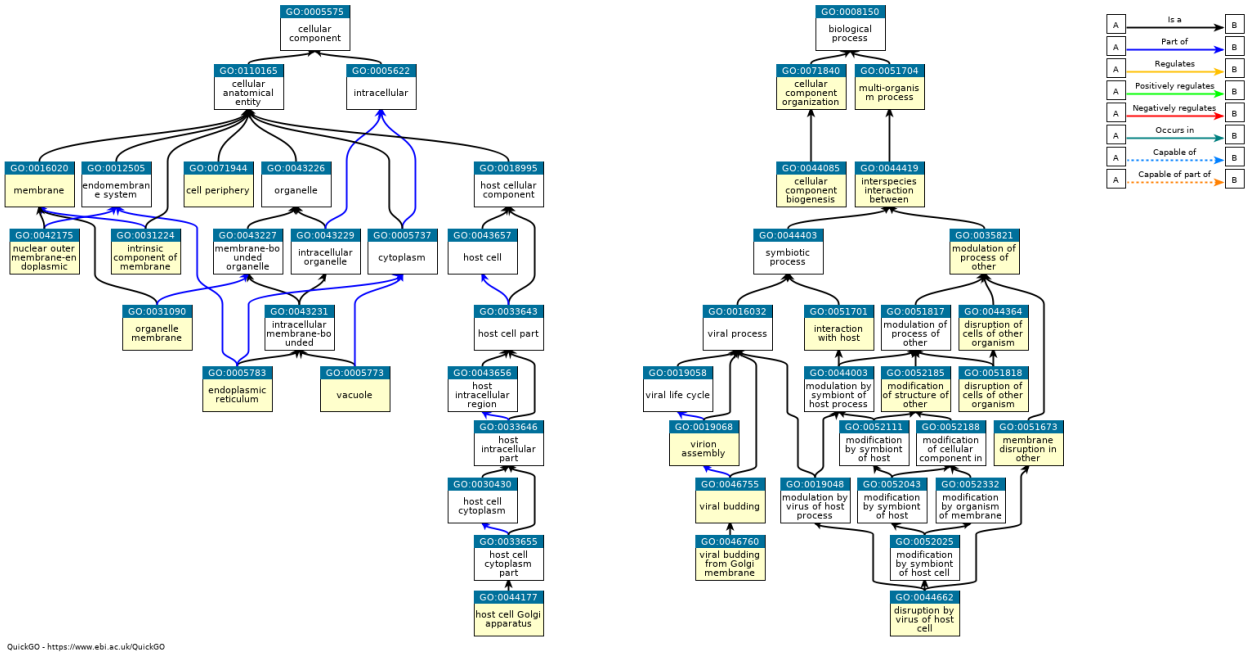


Figure 1. A. Breakdown of 37 gene clusters showing common and unique genes across four coronavirus clades, here clade-2 represents SARS, SARS-CoV-2 (COVID-2019) isolates and coronaviruses from two different bats. One of the gene cluster, orf10, marked in yellow, appears to be unique in SARS-CoV-2 according to annotations from GenBank, however a TBLASTN search of this protein against NCBI's Nucleotide database (NT) shows presence of this gene in other SARS-like coronaviruses. B. A species tree is shown based on SNPs from all core genes alignments, marked with clade 1-4 and SARS-like subclade.

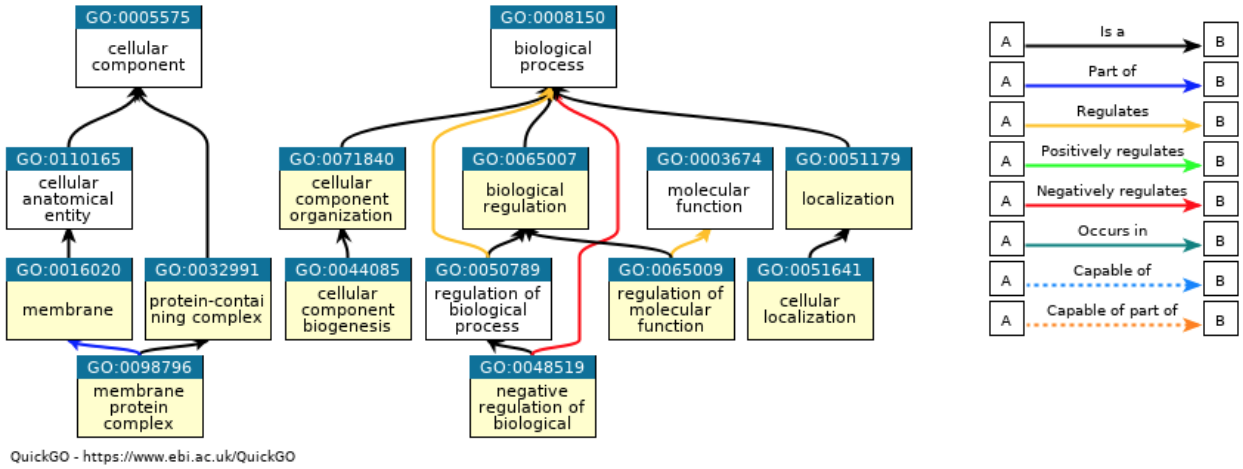


SARS-Cov-2 Whole Genome Comparison

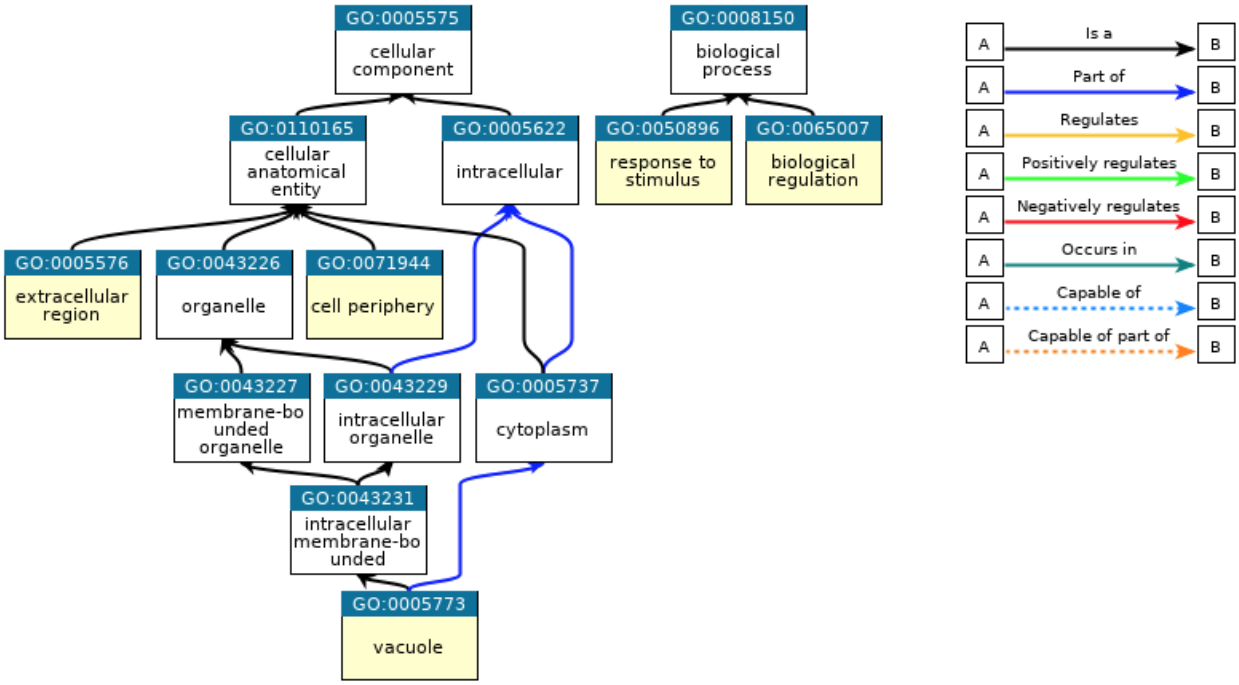
Figure 2. BLASTN based Whole genome comparison among SARS-CoV-2 (CDS shown in blue) and similar SARS-Cov (green) and two Bat SARS-like genomes (Bat-SL-CovZC45 in red and Bat-Hp-Zhejiang2013 in light blue). Regions with no match to SARS-CoV-2 are shown in white. This BLAST atlas was produced using CGview (http://cgview.ca)



Supplementary Figure 1.
Gene Ontology Graph for predicted functions for SARS-CoV-2 envelope protein, E.

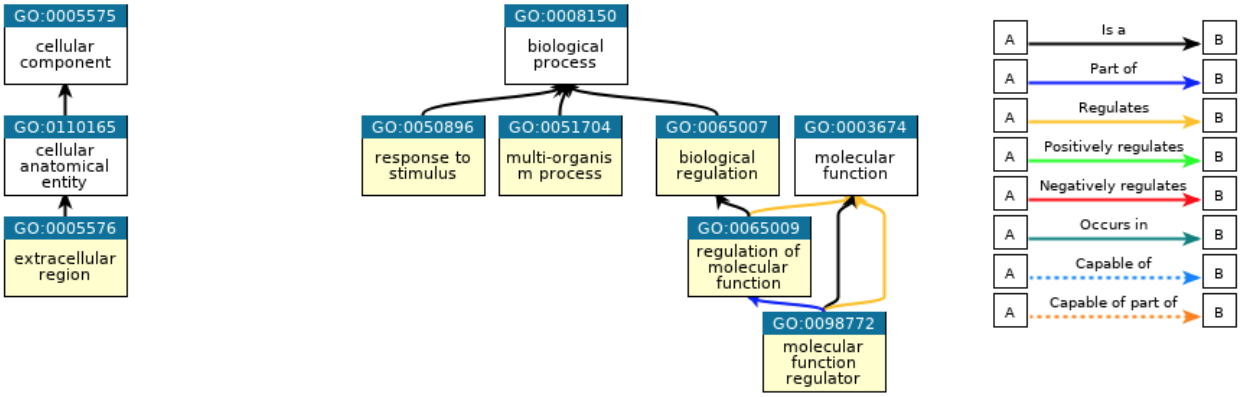


Supplementary Figure 2.
Gene Ontology Graph for SARS-CoV-2 ORF6.



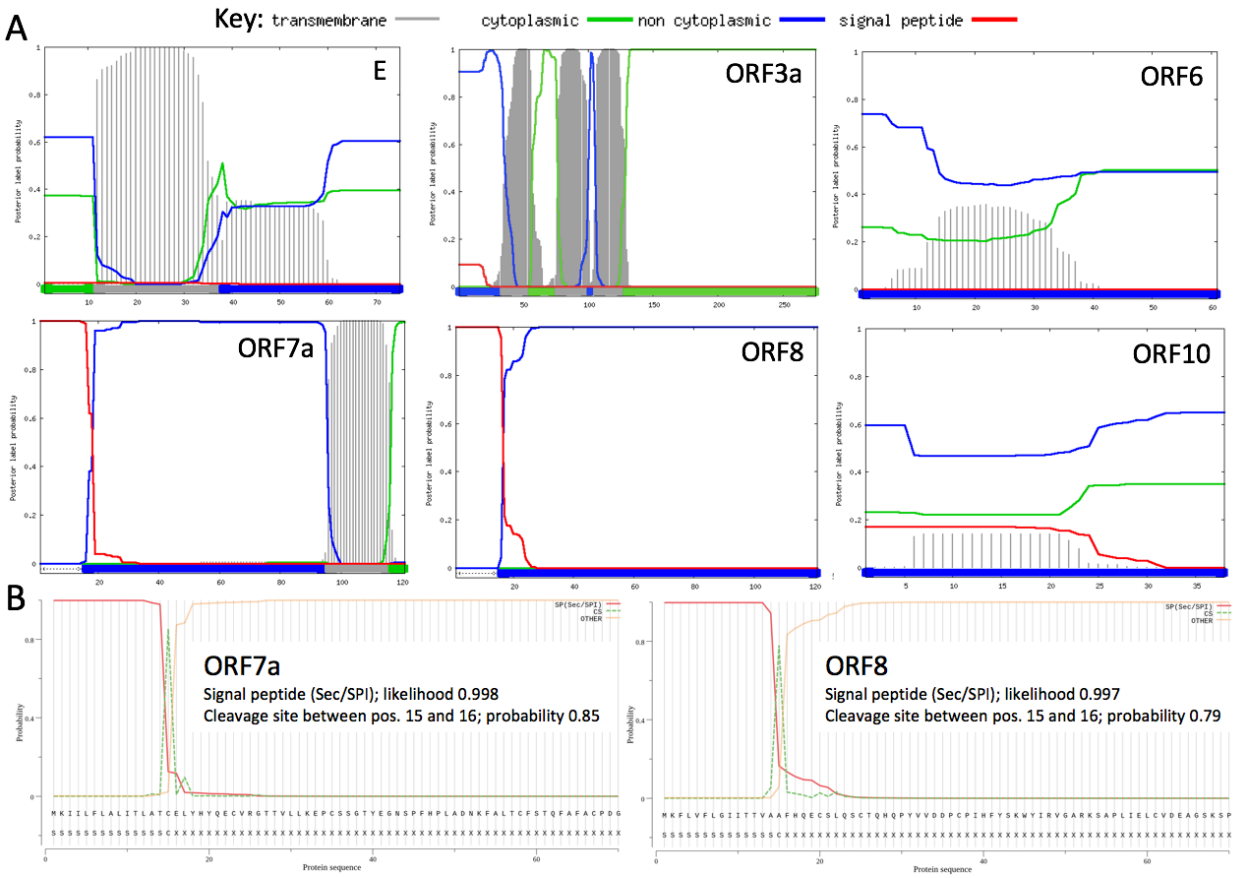
QuickGO - https://www.ebi.ac.uk/QuickGO

Supplementary Figure 3.
Gene Ontology Graph for SARS-CoV-2 ORF8



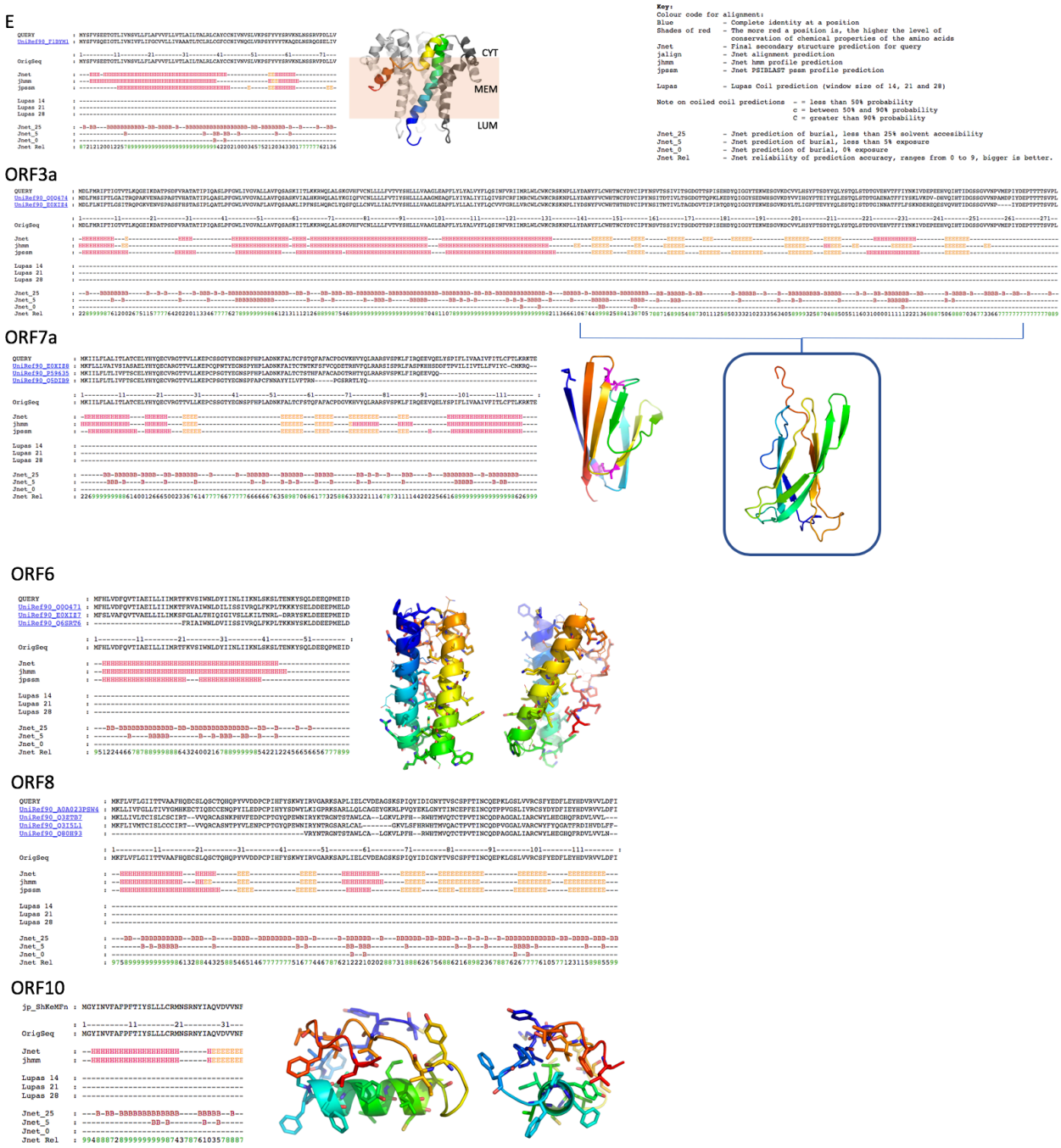
QuickGO - https://www.ebi.ac.uk/QuickGO

Supplementary Figure 4.
Gene Ontology Graph for SARS-CoV-2 gene, ORF10



Supplementary Figure 5

Prediction of signal peptides and transmembrane regions. A) combined prediction of signal peptides and transmembrane regions, and (B) an additional analysis to identify the type of signal peptide and cleavage site for ORF7a and ORF10.



Supplementary Figure 6

Prediction of secondary structure characteristics and tertiary architecture [colour-ramped from N-(blue) to C-terminus (red)]. E: homology model based on the pentameric SARS CoV E protein (PDB Id 5x29, 26% sequence identity). Only one chain is colour-ramped. The membrane (pale pink) is indicated with cytoplasmic and luminal sides labelled. Note that the N-terminal 7 and C-terminal 15 residues are not included in the model. ORF3a and ORF7a: structural models for the soluble domains. The two di-sulphide bridges conserved in SARS ORF3a are shown in magenta (template: 1xak). 90˚ views of *ab initio* models are shown for ORF6 and ORF10. Hydrophobic

398 and charged side chains are highlighted in ORF6, and hydrophobic side chains are highlighted in
399 ORF10.
400

Supplementary Table 1. Details of GenBank *Betacoronavirus* Genomes. Genomes with same taxonid are highlighted in yellow.

	Acession	Taxon	Name	Strain	Sample Location	Host
1	AF029248	11138	Murine hepatitis virus	A59	.	Murine hepatitis virus
1	AY700211	11138	Murine hepatitis virus	MHV-A59	.	Murine hepatitis virus
2	AF391541	11128	Bovine coronavirus	Bovine coronavirus	.	Bovine
3	AY274119	694009	Severe acute respiratory syndrome-related coronavirus	Severe acute respiratory syndrome-related coronavirus	Canada: Toronto	Homo sapiens
4	AY585228	31631	Human coronavirus OC43 (HCoV-OC43)	ATCC VR-759	USA	Home Sapiens
5	AY597011	290028	Human coronavirus HKU1 (HCoV-HKU1)	Human coronavirus HKU1 (HCoV-HKU1)	.	Homo sapiens
6	EF065505	424359	Bat coronavirus HKU4-1	B04f	China: Guangdong province	Bat
7	EF065509	424363	Bat coronavirus HKU5-1	LMH03f	China: Guangdong province	Bat
8	EF065513	424367	Bat coronavirus HKU9-1	BF_005I	China: Guangdong province	Bat
9	FJ938068	502102	Rat coronavirus Parker	Parker	.	Rat
10	JN874559	1160968	Rabbit coronavirus HKU14	HKU14-1	China	Oryctolagus cuniculus
11	JX869059	1235996	Human betacoronavirus 2c EMC/2012	HCoV-EMC	.	Homo sapiens
12	KC164505	1263720	Betacoronavirus England 1	England 1	United Kingdom	Homo sapiens
13	KC545383	1385427	Betacoronavirus Erinaceus/VMC/DEU/2012	Betacoronavirus Erinaceus/VMC/DEU/2012	Germany	Erinaceus europaeus
13	KC545386	1385427	Betacoronavirus Erinaceus/VMC/DEU/2012	Betacoronavirus Erinaceus/VMC/DEU/2012	Germany	Erinaceus europaeus
14	KF636752	1541205	Bat Hp-betacoronavirus/Zhejiang2013	Bat Hp-betacoronavirus/Zhejiang2013	China	Hipposideros pratti
15	KM349742	1590370	Betacoronavirus HKU24	HKU24-R05005I	China	Rattus norvegicus (Norway rat)
16	KU762338	1892416	Rousettus bat coronavirus	Rousettus bat coronavirus	China	Rousettus leschenaultia
17	MN938384	2697049	Wuhan seafood market pneumonia virus	Wuhan seafood market pneumonia virus	China: Shenzhen	Homo sapiens
17	MN975262	2697049	Wuhan seafood market pneumonia virus	Wuhan seafood market pneumonia virus	China	Homo sapiens
17	MN985325	2697049	Wuhan seafood market pneumonia virus	Wuhan seafood market pneumonia virus	USA	Homo sapiens
17	MN988713	2697049	Wuhan seafood market pneumonia virus	Wuhan seafood market pneumonia virus	USA: Illinois	Homo sapiens
18	MG772933	1508227	Bat SARS-like coronavirus	bat-SL-CoVZC45	China	Rhinolophus sinicus (Chinese rufous horseshoe bat)

References

1. Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. *Nucleic Acids Res* 2018; **46**(1): e5.
2. Schoeman D, Fielding BC. Coronavirus envelope protein: current knowledge. *Virol J* 2019; **16**(1): 69.
3. Minakshi R, Padhan K, Rani M, Khan N, Ahmad F, Jameel S. The SARS Coronavirus 3a protein causes endoplasmic reticulum stress and induces ligand-independent downregulation of the type 1 interferon receptor. *PLoS One* 2009; **4**(12): e8342.
4. 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**(3): 281-93.
5. Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* 2020; **36**(2): 422-9.
6. DeDiego ML, Nieto-Torres JL, Regla-Nava JA, et al. Inhibition of NF-kappaB-mediated inflammation in severe acute respiratory syndrome coronavirus-infected mice increases survival. *J Virol* 2014; **88**(2): 913-24.
7. Wu A, Peng Y, Huang B, et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* 2020.
8. Chan JF, Kok KH, Zhu Z, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microbes Infect* 2020; **9**(1): 221-36.
9. Liu W, Morse JS, Lalonde T, Xu S. Learning from the Past: Possible Urgent Prevention and Treatment Options for Severe Acute Respiratory Infections Caused by 2019-nCoV. *Chembiochem* 2020.
10. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J Virol* 2020.
11. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020.
12. Corman VM, Landt O, Kaiser M, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. *Euro Surveill* 2020; **25**(3).
13. Almagro Armenteros JJ, Tsirigos KD, Sonderby CK, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* 2019; **37**(4): 420-3.
14. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018; **46**(W1): W296-W303.
15. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* 2012; **80**(7): 1715-35.