An integrated *in silico* immuno-genetic analytical platform provides insights into COVID-19 serological and vaccine targets

Daniel Ward [1,*], Matthew Higgins[1], Jody E. Phelan[1], Martin L. Hibberd[1], Susana Campino[1], Taane G Clark[1,2]

[1] Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, London, United Kingdom

[2] Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom

* corresponding authors

Daniel Ward and Prof. Taane Clark

Department of Infection Biology

Faculty of Infectious and Tropical Diseases

London School of Hygiene and Tropical Medicine

Keppel Street, London WC1E 7HT

Daniel.ward1@lshtm.ac.uk, Taane.clark@lshtm.ac.uk

ABSTRACT

Background: The COVID-19 pandemic is causing a major global health and socio-economic burden, instigating the mobilisation of resources into the development of control tools, such as diagnostics and vaccines. The poor performance of some diagnostic serological tools has emphasised the need for up to date immune-informatic analyses to inform the selection of viable targets for further study. This requires the integration and analysis of genetic and immunological data for SARS-CoV-2 and its homology with other human coronavirus species to understand cross-reactivity.

Methods: We have developed an online tool for SARS-CoV-2 research, which combines an extensive epitope mapping and prediction meta-analysis, with an updated variant database (55,944 non-synonymous mutations) based on 16,087 whole genome sequences, and an analysis of human coronavirus homology. To demonstrate its utility, we present an integrated analysis of the SARS-CoV-2 spike and nucleocapsid proteins, which are candidate vaccine and serological diagnostic targets.

Results: Our analysis reveals that the nucleocapsid protein in its native form appears to be a sub-optimal target for use in serological diagnostic platforms. Whilst, a further analysis suggests that orf3a proteins may be a suitable alternative target for diagnostic assays.

Conclusions: The tool can be accessed online (http://genomics.lshtm.ac.uk/immuno) and will serve as a useful tool for biological discovery in the fight against SARS-CoV-2. Further, it may be adapted to inform on biological targets in future outbreaks, including new human coronaviruses that spill over from animal hosts.

Background

COVID-19, the disease caused by the SARS-CoV-2 virus, was first characterised in the city of Wuhan, Hubei, and has now spread to over 180 countries, instigating the most recent WHO public health emergency of international concern. With over four million confirmed cases worldwide and more than 280,000 deaths, the COVID-19 pandemic has placed an unfounded burden on the world's healthcare infrastructure and economies [1]. The majority of infections are either asymptomatic or result in mild flu-like disease, with severe cases of viral pneumonia affecting between 1.0% (≥20 years) and 18.4% (≥80 years) of diagnosed patients [2]. Its variable infection outcome, mode of transmission and incubation period together have enhanced the ability of the pathogen to spread efficiently worldwide. As a result, there has been an urgent push for the development of diagnostics, therapeutics and vaccines to aid control efforts.

Current front-line diagnostic strategies apply a quantitative reverse transcription PCR (RT-qPCR) assay on patient nasopharyngeal swabs, using primer/probe sets targeting the *nsp10, RdRp, nsp14*, envelope and nucleocapsid genes; tests endorsed by a number of agencies and health systems [3][4]. Patients hospitalised with severe respiratory disease who are RT-qPCR negative may be radiographically diagnosed (chest x-ray or computerised tomography scan), but in resource-poor and high infection rate settings these methods may prove unviable. Considering the inherent limitations in the sample collection process and transient viral load, RNA detection-based diagnostics may vary in their sensitivity. The demand for serological diagnostics is high, particularly because these tests are capable of detecting SARS-CoV-2 antibodies, a biomarker indicative of infection even if the virus is no longer present. This is essential to address crucial questions like how many people have been infected within a population, including those who may have been asymptomatic, and how long immunity can last.

Numerous lateral flow rapid diagnostic tests (RDTs) and enzyme-linked immunosorbent assays (ELISA) tests have been developed, included an approved IgM/IgG RDT which uses the nucleocapsid protein as a target for the detection of seroconverted individuals [5]. Other assays use the spike protein as an antigen, with some using the receptor binding domain (RBD) as a target, a region with a high level of diversity between alphacoronavirus species [6]. Unlike RNA detection methods, these platforms can identify convalescent patients, which is an important functionality to inform outbreak control efforts. Long-term control strategy will involve vaccine roll-out. There are more than 60 vaccines at different phases of development: pre-clinical, clinical evaluation and roll-out [7,8]. These include vaccines based on a non-replicating adenovirus vector base (Ad5-nCoV), an LNP-encapsulated mRNA, a spike DNA plasmid [7,8], and those using lentiviral-modified dendritic-cell/antigen presenting cells (DC/APC). These latter vaccine platforms utilise *in vivo* clustered APCs to present antigen to the host adaptive immune system. The Ad5-nCoV vaccine uses an adenovirus vector to deliver recombinant SARS-CoV-2 spike protein antigen to vaccinated individuals with a view to elicit a protective humoral immune response.

The discovery of efficacious vaccines along with sensitive and specific serological diagnostics are both dependant on the availability of up-to-date information on viral evolution and immune-informatic analyses. The identification of variable or conserved regions in the proteome of SARS-CoV-2 can inform the rational selection of reverse-design targets in both vaccinology and diagnostic fields, as well as indicate immunologically relevant regions of interest for further studies to characterise SARS-CoV-2 immune responses. Whilst there is some limited and narrow biological data for SARS-CoV-2 in the public domain, insights are most likely to come from its integration informatically. Here we present an online integrated immune-analytic resource for the visualisation and extraction of SARS-CoV-2 meta-analysis data within a circular framework [9]. This platform utilises an automated pipeline for the formation of a whole genome sequence variant database for SARS-CoV-2 isolates worldwide (as of May 10, 2020, n=16,087). We have integrated this dataset with

a suite of B-cell epitope prediction platform meta analyses, HLA-I and HLA-II peptide prediction, an 'epitope mapping' analysis of available experimental *in vitro* confirmed epitope data from The Immune Epitope Database (IEDB) and a protein orthologue sequence analysis of six relevant coronavirus species (SARS, MERS, OC43, HKU1, NL63 and 229E); with all data updated and annotated regularly with information from the UniProt database. With this resource users can browse the SARS-CoV-2 proteome annotated with the above analyses and easily extract meta data to inform further experiments. As a demonstration of the tools function, we present an analysis of the SARS-CoV-2 spike, nucleocapsid and orf3a proteins, which are vaccine and serological targets.

Methods

Whole genome sequence data analysis

SARS-CoV-2 nucleotide sequences were downloaded from NCBI (https://www.ncbi.nlm.nih.gov) and GISAID (https://www.gisaid.org). As a part of an automated in-house pipeline, sequences were aligned using MAFFT software (v7.2) [10] and trimmed to the beginning of the first reading frame (orf1ab-nsp1). Sequences with >20% missing were excluded from the dataset. Using data available from the NCBI COVID-19 resource, a modified annotation (GFF) file was generated and open reading frames (ORFs) for each respective viral protein were extracted (taking in to account ribosomal slippage) using bedtools 'getfasta' function [11]. Each ORF was translated using *EMBOSS transeq* software [12] and the variants for each protein sequence were identified using an in-house script. As a part of our analysis pipeline we generated consensus sequences for each SARS-CoV-2 protein from the nucleotide database using EMBOSS Cons CLI tool [12]. These canonical sequences were used as a reference for prediction, specificity and epitope mapping analyses.

B-cell epitope prediction meta-analysis

Six epitope prediction software platforms were chosen for this analysis (Bepipred [13], AAPpred [14] DRREP [15], ABCpred [16], LBtope [17] and BCEpreds [18]). The scores for by-residue analyses were

selected based on internal probability metrics of 0.8, and then collated. For platforms that output predicted sequences, we used a quality cut-off of 80% and mapped them to the amino acid sequence of each gene. Using a pragmatic approach, the scores across the predictive platforms were then normalised (minimum-maximum scaled) and combined to provide a consensus score.

HLA-I and HLA-II peptide prediction

We chose to use the netMHCpan server for our HLA-I peptide prediction analysis, due to its high overall performance and its extensive HLA-I allele database [19]. We ran predictions for a total of 2,915 alleles (HLA-A 886, HLA-B 1412 and HLA-C 617), across all peptide lengths (8-14 amino acids). The analysis generated 1.1 billion candidates. After QC we selected a total of 736,073 peptides based on strong binding affinity across the allele database for further analysis. We selected strong binding affinity peptides based on the tools internal binding scoring metrics. Only 'strong binding' alleles were selected for further analysis. For each position with a ligand with high binding affinity we analysed the percentage representation of the respective HLA-I type across the allele database. For predicting HLA-II peptides we used the MARIA online tool [20]. We pre-processed the SARS-CoV-2 canonical protein sequences using a 15 amino acid sliding window. We made predictions for all available HLA-II alleles. A 95% cut off was chosen for a positive HLA-II presentation. All data for each 15-mer is displayed on the tool.

Epitope mapping

B-cell epitopes for coronavirus species were sourced from the Immune Epitope Database (IEDB) resource (https://www.iedb.org). Using BLASTp [21] we mapped short amino acid epitope sequences onto the canonical sequence of SARS-CoV-2 proteins. A BLASTp bitscore of 25 with a minimum length of 8 residues was selected as a quality cut-off for mapped epitopes. The frequency of mapped epitopes was logged for each position in the protein and parsed for graphical representation.

Coronavirus homology analysis

Reference proteomes for SARS, MERS, OC43, 229E, HKU1 and NL63 α and β coronavirus (-CoV) species were sourced from UniProt database. These sequences were processed into 10-mers using the *pyfasta* platform and mapped on to the canonical sequences of SARS-CoV-2 proteins using the aforementioned 'epitope mapping' process. Homologous peptide sequences with a BLAST bitscore indicating 10 or more residues mapped to the target sequence were recorded and parsed for display on the graph.

Online SARS-CoV-2 "Immuno-analytics" resource and analysis software

We developed an online resource with an interactive plot that integrates SARS-CoV-2 genetic variation, epitope prediction and mapping, with other coronavirus homology, as well as a table for candidate proteome analysis. This tool is available online (from genomics.lshtm.ac.uk/immuno) (see **S1 figure** for screenshots). The BioCircos.js library [9] was used to generate the interactive plot and *Datatables.net* libraries for the table. The underlying browser software and in-house pipelines for data analysis are available (https://github.com/dan-ward-bio/COVID-immunoanalytics).

Results

Analysis of 16,087 SARS-CoV-2 sequences identified 55,944 non-synonymous mutations across 4,979 sites in protein coding regions. The most frequent mutations were the spike protein D614G (63.5%) and *nsp12*-L314P (63.2%) (**Table 1**). *Nsp12*-L314P is used to genotype the putative S and L strains of SARS-CoV-2, which have now been clustered into further groups [22]. Spike D614G lies 73 residues downstream from the spike RBD, a region of interest as it is a major target of protective humoral responses and bears immunodominant epitopes that play a possible role in antibody dependant enhancement [23–25]. Other high frequency mutations occur on the nucleocapsid gene (R203K, 18.8%; G204R 18.8%), which has been the target antigen for several serological RDTs currently in use

and in production. We have identified 229 non-synonymous variant sites across the nucleocapsid gene with mutations occurring 8,594 times in this dataset.

Using the SARS-CoV-2 immuno-analytics platform we can further query these polymorphic regions for immunological relevance. The 20 residues surrounding the spike mutation D614G (**Figure 1**) have a high epitope prediction meta-score (34% increase on the global median) with 204 IEDB epitope positions mapping to the surrounding residues, suggesting this region is of high interest and may elicit a strong immune response. On top of the high level of SARS-CoV sequence homology reported, we have identified multiple clusters in the S2 domain of the spike protein, with homology to MERS, OC43, 229E, HKU1 and NL63 human coronaviruses, which may result in a cross-reactive antibody response in immune sera. Human coronavirus sequence homology is greatly reduced in the S1 domain, with only two small 10-residue pockets of OC43 and HKU1 identity (see **Figure 1**). We observed a 17% increase over the median epitope meta-score in the receptor-binding motif (AA437-508), a region implicated in ACE2 recognition. HLA-II peptide binding prediction yielded a region within the RBD (S316-330) with high HLA-II ligand probability (0.6), as well as strong B-cell epitope prediction scores (28% above the global median). Meta data obtained from the UniProt database reveals 3 clusters of glycosylated residues, a characteristic that should be considered when choosing expression systems for producing protein/peptides based on these regions.

Looking at the metadata associated with the two high frequency non-synonymous nucleocapsid protein mutations, R203K and G204R, there is a total of 38 variant sites 30 residues either side of these positions with mutations occurring 1,271 times in the dataset (not including R203K and G204R). The average epitope meta-score for these variant sites is 30% above the global median prediction score, with the two aforementioned high frequency mutant residues scoring 35% above the global median epitope predictive score. The sequence homology analysis of the nucleocapsid protein revealed a high level of shared identity between SARS-CoV (90%) and MERS-CoV (45%) on a

per-residue basis. The k-mer mapping technique we have employed to identify homologous regions of >10 residues in proteins allows us to identify pockets of identity between orthologs - sufficiently long - to serve as HLA-bound peptides; something that is challenging using only pairwise multiple sequence alignments. Applying this analysis to the nucleocapsid protein reveals two clusters of shared human coronavirus orthologue identity (**Figure 1**), with an increased IEDB epitope mapping frequency, high polymorphism frequency and epitope meta-scores (23% above the global median) indicative of potential immunogenicity. The first area of interest is a 35-residue region within the nucleocapsid (102-137), which exhibits NL63, SARS, OC43, 229E, MERS and HKU1 human coronavirus homology. Within this region we observed an increase in mapped IEDB epitopes from other coronavirus species providing *in vitro* confirmation that these are indeed cross-reactive epitopes. The second region (167-206) contains the R203K and G204R mutations along with a cluster of high frequency variants. We detected homology with HKU1, NL63 and MERS human coronavirus species along with a high frequency of mapped IEDB epitopes, and a 34% increase on the median epitope prediction meta-score.

We have incorporated an HLA-I peptide prediction analysis into the tool to aid in the scrutiny and development of vaccine candidates. $CD8^+$ effector immunity has been reported to play a central role in the response to SARS-CoV infection, as well as infection mediated immunopathology [26–28]. We used a database of 2,915 HLA-A, HLA-B and HLA-C alleles to make HLA-I peptide binding predictions using netMHCpan 4.1 [19], with peptide lengths of 8 to 14 amino acids across the entire SARS-CoV-2 proteome. Previous studies of adaptive cellular effector immune responses to SARS-CoV infection have emphasised the importance of spike peptide presentation in the progression and severity of disease; regions of particular interest include: S436–443, S525–532, S366–374, S978, S1202 [26,27,29]. We analysed these regions for their performance as HLA-I ligands *in-silico* and found that all of the regions of interest had a high binding affinity score associated with that position. Moreover, these peptides were widely represented in the predictions made across the 2,915 HLA-A,

-B and -C alleles used in this analysis (**Figure 2**), with an average allele coverage across all peptides of 21%, 18% and 34% in HLA-A, -B and -C, respectively. These findings imply that these peptides as HLA-I ligands may indeed also serve to initiate a protective cellular response in SARS-CoV-2 infections across a significant proportion of the HLA-I population worldwide. We have identified another region of interest that scores highly in the HLA-I peptide binding analysis. S690-700 has a high frequency of strongly binding peptides with significant representation across HLA-I alleles: 40%, 23% and 60% for HLA-A, -B and -C, respectively. Furthermore, we have observed no mutations present in this region based on our SARS-CoV-2 variant analysis, implying this peptide appears to remain conserved making it a prime candidate for further study.

Protein 3a (orf3a), has been reported to play a role in host immune modulation by decreasing interferon alpha-receptor expression in SARS-CoV infected cells and activating the NLRP3 inflammasome [30] [31]; a response that may boost inflammation mediated pathology. Orf3a has been a target for SARS-CoV vaccinology studies, with reports of it eliciting potentially protective responses in both protein and DNA forms [32] [28]. These immunogenic properties appear conserved in the SARS-CoV-2 orthologue, with consistently strong antibody responses reported in COVID-19 patients [33]. Looking across the SARS-CoV-2 proteome, of the 50 residues with the highest B-cell epitope prediction meta-score, orf3a occupies 16%, despite only constituting 2.5% percent of the total SARS-CoV-2 protein sequence. Moreover, there are numerous high affinity HLA-II epitopes, which may serve to elicit strong antibody responses. Although protein orf3a shares a high level of identity with its SARS-CoV orthologue, we detected no amino acid sequence homology with OC43, NL63, HKU1 and 229E human coronavirus species. Our analysis of the 16,087 SAR-CoV-2 whole genome sequences detected 163 variant sites within orf3a, although only 17 sites have an alternative allele frequency greater than 0.1%, with non-synonymous mutations occurring in 7,042 samples across the dataset; a comparable level of polymorphism relative to the size of the gene compared to the nucleocapsid protein. The variant sites identified in the orf3a gene have a mean

epitope predictive meta-score of 2.3, which is equal to the median global, indicating that these sites may not form a part of a B-cell epitope. Comparing the predictive meta-scores of the nucleocapsid protein variant sites, we observed an increase of 26% over the global median, indicating that nucleocapsid protein non-synonymous mutations may impact epitope variability more than those found in orf3a. $CD8^+$ effector responses to protein 3a have been characterised in SARS-CoV patients and appear to play a significant role in immunity [27,34,35]. Notably, alongside two within the spike protein, a peptide in orf3a (orf3a36-50) has been found to form a part of the public (conserved) T-cell epitope repertoire across SARS-CoV patients [35]. This region scores highly in the HLA-II predictions with numerous HLA-A and HLA-B high affinity peptides with 19%, 41% and 48% coverage across the HLA-A, B and C database, respectively, and is relatively conserved with few low frequency non-synonymous mutations (mutant allele frequency maximum of 0.0017, N=28).

Discussion

We have developed an immune-analytical tool that combines *in silico* prediction data with *in vitro* epitope mapping, SARS-CoV-2 genome variation and a k-mer-based human coronavirus sequence homology with curated functional annotation data. The integration and co-visualisation of these data support the rational selection of diagnostic, vaccine targets with reverse-immunology and highlight regions for further immunological studies. Using the tool, we focused our analysis on three proteins that are of relevance to current SARS-CoV-2 research, highlighting important features that will inform decisions in producing targets.

Understanding the magnitude of transmission and patterns of infection, will lead to insights for post-isolation strategies. There has been a rush to deploy serological RDTs for the detection of SARS-CoV-2 IgG/IgM antibody responses. There are anecdotal accounts that some tests using spike and nucleocapsid antigens have been found to be insufficiently accurate, and have therefore not yet been deployed. Other assays have been based solely on the nucleocapsid protein and our analysis

reveal that in its native form this protein may prove a sub-optimal target for use in serological diagnostic platforms [5]. It possesses the greatest number of residues across all SARS-CoV-2 genes with high-frequency non-synonymous mutations, the majority of which have a high predictive epitope and IEDB epitope mapping scores when compared to variant positions of other genes. This implies that there may be an inherent variability in dominant antibody responses to different nucleocapsid protein isoforms, which may work to confound testing. We have located three regions of homology with other highly prevalent human coronavirus species, which could serve as non-specific SARS-CoV-2 epitopes if used in serological assays. Moreover, we have emphasised the high level of SARS-CoV identity across the SARS-CoV-2 proteome (except in orf8 and orf10), which may have implications for diagnostic deployment in countries that have had outbreaks involving SARS-CoV.

The spike protein has remained a focus of both vaccine and diagnostic research. Its functional role in viral entry imparts this antigen with immunodominant and neutralising antibody responses [23,36]. This role is confirmed in our analyses, with several clusters of high epitope meta-scores in functional regions, and IEDB epitope mapping counts. The S1 domain in particular has been the focus of a number of studies looking for specific antigens, not least because of its apparent lack of sequence homology with other human coronavirus species when compared to regions in the S2 domain and its apparent immunogenicity [6,33,36,37]. However, as vaccination programmes begin, most of which will target the spike protein in one way or another, it will become challenging to differentiate vaccination responses from those elicited by SARS-CoV-2 infection. There may then be a requirement for alternative viable targets for serological screening.

The broad nature of the analyses chosen for this tool may assist in the understanding of vaccine targets, both during design and testing phases. The prediction of HLA-I ligands is relevant not only to the study of functional viral targets, but the full range of potentially immunologically relevant

endogenous proteins analysed here that may be presented following intracellular processing, some of which may have less coverage in the literature. Our broad approach to HLA-I ligand prediction ensures that researchers understand the applicability of *in-silico* informed vaccine targets across different populations, a vital factor in pandemic situations.

Ensuring that targets are both specific and devoid of polymorphism is essential to ensuring the longevity of vaccine responses and diagnostic capabilities, the analysis of which is achieved easily with our tool. The humoral and cellular immune responses, as well as the effects of H-CoV protein homology to SARS-CoV-2 proteins have yet to be fully characterised. With the significant levels of amino-acid sequence identity between SARS-CoV-2 and other H-CoV species detected in our analysis, researchers should be wary of the potentially deleterious effects of both non-specific humoral and cellular responses in enhancing infection, a phenomenon observed in a number of other viral pathology models.

Using the SARS-CoV-2 immuno-analytics platform we were able to identify shortcomings in current targets for diagnostics and suggest orf3a as another target for further study. This protein has proven *in vitro* immunogenicity in COVID-19 patients, as well as an array of supportive results from analyses performed here. The database underpinning the online tool will be updated regularly with all IEDB epitopes, mutations and functional annotations as they become available. Importantly, this open-access platform and tool enables the acquisition of all of the aforementioned data associated with the SARS-CoV-2 proteome, assisting further important research on COVID-19 control tools.

Conclusions

The SARS-CoV-2 Immunoanalytics Platform enables the straightforward visualisation of ‘omic data to inform research in vaccine, diagnostic and immunology research. By integrating genomic and proteomic analyses with *in-silico* epitope predictions, we have highlighted important advantages and

shortcomings of two proteins at the foci of COVID-19 research (spike and nucleocapsid), while suggesting another candidate for further study (orf3a). Both spike and nucleocapsid proteins have regions of high identity shared with other endemic H-CoV species. Moreover, several high frequency mutations found in our dataset lie within putative T and B-cell epitopes, something that should be taken into consideration when designing vaccines and diagnostics.

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The sequencing data analysed during the current study are available from GISAID (https://www.gisaid.org) and NCBI (https://www.ncbi.nlm.nih.gov). Full analysis datasets can be downloaded from www.genomics.lshtm.ac.uk/imunno or https://github.com/dan-ward-bio/COVID-immunoanalytics.

Competing interests

The authors declare that they have no competing interests

Funding

DW is funded by a Bloomsbury Research PhD studentship. SC is funded by Medical Research Council UK grants (MR/M01360X/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1). TGC is funded by the Medical Research Council UK (Grant no. MR/M01360X/1, MR/N010469/1, MR/R025576/1, and MR/R020973/1) and BBSRC (Grant no. BB/R013063/1).

Authors' contributions

DW, SC and TGC conceived and directed the project. MH and JEP provided software and informatic support. DW and JEP performed bioinformatic and statistical analyses under the supervision of TGC.

DW, MLH, SC and TGC interpreted results. DW wrote the first draft of the manuscript. All authors commented and edited on various versions of the draft manuscript and approved the final manuscript. DW, TGC and SC compiled the final manuscript.

Acknowledgments

We gratefully acknowledge the laboratories who submitted the data to the IEDB, NCBI and GISAID public databases on which this research is based. We also thank IEDB, NCBI and GISAID for developing and curating their databases. We gratefully acknowledge the availability of the Medical Research Council UK funded eMedLab (HDR UK) computing resource. DW wishes to thank Fraser, Josh, Stu and Paul for their support.

References

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. 2020;0. doi:10.1016/S1473-3099(20)30120-1
2. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019 : a model-based analysis. Lancet Infect Dis. 2020;3099: 1–9. doi:10.1016/S1473-3099(20)30243-7
3. Vogels CBF, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC, et al. Analytical sensitivity and efficiency comparisons of SARS-COV-2 qRT-PCR assays. medRxiv. 2020; 2020.03.30.20048108. doi:10.1101/2020.03.30.20048108
4. CDC. Processing of Sputum Specimens for Nucleic Acid Extraction. 2020. Available: www.cdc.gov/lab
5. JHU C for HS. Global Progress on COVID-19 Serology-Based Testing. [cited 3 Apr 2020]. Available: http://www.centerforhealthsecurity.org/resources/COVID-19/Serology-based-tests-for-COVID-19.html#sec1
6. Lassaunière R, Frische A, Harboe ZB, Nielsen ACY, Fomsgaard A, Krogfelt KA, et al. Evaluation

of nine commercial SARS-CoV-2 immunoassays. [cited 17 Apr 2020]. doi:10.1101/2020.04.09.20056325

7. Thanh Le T, Andreadakis Z, Kumar A, Gómez Román R, Tollefsen S, Saville M, et al. The COVID-19 vaccine development landscape. Nat Rev Drug Discov. 2020 [cited 15 Apr 2020]. doi:10.1038/d41573-020-00073-5

8. WHO. DRAFT landscape of COVID-19 candidate vaccines. [cited 3 Apr 2020]. Available: https://www.who.int/blueprint/priority-diseases/key-action/novel-coronavirus-landscape-ncov.pdf?ua=1

9. Cui Y, Chen X, Luo H, Fan Z, Luo J, He S, et al. BioCircos.js: an interactive Circos JavaScript library for biological data visualization on web applications. Bioinformatics. 2016;32: 1740–2. doi:10.1093/bioinformatics/btw041

10. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010

11. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

12. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. Elsevier Ltd; 2000. pp. 276–277. doi:10.1016/S0168-9525(00)02024-2

13. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 2017;45: W24–W29. doi:10.1093/nar/gkx346

14. Davydov YI, Tonevitsky AG. Prediction of linear B-cell epitopes. Mol Biol. 2009;43: 150–158. doi:10.1134/S0026893309010208

15. Sher G, Zhi D, Zhang S. DRREP: deep ridge regressed epitope predictor. BMC Genomics. 2017;18: 676. doi:10.1186/s12864-017-4024-8

16. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent

neural network. Proteins Struct Funct Bioinforma. 2006;65: 40–48. doi:10.1002/prot.21078

17. Singh H, Ansari HR, Raghava GPS. Improved Method for Linear B-Cell Epitope Prediction Using Antigen's Primary Sequence. Schönbach C, editor. PLoS One. 2013;8: e62216. doi:10.1371/journal.pone.0062216

18. Saha S, Raghava GPS. ICARIS 2004, LNCS 3239. 2004.

19. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. J Immunol. 2017;199: 3360–3368. doi:10.4049/jimmunol.1700893

20. Chen B, Khodadoust MS, Olsson N, Wagar LE, Fast E, Liu CL, et al. Predicting HLA class II antigen presentation through integrated deep learning. Nat Biotechnol. 2019;37: 1332–1343. doi:10.1038/s41587-019-0280-2

21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

22. Phelan J, Deelder W, Ward D, Campino S, Hibberd ML, Clark TG. Controlling the SARS-CoV-2 outbreak, insights from large scale whole genome sequences generated across the world. bioRxiv. 2020; 2020.04.28.066977. doi:10.1101/2020.04.28.066977

23. Sui J, Deming M, Rockx B, Liddington RC, Zhu QK, Baric RS, et al. Effects of human anti-spike protein receptor binding domain antibodies on severe acute respiratory syndrome coronavirus neutralization escape and fitness. J Virol. 2014;88: 13769–80. doi:10.1128/JVI.02232-14

24. Wang SF, Tseng SP, Yen CH, Yang JY, Tsao CH, Shen CW, et al. Antibody-dependent SARS coronavirus infection is mediated by antibodies against spike proteins. Biochem Biophys Res Commun. 2014;451: 208–214. doi:10.1016/j.bbrc.2014.07.090

25. Korber B, Fischer W, Gnanakaran SG, Yoon H, Theiler J, Abfalterer W, et al. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv. 2020; 2020.04.29.069054. doi:10.1101/2020.04.29.069054

26. Zhi Y, Kobinger GP, Jordan H, Suchma K, Weiss SR, Shen H, et al. Identification of murine CD8 T cell epitopes in codon-optimized SARS-associated coronavirus spike protein. Virology. 2005;335: 34–45. doi:10.1016/j.virol.2005.01.050

27. Channappanavar R, Fett C, Zhao J, Meyerholz DK, Perlman S. Virus-Specific Memory CD8 T Cells Provide Substantial Protection from Lethal Severe Acute Respiratory Syndrome Coronavirus Infection. J Virol. 2014;88: 11034–11044. doi:10.1128/jvi.01505-14

28. Lu B, Tao L, Wang T, Zheng Z, Li B, Chen Z, et al. Humoral and cellular immune responses induced by 3a DNA vaccines against severe acute respiratory syndrome (SARS) or SARS-like coronavirus in mice. Clin Vaccine Immunol. 2009;16: 73–77. doi:10.1128/CVI.00261-08

29. Chen H, Hou J, Jiang X, Ma S, Meng M, Wang B, et al. Response of Memory CD8 + T Cells to Severe Acute Respiratory Syndrome (SARS) Coronavirus in Recovered SARS Patients and Healthy Individuals . J Immunol. 2005;175: 591–598. doi:10.4049/jimmunol.175.1.591

30. 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. doi:10.1371/journal.pone.0008342

31. Siu KL, Yuen KS, Castano-Rodriguez C, Ye ZW, Yeung ML, Fung SY, et al. Severe acute respiratory syndrome Coronavirus ORF3a protein activates the NLRP3 inflammasome by promoting TRAF3-dependent ubiquitination of ASC. FASEB J. 2019;33: 8865–8877. doi:10.1096/fj.201802418R

32. Zhong X, Guo Z, Yang H, Peng L, Xie Y, Wong TY, et al. Amino terminus of the SARS coronavirus protein 3a elicits strong, potentially protective humoral responses in infected patients. J Gen Virol. 2006;87: 369–374. doi:10.1099/vir.0.81078-0

33. 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; 2020.03.26.994756. doi:10.1101/2020.03.26.994756

34. Oh H-LJ, Chia A, Chang CXL, Leong HN, Ling KL, Grotenbreg GM, et al. Engineering T Cells

Specific for a Dominant Severe Acute Respiratory Syndrome Coronavirus CD8 T Cell Epitope. J Virol. 2011;85: 10464–10471. doi:10.1128/jvi.05039-11

35. Li CK, Wu H, Yan H, Ma S, Wang L, Zhang M, et al. T Cell Responses to Whole SARS Coronavirus in Humans. J Immunol. 2008;181: 5490–5500. doi:10.4049/jimmunol.181.8.5490

36. Wu F, Wang A, Liu M, Wang Q, Chen J, Xia S, et al. Neutralizing antibody responses to SARS-CoV-2 in a COVID-19 recovered patient cohort and their implications. medRxiv. 2020; 2020.03.30.20047365. doi:10.1101/2020.03.30.20047365

37. Okba NMA, Müller MA, Li W, Wang C, Geurtsvankessel CH, Corman VM, et al. SARS-CoV-2 specific antibody responses in COVID-19 patients. [cited 16 Apr 2020]. doi:10.1101/2020.03.18.20038059

Table 1: Most frequent non-synonymous mutations found in the 16,087 global SARS-CoV-2 whole genome sequences

Protein*	Pos.	Ref. Allele	Alt. Allele	Freq.	Dominant Alt. Allele Freq. (N=16,087)	Europe (9,101)	Asia (1,018)	NAm (4,424)	SAm (108)	AFR (110)	OCE (1326)	UK** (5,531)
S	614	D	G	10193	0.635	0.709	0.601	0.633	0.630	0.645	0.625	0.639
nsp12	323	P	T:F:L	1:3:10163	0.632	0.707	0.597	0.632	0.630	0.655	0.621	0.638
orf3a	57	Q	H	3819	0.238	0.263	0.216	0.241	0.213	0.245	0.231	0.241
nsp2	85	T	I	3179	0.199	0.220	0.181	0.202	0.185	0.200	0.193	0.200
N	203	R	M:K:S	3:3008:2	0.188	0.212	0.196	0.184	0.185	0.200	0.185	0.190
N	204	G	R	3003	0.188	0.211	0.196	0.183	0.185	0.200	0.185	0.189
nsp6	37	L	F	2132	0.134	0.145	0.151	0.125	0.167	0.118	0.151	0.126
orf8	84	L	S	1982	0.124	0.134	0.135	0.125	0.139	0.118	0.121	0.124
orf3a	251	G	C:V	3:1609	0.100	0.110	0.106	0.098	0.120	0.100	0.109	0.097
nsp13	541	Y	C	1219	0.076	0.084	0.080	0.077	0.093	0.073	0.065	0.078
nsp13	504	P	L:S	1190:5	0.075	0.082	0.082	0.075	0.093	0.073	0.064	0.075
nsp2	585	P	S	696	0.044	0.047	0.035	0.043	0.065	0.064	0.050	0.042
nsp2	559	I	V	641	0.041	0.043	0.034	0.040	0.065	0.055	0.047	0.038
nsp2	212	G	D:C	505:2	0.031	0.035	0.036	0.031	0.056	0.045	0.025	0.030
nsp3	58	A	T:S	464:1	0.029	0.032	0.031	0.028	0.056	0.045	0.023	0.028
orf8	24	S	L	372	0.023	0.025	0.021	0.023	0.009	0.027	0.026	0.022
M	175	T	A:M	1:342	0.021	0.025	0.023	0.021	0.009	0.027	0.020	0.021
nsp5	15	G	D:S	1:328	0.020	0.021	0.022	0.021	0.009	0.018	0.023	0.020
orf3a	13	V	L:A	323:1	0.020	0.023	0.016	0.020	0.019	0.045	0.020	0.023
N	193	S	I	274	0.017	0.020	0.017	0.017	0.037	0.018	0.010	0.018

Pos. = position; Freq. = Frequency; NAm = North America, SAm = South America, AFR = Africa, OCE = Oceania; REF = Reference; ALT= Alternative; * S = Spike, M = Membrane, N = Nucleocapsid, ** included in Europe

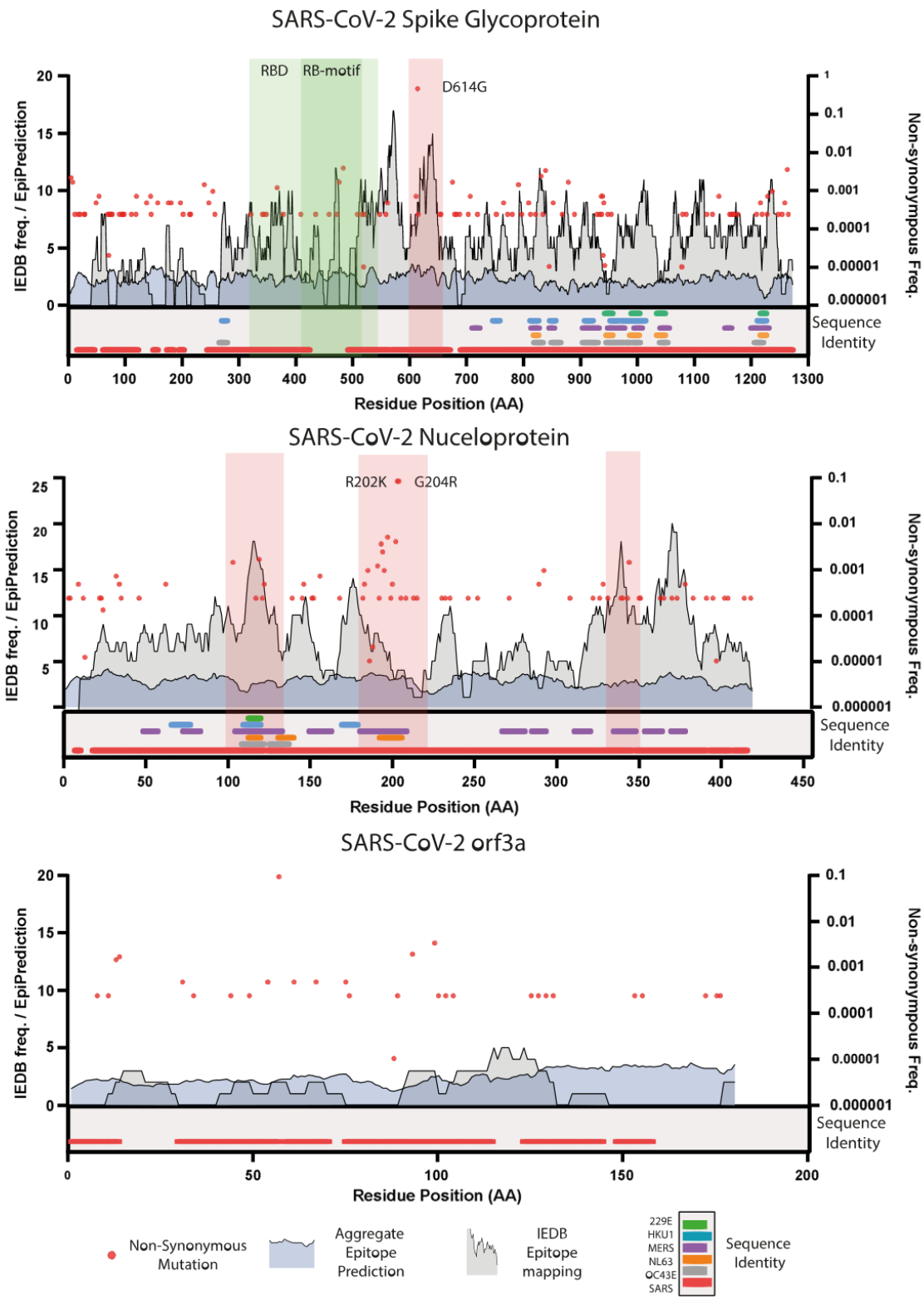


Figure 1. Linearised extracts from the SARS-CoV-2 immuno-analytics resource database; Spike, Nucleocapsid and orf3a proteins. Non-synonymous mutations, epitope prediction meta-score, IEDB epitope mapping and sequence identity analyses were plotted (see key). **Left axis** denotes scale for epitope prediction and IEDB mapping. **Right axis** denotes the relative allele frequency found in the genome dataset; AA amino acid

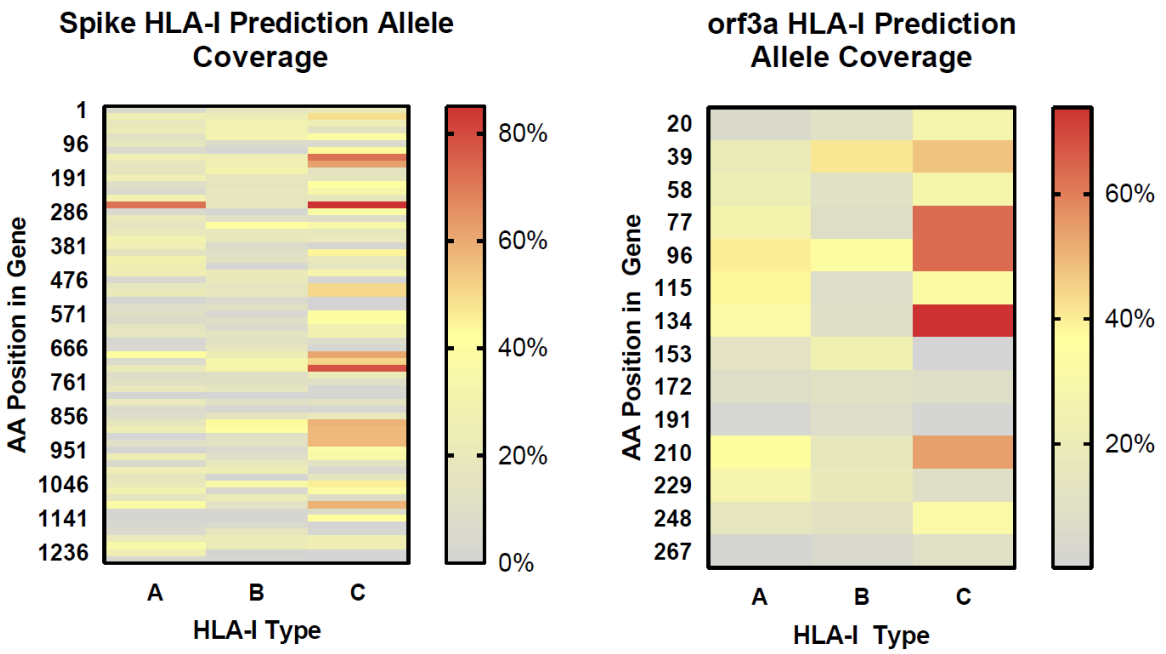
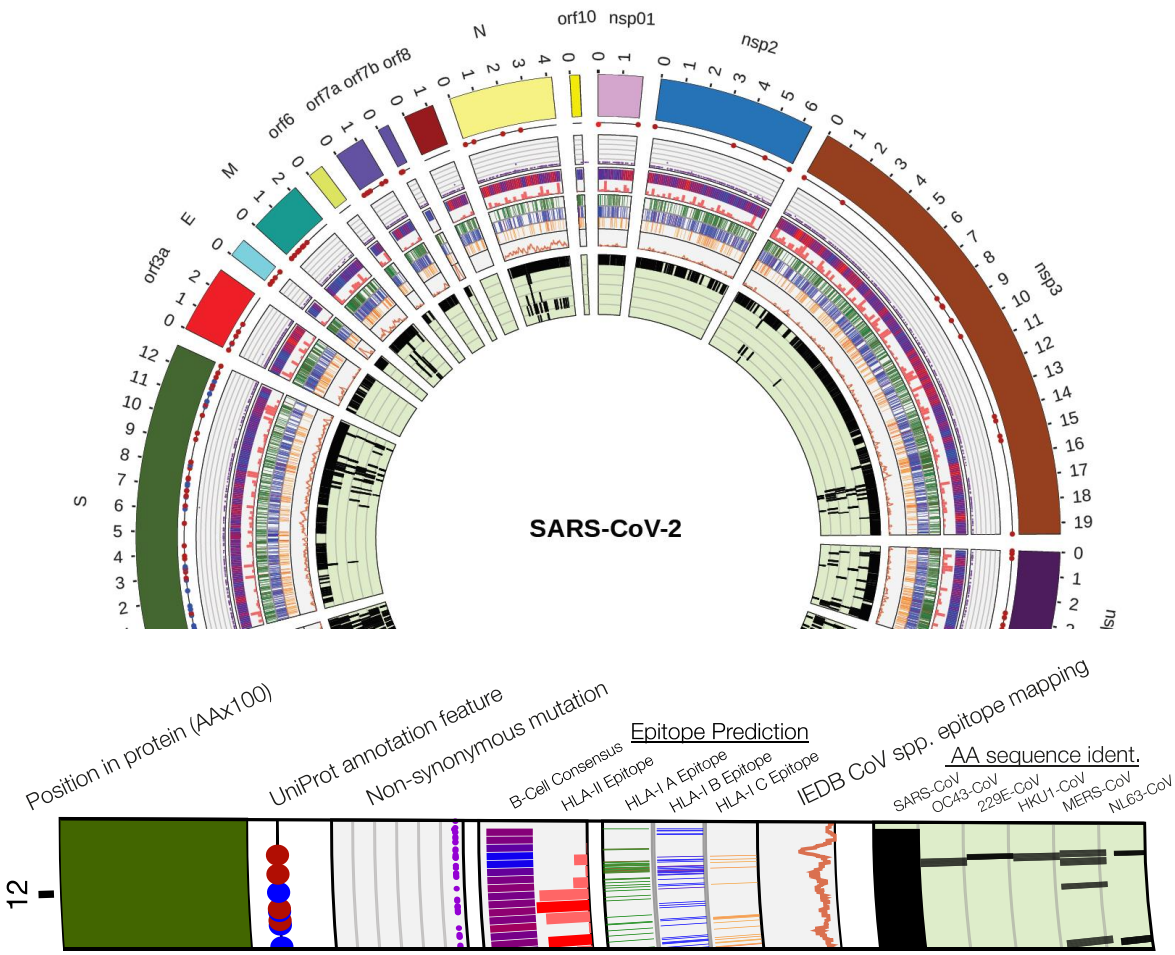


Figure 2. HLA-I A, B and C allele representation in strong binding epitopes. The data was subsampled using a 20 AA sliding window (10 AA upstream/downstream of each position) displaying the maximum representation in a single epitope within the window. Displayed values were calculated based on the allele dataset used for epitope prediction (HLA-A 886, HLA-B 1412 and HLA-C 617).

S1 Figure

The Immunoanalytics webpage (http://genomics.lshtm.ac.uk/immuno)

(A) Interactive circular view with informative tracks



(B) A search tool in a table format

SARS-CoV-2 Residue Metadata

AA-Pos Start: AA-Pos End:

Copy CSV Excel PDF Print Search:

Gene	Protein CDS Position	Mutation Occurances	Mutation Freq.	REF	ALT	Bepipred	AAPPRed	bcepreds	Lbtope	DRREP	ABCpreds	Epitope Prediction Total	CoV Homology
nsp01	1	NA	NA	M	NA	0.2546666667		0.02525	0.4139	0	0		(SARS)
nsp2	1	NA	NA	Y	NA	0.242	0.378365	-0.085	0.5842	0	0.25	1.369565	(OC43_SARS)
nsp3	1	NA	NA	P	NA	0.2281111111	0.71646	0.157	0.4743	0.5	0.25	2.3258711111	(SARS_SARS)
nsp4	1	NA	NA	I	NA	0.2692222222	0.2925375	-0.296	0.5148	0	0	0.7805597222	(HKU1_MERS_OC43_SARS)
nsp5	1	NA	NA	G	NA	0.2403333333	0.0316475	0.1775	0.6974	0	0.25	1.3968808333	(HKU1_MERS_229E_NL63_OC
nsp6	1	NA	NA	A	NA	0.2114444444	-0.0237725	0.11025	0.4787	0.5	0.25	1.5266219444	NA
nsp7	1	NA	NA	K	NA	0.2322222222	0.5130975	0.39325	0.581	0	0.25	1.9695697222	NA
nsp8	1	NA	NA	I	NA	0.2595555556	0.3265	0.12825	0.4944	0	0.25	1.4587055556	NA
nsp9	1	NA	NA	N	NA	0.2716666667	0.1959525	0.206	0.6533	0	0	1.3269191667	NA
nsp10	1	NA	NA	G	NA	0.2515555556	0.4671925	0.233	0.6083	0	0.25	1.8100480556	NA