Prediction of T and B cell epitopes in the proteome of SARS-CoV-2 for potential use in diagnostics and vaccine design

Parvez Singh Slathia*, Preeti Sharma

School of Biotechnology, Shri Mata Vaishno Devi University, Kakrial, Katra, J&K, India

*corresponding author: parvezsingh@gmail.com, parvez.singh@smvdu.ac.in.

Abstract:

The world is currently battling the Covid-19 pandemic for which there is no therapy available. Prophylactic measures like vaccines can effectively thwart the disease burden. The current methods of detection are PCR based and require skilled manpower to operate. The availability of cheap and ready to use diagnostics like serological methods can ease the detection of SARS-CoV-2 virus. In the current study, immunoinformatics tools have been used to predict T and B cell epitopes present in all the proteins of this virus. NetMHCPan, NetCTL and NetMHCII servers were used for T cell epitope prediction while BepiPred and ABCPred were used for B cell epitope prediction. Population coverage analysis for T cell epitopes revealed that these could provide protection to the people throughout world. The T cell epitopes can exclusively used for vaccine design whereas B cell epitopes can be used for both vaccine design and developing diagnostic kits.

Introduction:

World Health Organization (WHO) was apprised on 31 December, 2019 about pneumonia of unknown cause in Wuhan city of Hubei province, China. The disease caused by a novel coronavirus has been termed as COVID-19 and was declared a pandemic on 11 March, 2020 by WHO. According to WHO till 23 March, 2020 294,110 people are infected and 12,944 killed with 187 countries areas or territories being affected (https://www.who.int/emergencies/diseases/novel-coronavirus-2019). The virus was named as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. The symptoms of COVID-19 generally are fever, cough, fatigue, and can also include sputum production, headache, haemoptysis, diarrhoea, dyspnoea, and lymphopenia [2]. In severe cases the symptoms include pneumonia, acute respiratory distress syndrome, acute cardiac injury and incidence of grand-glass opacities that lead to death [3]. Person to person transmission is the

major mode of transmission which occurs either through direct contact or through large droplets spread by coughing or sneezing from an infected individual [2] . The virus is also found in stool and blood which necessitates to investigate these routes of infection as well [4]. The reproduction number, R_0 indicates average number of new infections generated by an infectious person in a totally naïve population and R_0 of this virus is estimated between 2 and 3. $R_0 > 1$ signifies that the infection is likely to increase and could result in the spread of epidemic [5]. The diagnosis of the virus is currently carried out by detecting viral nucleic acid by PCR in nasopharyngeal swabs [6]. No serological diagnostic method has been licensed yet. Though there are some rapid test kits developed but none has been recommended by WHO. The first genome sequence of SARS-CoV-2 was available on 11 January, 2020 and efforts for developing vaccines are in rapid progression. The first vaccine mRNA-1273 developed by NIAID and Moderna, Inc.entered clinical trials on 16 March, 2020. The vaccine is a novel lipid nanoparticle (LNP)-encapsulated mRNA-based vaccine that encodes for a full-length, prefusion stabilized spike (S) protein of SARS-CoV-2 (https://www.nih.gov/news-events/news-releases/nih-clinical-trial-investigational-vaccine-covid-19-begins). There are many other vaccine candidates that are in pipeline including live attenuated, inactivated, DNA, peptide based, recombinant vector, recombinant protein which are either at development stage and some have entered clinical trials as well.

The genome sequences of organisms make it possible to ascertain the encoded protein repertoire. Bioinformatics tools can be used to find out the epitopic regions in these proteins. The epitopes are the regions that are identified by T and B cells. Immunoinformatics has developed as a branch of bioinformatics which deals with the prediction of T and B cell epitopes which can be used for vaccine design [7]. There are many studies on vaccine design using this approach for different pathogens; virus (West Nile virus and Japanese Encephalitis virus) [8], parasites (*Trypanosoma cruzi*) [7], bacteria (*Listeria monocytogenes*) [9]. B cell epitopes can also be used in design of diagnostics wherein the epitopic regions can be mapped to their respective antibodies. These antibodies if having high titre in the patient's serum can be used for developing serological kits. In the current study, we have predicted T and B cell epitopes in the proteome of SARS-CoV-2 using different bioinformatics tools which can find use in designing of vaccines and diagnostics.

Materials and Methods:

Source of sequences:

NCBI genome database was used to find the genome of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). From the genome sequence, all protein sequences were obtained. Each protein sequence was used to find its identical protein sequences in Identical Protein Groups (IPG) database (https://www.ncbi.nlm.nih.gov/ipg/). Identical protein groups include protein accessions from GenBank, RefSeq, SwissProt, PDB and other sources. IPG makes it easier to find protein information by searching against groups of protein records where each group represents a unique protein sequence.

Class I MHC epitope prediction:

NetMHCPan 4.0 server [10] (http://www.cbs.dtu.dk/services/NetMHCpan/) was used to find the high affinity binding peptides to different HLA alleles. This server uses artificial neural networks (ANNs) for finding the binding affinity of peptides for Class I MHC. Ten protein sequences in table 1 were used for carrying out predictions. In this case HLA supertype alleles available at the server were used for running predictions. The classification of HLA alleles into supertypes is based on that alleles within supertypes are expected to exhibit widely overlapping peptide repertoires, whereas alleles from different supertypes would more frequently bind non-overlapping sets of peptides [11]. The peptides with high binding affinity for Class I MHC may not necessarily activate Tc cells. Amenability to proteasomal cleavage and affinity for transport by TAP (Transporters associated with antigen processing) are the other two important features required for a peptide to act as Tc cell epitope. NetCTL 1.2 server, [12] which integrates prediction of peptide MHC class I binding, proteasomal C terminal cleavage and TAP transport efficiency can serve this purpose. The peptides obtained from NetMHCPan 4.0 server were used for analysis by NetCTL server. Those peptides that were predicted as epitopes by NetCTL have greater chance of activating Tc cells into cytotoxic T lymphocyte (CTL).

Class II MHC epitope prediction:

NetMHCII 2.3 server [13] was used for prediction of Class II MHCepitopes. This server uses ANN approach for predicting binding of peptides to HLA-DR, HLA-DQ, HLA-DP class II MHC molecules. All the protein sequences mentioned in table 1 were used for predictions.

Population Coverage:

The population coverage analysis helps in finding the efficacy of epitopes in different population groups of the world. This is based on HLA allele distribution across different populations in the world. Population coverage analysis tool at IEDB analysis resource was used to check the coverage of different population groups by the epitopes (http://tools.immuneepitope.org/population/). The prediction for coverage was carried out combined for Class I and II MHC. BLAST search against the NCBI protein database for the epitopes was performed to check if there exists any similarity to any of the human proteins.

B cell Epitope prediction:

BepiPred-2.0 [14] and ABCPred [15] servers were used to find Linear B cell epitopes. BepiPred server utilizes a random forest algorithm trained on epitopes annotated from antibody-antigen protein structures whereas ABCPred is ANN based B cell epitope prediction server. The peptides that were predicted as epitopes by both the servers were considered.

Results and Discussion:

Protein sequences:

The genome sequence of SARS-CoV-2 isolate WIV07 bearing accession no. MN996531 was used as source. The protein sequences from this genome were used in this study. The protein sequences were retrieved from NCBI Protein database. Each of the protein sequences was also searched in Identical Protein Sequence (IPG) Database and all the sequences for each protein were aligned by Clustal Omega. Interestingly, all the sequences were identical which implies that the sequences of all the proteins were conserved. Since this is a novel virus the sequence entries for it are low in number, which could be the reason for identical protein sequences. IPG database has benefit over other protein sequence databases as one sequence can be used as representative for a group of identical sequences. Thus, unnecessary redundancy of sequence data can be avoided. The list of protein sequences used for epitope prediction is given in table 1. The variation in the sequence is one of the main reasons for immune evasion by the pathogens. The conservancy in the protein sequences means that the epitopes remain consistent and therefore the immune system continues to recognize them.

Prediction of Class I MHC epitopes:

NetMHCPan 4.0 server predicted the peptides that bind to different Class I MHC alleles. The selection of binding peptides was made on the basis of binding score. The peptides chosen as epitopes had threshold score value of 0.8 or more. Such peptides showed predicted binding affinity of less than 10nM. The server uses a default threshold score value of 0.5 however, a higher score value was used to find epitopes having higher binding affinity for MHC alleles. The predicted epitopes with their respective scores are shown in table 2 (for polyprotein Supplementary table 1). The highest numbers of epitopes were predicted for spike protein sequence. NetMHCPan is one of the better performing methods among the available methods for class I MHC binding predictions [16]. Thus, this method was chosen for running predictions. There were very less promiscuous epitopes predicted in the study (indicated in red in table 2). Promiscuous nature of epitopes is a desirable property as one epitope can bind to different alleles. Only spike protein showed the presence of four promiscuous epitopes. All the peptides showing high binding affinity may not be able to activate Tc cells. For a peptide to act as Tc cell epitope it must be amenable to processing by proteasome, show affinity to be transported by TAP in addition to affinity for Class I MHC. NetCTL 1.2 server predicted the CTL epitopes among the cohort of high affinity Class I MHC binding peptides. The combined score value of greater than 1 was used as selection criterion for CTL epitopes (though default threshold value is 0.75). Since NetCTL has been trained on human data therefore it is supposed to give better performance for human proteasome and TAP [12]. Thus, the peptides fulfilling both the criteria (NetMHCPan score 0.8 or more and NetCTL score more than 1) were selected as epitopes (table 2). These epitopes having high binding affinity for MHC molecules and having ability to be processed by cytosolic pathway, transported by TAP have higher probability to function as CTL epitopes.

Prediction of Class II MHC epitopes:

NetMHCII 2.3 server predicted epitopes are shown on table 3 (for polyprotein Supplementary table 2). The selection criterion used for epitopes was score of 0.8 or more. Further, the set of epitopes having the identical core amino acid residues were represented only once in the set of selected epitopes. The epitopes were further analyzed for overlaps in sequence upon which it was found that many of Class II MHC epitopes had significant overlaps. Such regions of different proteins are depicted in table 4. Class II MHC epitopes are processed by the endocytic pathway passing through endosomes and lysosomes before finally being presented to Th cells along with MHC II. However, the methods for these

predictions are not available and therefore, the binding affinity to Class II MHC is the only criterion which can be used to predict such epitopes [17].

Population Coverage:

Population coverage gives information about the efficiency of epitopes to generate immune response in different population groups located in geographical locations across the world. This analysis is based on the distribution of HLA alleles among different population groups throughout the world [18]. The population coverage tool utilizes the HLA allele frequencies of population groups of the world obtained from Allele Frequency database (http://www.allelefrequencies.net/). In this case the MHC restricted epitopes were able to cover the populations throughout the world (figure 1). The highest coverage was that for European population (99.08%) whereas lowest for South Africa (44.31%). The epitopes could achieve more than 90% coverage in ten population groups. The predicted epitopes can bind to MHC alleles which show distribution pan world. The vaccine designed using these epitopes can provide immunity to all the population groups in the world. For determining the homology of these epitopes with human proteins BLAST analysis was done. None of the epitopes shared identity with human protein sequences eliminating the chances of these being recognized as self epitopes.

B cell epitope prediction:

The threshold score used for both the servers was 0.5 and the peptides having higher score were taken as epitopes. The common epitopes predicted by both servers were used. The prediction for orf1ab/polyprotein could not be run as BepiPred server can run predictions for a protein of $\leq$ 6000 amino acid residues. B cell epitopes fall into two categories – linear and conformational. For the want of all protein structures of SARS-CoV-2, in our study we have only predicted linear epitopes. The linear B cell epitopes are depicted in table 5. The prediction as B cell epitopes by two servers strengthen the chances of being recognized by antibodies.

Designing of vaccine and diagnostics:

T and B cell epitopes predicted during the study can be used for vaccine design. MHC binding assays for T cell epitopes can be carried out to find out their binding efficiency to respective HLA alleles. Both T and B cell epitopes can be used for designing synthetic peptide vaccines and their efficacy needs to be checked. Alternatively, synthetic genes can be

constructed encoding these epitopes which can be cloned in vectors like pcDNA, pVAX to generate DNA vaccines. The use of adjuvant can further enhance the immunogenicity of the vaccines. Suitable adjuvant particularly cytokines can be used along with the designed constructs to check their effect [19]. These genes can be introduced into adenovirus based vector vaccine to produce recombinant vector vaccines. Adenovirus based vaccines have the capacity to elicit sustained T cell response and increase the protective immunity [20].

B cell epitopes could have potential for designing of future vaccines. Not only vaccines, these epitopes can find use in diagnostics as well. Antigen antibody interactions form the basis of majority of diagnostics, as they are very easy to perform and are very sensitive and specific in nature. In some cases the antigen can directly be evaluated like in case of pregnancy kit where presence of human chorionic gonadotrophin hormone (HCG) is ascertained in urine. However, sometimes the antigen concentration is very low so alternative means have to be devised. This approach is used in some HIV test kits where antibodies raised against HIV protein function as antigen like Biorad-Genie Fast HIV kit [21]. The antibodies now detect the antibodies generated against HIV protein as antigen. If the amount of antigens required for detection is low in serum then the second approach seems more plausible where antibodies secreted against the virus can be targeted. The predicted B cell epitopes in our study can be synthesized and immobilized to a matrix. The serum of the infected persons can be used to screen the antibodies that can bind to these epitopes. The antibodies thus found can form the basis of a future diagnostic kit. The antibodies with high titre can be slelected. Generation of monoclonal antibodies for detection of these antibodies would form the next step in devising a diagnostic kit. The epitopes can lead to the generation of easy to use serological kit which can be designed on the platform of previously available kits which can be used out of the laboratory. In conclusion, the epitopes predicted in this study may be used for designing of vaccines and diagnostic kits.

References:

1. Gorbalenya, A.E., *Severe acute respiratory syndrome-related coronavirus–The species and its viruses, a statement of the Coronavirus Study Group.* BioRxiv, 2020.
2. Rothan, H.A. and S.N. Byrareddy, *The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak.* Journal of autoimmunity, 2020: p. 102433.
3. Huang, C., et al., *Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China.* The Lancet, 2020. **395**(10223): p. 497-506.
4. del Rio, C. and P.N. Malani, *COVID-19—new insights on a rapidly changing epidemic.* Jama, 2020.
5. Liu, Y., et al., *The reproductive number of COVID-19 is higher compared to SARS coronavirus.* Journal of travel medicine, 2020.
6. Carver, C. and N. Jones, *Comparative accuracy of oropharyngeal and nasopharyngeal swabs for diagnosis of COVID-19*. 2020, Oxford COVID-19 evidence service team centre for evidence based medicine.
7. Slathia, P.S. and P. Sharma, *Conserved epitopes in variants of amastin protein of Trypanosoma cruzi for vaccine design: A bioinformatics approach.* Microbial pathogenesis, 2018. **125**: p. 423-430.
8. Slathia, P.S. and P. Sharma, *A common conserved peptide harboring predicted T and B cell epitopes in domain III of envelope protein of Japanese Encephalitis Virus and West Nile Virus for potential use in epitope based vaccines.* Comparative immunology, microbiology and infectious diseases, 2019. **65**: p. 238-245.
9. Jahangiri, A., et al., *An in silico DNA vaccine against Listeria monocytogenes.* Vaccine, 2011. **29**(40): p. 6948-6958.
10. Jurtz, V., et al., *NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data.* The Journal of Immunology, 2017. **199**(9): p. 3360-3368.
11. dos Santos Francisco, R., et al., *HLA supertype variation across populations: new insights into the role of natural selection in the evolution of HLA-A and HLA-B polymorphisms.* Immunogenetics, 2015. **67**(11-12): p. 651-663.
12. Larsen, M.V., et al., *Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction.* BMC bioinformatics, 2007. **8**(1): p. 424.
13. Jensen, K.K., et al., *Improved methods for predicting peptide binding affinity to MHC class II molecules.* Immunology, 2018. **154**(3): p. 394-406.
14. Jespersen, M.C., et al., *BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes.* Nucleic acids research, 2017. **45**(W1): p. W24-W29.
15. Saha, S. and G.P.S. Raghava, *Prediction of continuous B-cell epitopes in an antigen using recurrent neural network.* Proteins: Structure, Function, and Bioinformatics, 2006. **65**(1): p. 40-48.
16. Trolle, T., et al., *Automated benchmarking of peptide-MHC class I binding predictions.* Bioinformatics, 2015. **31**(13): p. 2174-2181.
17. Backert, L. and O. Kohlbacher, *Immunoinformatics and epitope prediction in the age of genomic medicine.* Genome medicine, 2015. **7**(1): p. 119.
18. Bui, H.-H., et al., *Predicting population coverage of T-cell epitope-based diagnostics and vaccines.* BMC bioinformatics, 2006. **7**(1): p. 153.
19. Suschak, J.J., J.A. Williams, and C.S. Schmaljohn, *Advancements in DNA vaccine vectors, non-mechanical delivery methods, and molecular adjuvants to increase*

immunogenicity. Human vaccines & immunotherapeutics, 2017. **13**(12): p. 2837-2848.
20. Barnes, E., et al., *Novel adenovirus-based vaccines induce broad and sustained T cell responses to HCV in man.* Science translational medicine, 2012. **4**(115): p. 115ra1-115ra1.
21. http://www.bio-rad.com/webroot/web/pdf/inserts/CDG/en/883602_EN.pdf

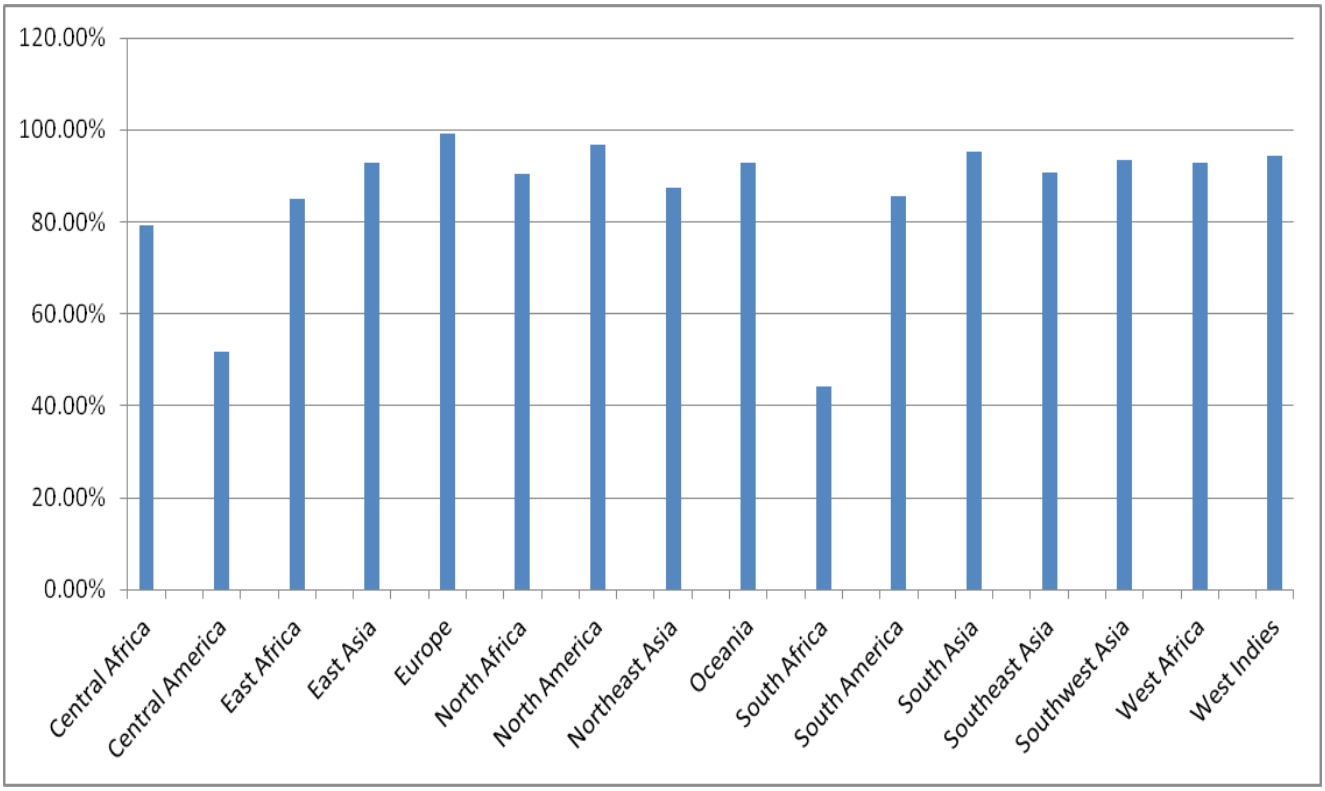

Figure 1: Population coverage of T cell epitopes for different populations of world

Table 1: Accession nos. and information of the proteins used in the study

Replicon Accession	Start	Stop	Protein product	Protein Accession	Length	Protein name
MN996531.1	253	21542	orf1ab	QHR63289.1	7096	orf1ab polyprotein
MN996531.1	21550	25371	S	QHR63290.2	1273	spike glycoprotein
MN996531.1	25380	26207	NS3	QHR63291.1	275	nonstructural protein NS3
MN996531.1	26232	26459	E	QHR63292.1	75	envelope protein
MN996531.1	26510	27178	M	QHR63293.1	222	membrane protein
MN996531.1	27189	27374	NS6	QHR63294.1	61	nonstructural protein NS6
MN996531.1	27381	27746	NS7a	QHR63295.1	121	nonstructural protein NS7a
MN996531.1	27743	27874	NS7b	QHR63296.1	43	nonstructural protein NS7b
MN996531.1	27881	28246	NS8	QHR63297.1	121	nonstructural protein NS8
MN996531.1	28261	29520	N	QHR63298.1	419	nucleocapsid protein

Table 2: Class I MHC epitopes for all viral proteins except polyprotein

Protein	Allele	Epitope	NetMHCpan Score	NetCTL Score
Envelope	HLA-A*02:01	FLAFVVFLL	0.8244220	1.4506
		SLVKPSFYV	0.8004280	1.2840
Membrane	HLA-A*02:01	GLMWLSYFI	0.8681480	1.3561
		KLLEQWNLV	0.8247570	1.0357
	HLA-A*24:02	YFIASFRLF	0.8082140	1.7638
	HLA-B*58:01	MACLVGLMW	0.8119780	1.5697
		LAAVYRINW	0.8018230	1.9702
Orf3a/NS3	HLA-A*01:01	FTSDYYQLY	0.8177230	3.7483
	HLA-A*02:01	YLYALVYFL	0.9156070	1.5407
		LLYDANYFL	0.9112100	1.5693
	HLA-B*27:05	KRWQLALSK	0.8220400	2.0143
Orf7b/NS7b	HLA-A*02:01	FLAFLLFLV	0.8341510	1.4810
		YLCFLAFLL	0.8064860	1.2086
Nucleocapsid	HLA-B*07:02	FPRGQGVPI	0.8570280	1.6399
Spike	HLA-A*01:01	LTDEMIAQY	0.9987410	3.6616
		WTAGAAAYY	0.9621710	3.1128
		TSNQVAVLY	0.9584610	3.0758
		NIDGYFKIY	0.8607610	1.9606
		CVADYSVLY	0.8232600	2.5759
		GAEHVNNSY	0.8084830	1.9960
	HLA-A*02:01	YLQPRTFLL	0.9730320	1.5152
		VLNDILSRL	0.9507120	1.3533
		KIADYNYKL	0.9089980	1.4347
		RLDKVEAEV	0.8609410	1.0612
		FIAGLIAIV	0.8207200	1.2124
	HLA-A*03:01	RLFRKSNLK	0.9588540	1.7563
		GVYFASTEK	0.9486550	1.4615
		VTYVPAQEK	0.9086670	1.3960
		TLKSFTVEK	0.8988250	1.3483
		QIYKTPPIK	0.8933990	1.4526
		GVYYHKNNK	0.8870630	1.3335
	HLA-A*24:02	VYSTGSNVF	0.9696490	1.8571
		QYIKWPWYI	0.9613400	1.8109
		YFPLQSYGF	0.9441900	1.6931
		NYNYLYRLF	0.9352580	1.9482
		PFFSNVTWF	0.8728450	1.4947
		RFDNPVLPF	0.8537630	1.0633
		EYVSQPFLM	0.8483680	1.7025
		YYHKNNKSW	0.8415660	1.2214

		PYRVVVLSF	0.8166080	1.8786
		VYDPLQPEL	0.8101800	1.1735
	HLA-A*26:01	WTAGAAAYY	0.9337910	2.0048
		FVFKNIDGY	0.9119180	2.2795
		CVADYSVLY	0.8578490	1.5087
	HLA-B*27:05	GRLQSLQTY	0.9825460	1.7162
		VRFPNITNL	0.9817510	1.0199
		TRFQTLLAL	0.9782090	1.4733
		TRFASVYAW	0.9414270	1.1380
		TRTQLPPAY	0.8725600	1.0554
	HLA-B*39:01	TRFQTLLAL	0.9487980	1.6214
		VRFPNITNL	0.8628590	1.0215
	HLA-B*58:01	HADQLTPTW	0.9737440	1.7854
		RSFIEDLLF	0.9372980	1.9914
		GTITSGWTF	0.8789670	1.5724
		LAGTITSGW	0.8666350	1.4057
		NSIAIPTNF	0.8133100	1.6445
		QSAPHGVVF	0.8091400	1.6400
		IAIPTNFTI	0.8078680	1.5865

Table 3: Class II MHC epitopes for all viral proteins except polyprotein

Protein	Allele	Epitope	Core
Envelope	DRB1_0101	LVTLAILTALRLCAY	LAILTALRL
	HLA-DPA10103-DPB10601	LLFLAFVVFLLVTLA	FLAFVVFLL
		NSVLLFLAFVVFLLV	VLLFLAFVV
		LFLAFVVFLLVTLAI	AFVVFLLVT
		LAFVVFLLVTLAILT	VFLLVTLAI
Membrane	DRB1_0101	TLSYYKLGASQRVAG	YKLGASQRV
		ILRGHLRIAGHHLGR	HLRIAGHHL
	HLA-DQA10201-DQB10301	INWITGGIAIAMACL	ITGGIAIAM
		RINWITGGIAIAMAC	WITGGIAIA
		WITGGIAIAMACLVG	TGGIAIAMA
		PKEITVATSRTLSYY	TVATSRTLS
Orf3/NS3	DRB1_0101	KKRWQLALSKGVHFV	WQLALSKGV
		FTSDYYQLYSTQLST	YYQLYSTQL
		RWQLALSKGVHFVCN	LALSKGVHF
		ALVYFLQSINFVRII	FLQSINFVR
	HLA-DPA10103-DPB10401	VLHSYFTSDYYQLYS	FTSDYYQLY
		VYFLQSINFVRIIMR	LQSINFVRI
	HLA-DPA10103-DPB10601	LYLYALVYFLQSINF	LYALVYFLQ
		EAPFLYLYALVYFLQ	FLYLYALVY
		LLLLFVTVYSHLLLV	LLFVTVYSH
		ALVYFLQSINFVRII	FLQSINFVR
		FVCNLLLLFVTVYSH	CNLLLLFVT
	HLA-DPA10301-DPB10402	VYFLQSINFVRIIMR	FLQSINFVR
	HLA-DQA10201-DQB10301	IPYNSVTSSIVITSG	SVTSSIVIT
		DFVRATATIPIQASL	FVRATATIP
Orf6/NS6	DRB1_0101	EILLIIMRTFKVSIW	LIIMRTFKV
	HLA-DPA10103-DPB10601	MFHLVDFQVTIAEIL	MFHLVDFQV
	HLA-DQA10101-DQB10501	FKVSIWNLDYIINLI	KVSIWNLDY
		KVSIWNLDYIINLII	WNLDYIINL
Orf7a/NS7a	DRB1_0101	VKHVYQLRARSVSPK	VYQLRARSV
	HLA-DPA10103-DPB10401	NKFALTCFSTQFAFA	CFSTQFAFA
		DNKFALTCFSTQFAF	FALTCFSTQ
	HLA-DPA10103-DPB10601	VAAIVFITLCFTLKR	VFITLCFTL
		DNKFALTCFSTQFAF	KFALTCFST
		LIVAAIVFITLCFTL	IVAAIVFIT
		MKIILFLALITLATC	IILFLALIT
		KFALTCFSTQFAFAC	TCFSTQFAF
Orf7b/NS7b	HLA-DPA10103-DPB10601	DFYLCFLAFLLFLVL	YLCFLAFLL
		FYLCFLAFLLFLVLI	FLAFLLFLV

		SLIDFYLCFLAFLLF	DFYLCFLAF
		FLAFLLFLVLIMLII	FLLFLVLIM
		AFLLFLVLIMLIIFW	LFLVLIMLI
		LIMLIIFWFSLELQD	MLIIFWFSL
		LLFLVLIMLIIFWFS	FLVLIMLII
		LFLVLIMLIIFWFSL	LIMLIIFWF
		MLIIFWFSLELQDHN	IIFWFSLEL
		IELSLIDFYLCFLAF	SLIDFYLCF
Orf8/NS8	HLA-DQA10101-DQB10501	LVVRCSFYEDFLEYH	VRCSFYEDF
		VRCSFYEDFLEYHDV	RCSFYEDFL
	HLA-DQA10102-DQB10501	LVFLGIITTVAAFHQ	IITTVAAFH
Nucleocapsid	DRB1_0101	GIIWVATEGALNTPK	WVATEGALN
Spike	HLA-DRB1_0101	VLSFELLHAPATVCG	FELLHAPAT
		LQTYVTQQLIRAAEI	YVTQQLIRA
		YVGYLQPRTFLLKYN	YLQPRTFLL
		TGRLQSLQTYVTQQL	LQSLQTYVT
		IIAYTMSLGAENSVA	YTMSLGAEN
	HLA-DPA10103-DPB10401	FGEVFNATRFASVYA	FNATRFASV
	HLA-DPA10103-DPB1060	MFVFLVLLPLVSSQC	FLVLLPLVS
		PYRVVVLSFELLHAP	RVVVLSFEL
		FVFLVLLPLVSSQCV	FLVLLPLVS
		GVVFLHVTYVPAQEK	VFLHVTYVP
	HLA-DQA10201-DQB10301	DSSSGWTAGAAAYYV	SGWTAGAAA
		SSSGWTAGAAAYYVG	TAGAAAYYV
		IQDSLSSTASALGKL	LSSTASALG

Table 4: Class MHC II epitopes combined after analysis of overlaps and position in the respective protein

Peptide Sequence	Viral Protein	Position
NSVLLFLAFVVFLLVTLAI	Envelope	15-34
RINWITGGIAIAMACLVG	Membrane	72-90
EAPFLYLYALVYFLQSINFVRIIMR	Orf3/NS3	102-120
KKRWQLALSKGVHFVCN	Orf3/NS3	126-142
VLHSYFTSDYYQLYSTQLST	Orf3/NS3	202-221
MFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINLII	Orf6/NS6	1-37
MKIILFLALITLATC	Orf7a/NS7a	1-15
DNKFALTCFSTQFAFAC	Orf7a/NS7a	51-67
VKHVYQLRARSVSPK	Orf7a/NS7a	71-84
LIVAAIVFITLCFTLKR	Orf7a/NS7a	102-118
IELSLIDFYLCFLAFLLFLVLIMLIIFWFSLELQDHN	Orf7b/NS7b	2-38
LVFLGIITTVAAFHQ	Orf8/NS8	4-18
LVVRCSFYEDFLEYHDVR	Orf8/NS8	98-115
MFVFLVLLPLVSSQCV	Spike	1-16
DSSSGWTAGAAAYYVGYLQPRTFLLKYN	Spike	253-280
FGEVFNATRFASVYA	Spike	338-352
PYRVVVLSFELLHAPATVCG	Spike	506-526
IIAYTMSLGAENSVA	Spike	692-706
IQDSLSSTASALGKL	Spike	934-948
TGRLQSLQTYVTQQLIRAAEI	Spike	998-1018
GVVFLHVTYVPAQEK	Spike	1059-1073

Table 5: B cell epitopes predicted by both servers and their position in the respective protein

Protein	B cell Epitope	Position
Envelope	YVYSRVKNLNSSRVPD	57-72
Membrane	NGTITVEELKKLLEQW	5-21
	KLGASQRVAGDS	180-191
	RYRIGNYKLNTDHSSSSDNIA	198-218
NS3	QGEIKDATPSDF	17-28
	KIITLKKRWQL	61-71
	GDGTTSPISEHDYQIGGYTEKWESGV	172-197
	DEPEEHVQIHTIDGSSGVVNPVMEPIYDEPTTTTS	238-272
NS6	LTENKYSQLDEEQP	44-57
NS7a	EPCSSGTYEGNSPFHPLAD	33-51
	KHVYQLRARSVSPKLFIRQEEVQEL	72-96
NS8	QSCTQHQPYVVDDPCPIHFYSKW	23-45
	RVGARKSAPL	48-57
	VDEAGSKSPIQYIDIGN	62-78
Nucleocapsid	NGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNN	4-49
	HGKEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGKMKDLS	59-105
	GALNTPKDHIGTRNPANNAAI	137-157
	TLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD	166-216
	LNQLESKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATK	227-266
	RRGPEQTQGNFGDQELIRQGTDYK	276-299
	DAYKTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADS	358-416
Spike	QCVNLTTRTQLPPAYTNSFTRGV	14-36
	LGVYYHKNNKSWMESEFRVYSSA	141-163
	DLEGKQGNFKNLRE	178-191
	HTPINLVRDLPQGFSA	207-222
	YLTPGDSSSGWTA	248-260
	YQTSNFRVQP	313-322
	FGEVFNATRFASVYAWNRK	338-356
	NSASFSTFKCYGVSPTKLNDLCFTNV	370-395
	GDEVRQIAPGQTGKIADYNYK	404-424
	NNLDSKVGGNYNY	440-451
	LFRKSNLKPFERDISTEIYQAGST	455-478
	VEGFNCYFPLQ	483-493
	ELLHAPATVCGPKKSTNLVK	516-535
	ADQLTPTWRVYSTGSNVFQT	626-645
	VNNSYECDIP	656-665

	SYQTQTNSPRRARSVASQS	673-691
	AYTMSLGAENSVAYSN	694-709
	KQIYKTPPIKDFGGF	786-800
	LPDPSKPSKR	806-815
	LADAGFIKQYGDCLGD	828-843
	GQSKRVDFC	1035-1043
	FYEPQIITTD	1109-1118
	VNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGI	1133-1172
	SCCKFDEDDSEPVLKGVKL	1252-1270

Supplementary Table 1: Class I MHC epitopes in orf1ab/polyprotein

Allele	Epitope	NetMHCpan Score	NetCTL Score
HLA-A*01:01	CTDDNALAY	0.8961040	3.7761
	DTDFVNEFY	0.8003860	3.6267
HLA-A*02:01	FLAHIQWMV	0.9443830	1.4937
	ILFTRFFYV	0.9258750	1.4756
	FLLPSLATV	0.9145480	1.4888
	FLLNKEMYL	0.9128610	1.4498
	YLDAYNMMI	0.8985970	1.4815
	TMADLVYAL	0.8817300	1.4389
	TLMNVLTLV	0.8777620	1.3765
	YLNTLTLAV	0.8680030	1.4705
	SMWALIISV	0.8608980	1.4970
	LLLDDFVEI	0.8573470	1.4579
	FLARGIVFM	0.8532880	1.2718
	FVNEFYAYL	0.8508950	1.3744
	NLIDSYFVV	0.8506660	1.3372
HLA-A*03:01	RLISMMGFK	0.8102790	1.6603
HLA-A*24:02	NYMPYFFTL	0.8651630	1.9562
HLA-A*26:01	ETISLAGSY	0.8870150	2.5695
HLA-B*07:02	IPRRNVATL	0.8455180	1.6725
	KPNELSRVL	0.8054780	1.6226
HLA-B*08:01	YLRKHFSMM	0.8473990	2.1362
	LMIERFVSL	0.8255850	1.4415
HLA-A*39:01	SHFAIGLAL	0.8242740	2.1706
	MHAASGNLL	0.8012630	2.0434
HLA-B*40:01	AEWFLAYIL	0.8116560	2.0555
	GEAANFCAL	0.8055890	1.9928
HLA-B*58:01	KSHNIALIW	0.8755450	1.8757
	RTIKGTHHW	0.8699230	1.9496
	LAAVNSVPW	0.8389900	1.8945
	VSFLAHIQW	0.8289440	2.0111
	TAFGLVAEW	0.8128710	1.7994
	KAYKIEELF	0.8112020	1.8582
	WSMATYYLF	0.8010690	1.5112
HLA-B*15:01	VMYMGTLSY	0.8464560	1.5666
	VQMAPISAM	0.8194500	1.4035
	RLYYDSMSY	0.8085090	1.5023
	VMFTPLVPF	0.8001060	1.4760

Supplementary Table 2: Class II MHC epitopes in orf1ab/polyprotein

Allele	Epitope	Core
DRB1_0101	ESPFVMMSAPPAQYE	FVMMSAPPA
	YFNMVYMPASWVMRI	MVYMPASWV
	KYKFVRIQPGQTFSV	FVRIQPGQT
	TPVYSFLPGVYSVIY	YSFLPGVYS
	PEFYEAMYTPHTVLQ	YEAMYTPHT
	LCLFLLPSLATVAYF	LLPSLATVA
	SHRFYRLANECAQVL	FYRLANECA
	QKLLKSIAATRGATV	LKSIAATRG
	CLFLLPSLATVAYFN	LLPSLATVA
	MTYRRLISMMGFKMN	YRRLISMMG
	RRNVATLQAENVTGL	VATLQAENV
	ASKILGLPTQTVDSS	ILGLPTQTV
	SHFVNLDNLRANNTK	FVNLDNLRA
	LKGKYVQIPTTCAND	YVQIPTTCA
	PHTVLQAVGACVLCN	VLQAVGACV
	KFVRIQPGQTFSVLA	FVRIQPGQT
	LIVTALRANSAVKLQ	VTALRANSA
	KEMYLKLRSDVLLPL	YLKLRSDVL
DRB1_0404	GLVASIKNFKSVLYY	SIKNFKSVL
	HSLSHFVNLDNLRAN	HFVNLDNLR
	AMGIIAMSAFAMMFV	GIIAMSAFA
HLA-DPA10103-DPB10301	MIERFVSLAIDAYPL	ERFVSLAID
	MTYRRLISMMGFKMN	YRRLISMMG
HLA-DPA10103-DPB10401	LFFFLYENAFLPFAM	LYENAFLPF
	AVNLLTNMFTPLIQP	LLTNMFTPL
HLA-DPA10103-DPB10601	FLAYILFTRFFYVLG	YILFTRFFY
	VAEWFLAYILFTRFF	FLAYILFTR
	SLFFFLYENAFLPFA	FLYENAFLP
	ASRELKVTFFPDLNG	ELKVTFFPD
	RMYIFFASFYYVWKS	YIFFASFYY
	YMPYFFTLLLQLCT	PYFFTLLLQ
	MPYFFTLLLQLCTFT	YFFTLLLQL
	SVIYLYLTFYLTNDV	YLYLTFYLT
	MYIFFASFYYVWKSY	IFFASFYYV
	KHAFLCLFLLPSLAT	AFLCLFLLP
	NFNVLFSTVFPPTSF	VLFSTVFPP

	QWSLFFFLYENAFLP	SLFFFLYEN
	MLVYCFLGYFCTCYF	YCFLGYFCT
	VYSVIYLYLTFYLTN	IYLYLTFYL
	ILFTRFFYVLGLAAI	FFYVLGLAA
	YILFTRFFYVLGLAA	LFTRFFYVL
HLA-DQA10102-DQB10501	VPLNIIPLTTAAKLM	IIPLTTAAK
	WLIINLVQMAPISAM	LVQMAPISA
	AIILASFSASTSAFV	LASFSASTS
	MWLIINLVQMAPISA	IINLVQMAP
	MPNMLRIMASLVLAR	MLRIMASLV
	LLMPILTLTRALTAE	ILTLTRALT
	MWALIISVTSNYSGV	IISVTSNYS
	PLIVTALRANSAVKL	VTALRANSA
	PFVMMSAPPAQYELK	VMMSAPPAQ
	AALGVLMSNLGMPSY	LGVLMSNLG
	ALGVLMSNLGMPSYC	LMSNLGMPS
	KTTVASLINTLNDLN	VASLINTLN
	SPFVMMSAPPAQYEL	VMMSAPPAQ
HLA-DQA10201-DQB10301	LASFSASTSAFVETV	FSASTSAFV
	IILASFSASTSAFVE	ASFSASTSA
	EYTDFATSACVLAAE	FATSACVLA
	SFSASTSAFVETVKG	ASTSAFVET
	PPQTSITSAVLQSGF	TSITSAVLQ
	GSLIYSTAALGVLMS	LIYSTAALG
	RQMSCAAGTTQTACT	CAAGTTQTA
	ISASIVAGGIVAIVV	IVAGGIVAI
	DISASIVAGGIVAIV	SIVAGGIVA
	TSITSAVLQSGFRKM	SITSAVLQS
	KPLEFGATSAALQPE	FGATSAALQ
	IVYTACSHAAVDALC	YTACSHAAV
	NIIPLTTAAKLMVVI	LTTAAKLMV
	LCTHTGTGQAITVTP	TGTGQAITV
	EFSSLPSYAAFATAQ	PSYAAFATA
	ISVTSNYSGVVTTVM	SNYSGVVTT
	TSSGDATTAYANSVF	TSSGDATTA
	NVLSTFISAARQGFV	VLSTFISAA
HLA-DPA10301-DPB10402	ETKFLTENLLLYIDI	KFLTENLLL